MENTAL HEALTH SENSING USING MOBILE PHONES

A Thesis

Submitted to the Faculty

in partial fulfillment of the requirements for the

degree of

Doctor of Philosophy

in

Computer Science

by

Rui Wang

DARTMOUTH COLLEGE

Hanover, New Hampshire

September 28, 2018

Examining Committee:

Andrew T. Campbell, Ph.D., Chair

Xia Zhou, Ph.D.

Venkatramanan S. Subrahmanian, Ph.D.

Gregory D. Abowd, Ph.D.

F. Jon Kull, Ph.D.
Dean of the Guarini School of Graduate and Advanced Studies

Abstract

Effectively assessing and monitoring mental health is critical to the detection and treatment of mental illness. Mobile technologies, specifically mobile sensing using smartphones, have potential to provide abundant real time information about a person's behaviors, lifestyle, and symptoms. In this thesis, we present two studies using smartphone sensing to assess mental wellbeing. The *StudentLife study* uses the StudentLife sensing system to collect passive sensing data, EMA (ecological momentary assessment), and a number of well-known pre-post behavioral and mental health surveys from 48 Dartmouth students during the spring term in 2013. We identify a Dartmouth term lifecycle in the data that shows how students' stress, sleep, and daily activity patterns change as the term progresses. We find a number of significant correlations between the automatic objective sensor data from smartphones and mental health and educational outcomes of the student body. We discuss a follow-up study in which we propose a set of "symptom features" that proxy the DSM-5 (Diagnostic and Statistical Manual of Mental Disorders) defined depression symptoms specifically designed for college students. We identify a number of important new associations between symptom features and student self reported PHQ-8 (Personal Health Questionnaire Depression Scale) and PHQ-4 depression scores. We show that symptom features derived from phone and wearable sensors can predict whether or

not a student is depressed on a week by week basis with 81.5% recall and 69.1% precision. In the second part of the thesis, we discuss our contribution to assessing serious mental health using mobile phone sensing. We discuss *CrossCheck*, a randomized control trial that aims to use smartphones to assess schizophrenia patients' symptoms and ultimately predict schizophrenia relapses using passive sensing from smartphones. 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. We build a symptom prediction system to track participants' symptoms in weekly basis and reach out to participants who are determined by the system as at risk. Finally, we evaluate relapse prediction models that predict whether or not a participant is going to relapse based on smartphone passive sensing data and self-report EMA.

Acknowledgments

First, I thank my advisor Prof. Andrew T. Campbell for the continuous support of my Ph.D study and research. He has profoundly influenced my view of research and taught me the importance of producing high-quality research that reach disciplines outside computer science. He provided an inspiring environment where I can connect and collaborate with other researchers. I value his honest opinions, patience, and understanding over the past six years. I could not finish this thesis without his invaluable mentorship, guidance, and support.

I feel fortunate and honored to have had the opportunity to work closely with Prof. Dror Ben-Zeev, Rachel Brian, Prof. Tanzeem Choudhury, Hane Aung, Prof. Mi Zhang, and Prof. Emily Scherer in the CrossCheck project.

I enjoyed working with Bret Peterson, Arthur Brant, Yuanwei Chen, Menachem Fromer, Honor Hsin, and Collin Walter from Verily, and Jun Yang from Huami for my internships.

I would like to thank Prof. Xia Zhou and Prof. Gabriella Harari for their support and guidance in different stages of my PhD. life. I would like to thank all present and past members from the dartnets and the Smartphone Sensing Group: Chuang-Wen You, Fanglin Chen, Zhenyu Chen, Tianxing Li, Peilin Hao, Weichen Wang, Tianxing Li, Zhao Tian, Kizito Masaba, Shayan Mirjafari, Yichen Li, Ruibo Liu, Xi Xiong,

and Chuankai An. I am thankful to all the time they spent with me.

Finally, I am thankful to have the support from my family. I especially thank my wife Jing Yang for her love, understanding, and support. Finally, I can not thank my parents enough. They provided me good education and also supported me.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

Effectively assessing and monitoring mental health is critical to the detection and treatment of mental illness. Early detection of mental illness warning signs could facilitate time-sensitive interventions. However, traditional clinical practices are inefficient in detecting early warning signs. Standard methods are based on face to face interaction and assessment by clinicians, conducted at set times and locations. Such assessments are limited to the information the patients report back about their behaviors and symptoms. As a result, such data are often influenced by biases in self-report thus may not accurately capture patients' day-to-day functioning. Mobile technologies, specifically mobile sensing using smartphones, have the potential to overcome drawbacks with traditional mental health assessment and provide more abundant real time information about patients' behaviors, lifestyle, and symptoms.

About 77% of Americans in 2018 own smartphones according to Pew Research[1].

[1] http://www.pewinternet.org/fact-sheet/mobile/

1.1 Overview

Smartphones are equipped with many sensors that are capable of detecting people's behaviors, and are nearly constantly carried by their owners, which provide unparalleled access to people's daily lives. Smartphones can also be used to query people about their psychological states (e.g., stress, emotion, mood) by notifying users to answer survey questions. Researchers have already begun to use smartphones as behavioral data collection tools and investigated using smartphone sensing to monitor mental health outcomes. The MONARCA project [150, 149, 96, 162] first report on findings from mobile sensing and bipolar disorder. The authors [150] discuss correlations between the activity levels over different periods of the day and psychiatric evaluation scores associated with the mania-depression spectrum. Early work on mobile mental health for schizophrenia patients by Ben-Zeev et al. [27, 30, 29, 28, 32] studies the feasibility and acceptability of using mobile devices for intervention and self-management among individuals with schizophrenia. In [33] the authors find that participants feel comfortable using mobile phones with passive sensing apps. Participants also report that they are interested in receiving feedback and suggestions regarding their health. In [78], the authors present a study of 79 college-age participant from October 2015 to May 2016 in which they find a number of mobility features (e.g., home stay duration, normalized entropy) correlate with PHQ-9 (patient health questionnaire-9) [174, 122, 121], a widely used depression screening survey. Canzian et al [49] develop an expanded set of mobility features [164, 163] and find that location data correlates with PHQ-9. The authors show that the maximum distance traveled between two places strongly correlates with the PHQ score.

There are a number of questions that need to be addressed before running a smartphone sensing study for mental health [95]. Which device and sensing applica-

tion should we use? How long should the study run? How often should we sample the sensors? How do we obtain behavioral variables from the smartphone data? What are the mental health outcomes? This thesis makes contributions that address these questions. Specifically, we present two studies, *StudentLife* [196, 197, 200] and *CrossCheck* [195, 199], in which we discuss our smartphone sensing systems, study design, backend data analytics systems, behavioral features derived from passive sensing data, and mental health outcome predictions. In what follows, we review smartphone sensing systems, behavioral modeling methods, and mental health outcomes.

1.1.1 Smartphone Sensing Systems

A smartphone sensing system usually consists of a sensing app running on the phone and a backend service running in the cloud. The sensing app collects data by sampling from a series of sensors, apps, and phone logs and uploads the data to the backend service. The backend service consists a number of behind-the-scenes features to facilitate data collection. For example, the backend service stores the uploaded sensing data in a database and provide tools to manage participants and monitor study adherence. In what follows, we review the sensing app and the backend service.

The Smartphone Sensing App

A smartphone is a device that combines sensing, computing, and communication. Phone manufactures have equipped smartphones with a number of sensors that are capable to perceive its user's activities and surroundings. For example, accelerometers are used to infer a user's physical activities (e.g., stationary, walking, running, in a vehicle), GPS is used to record a user's mobility (e.g., places visited, distance traveled,

mobility routines), microphones are used to infer social interaction, light sensors are used to measure the ambient light environment, and smartphone operating systems record lock/unlock events that are used to measure phone use. The sensing app leverages all the sensors on the phone to collect various sensing data from smartphones and apply machine learning models to infer users' behaviors. Specifically, a sensing app collects a user's physical activity, social interaction, mobility, sleep, phone use, and self-report EMAs.

Activity detection. There is a large amount of work on activity recognition using smartphones and wearable sensors [22, 59, 107, 106, 183, 39]. Specifically, prior work [127, 131] develop physical activity classifiers for smartphones to infer stationary, walking, running, driving and cycling based on features extracted from accelerometer streams. The activity classifier extracts features from the preprocessed accelerometer stream, then applies a decision tree to infer the activity using the features. The activity classifier achieves overall 94% of accuracy [131]. In [183, 39], researchers use wearable sensors to detect eating behaviors. Researchers assess sensor-, device- and workload-specific heterogeneities (e.g., heterogeneities across devices and their configurations) for activity recognition in [177]. Smartphone operating systems now provide built-in activity recognition APIs (e.g., Android Activity Recognition[2] and iOS Core Motion[3]) that infer users' physical activities as well as step counts.

Conversation detection. Smartphones are able to infer whether or not there is a conversation around a user aka social interaction. Prior work [155, 127] develop privacy-sensitive audio and conversation classifiers. The classifiers process audio recorded using the phone's microphone on the fly to extract and record features,

[2]https://developers.google.com/location-context/activity-recognition/
[3]https://developer.apple.com/documentation/coremotion

and use a two-state hidden Markov model (HMM) to infer speech segments. The speech segments are then grouped into distinct conversations. In order to protect privacy, only the conversation inferences are recorded.

Mobility detection. The GPS sensor on the smartphoens identifies locations, which can infer the significant places (e.g., home, work, traveling) in a user's daily routine [17]. We can also use GPS traces to compute distance traveled and dwell time at different types of locations (e.g., home, gym, work places, study places). To obtain more fine-grained mobility information, we can also use WiFi scan logs to identify a user's indoor locations [194, 157].

Sleep detection. Our previous work [53, 127] implements a sleep classifier that unobtrusively infers sleep duration without any special interaction with the phone. The sleep classifier extracts four types of features: light features, phone usage features including the phone lock state, activity features (*e.g.*, stationary), and sound features from the microphone. Our sleep model combines these features to form a more accurate sleep model and predictor. We train the model using the method described in [53] with an accuracy of +/- 32 mins to the ground truth.

Phone use detection. Smartphone operating systems record lock/unlock events, app usage, and call/SMS logs. User interaction with the phone is potentially indicative of general daily functioning. Many studies explore the relationship between smartphone use and mental health outcomes [66, 75, 76]. In a prior study [172], researchers collect a wide range of phone usage data from smartphones and identify a number of usage features (e.g., the number of apps used per day, the ratio of SMSs to calls, the number of event-initiated sessions) that are relevant to problematic smartphone use.

MobileEMA. In-situ EMA (ecological momentary assessment) [171] on smartphones is used in studies to capture additional human behavior beyond what the surveys and automatic sensing provide. The user is prompted by a short survey scheduled at some point during their day. We can use MobileEMA to collect a large range of self-reports, such as stress, anxiety, depression (e.g., PHQ-4 [123]), schizophrenia symptoms [31], and general behaviors [196].

The sensing app runs in the background and does not need users' input except periodic MobileEMA. The collected data are first stored in the phone and uploaded to the backend service when possible.

The Backend Service

The back-end service usually consists of a participant manager, a portal server, a data storage server, and a data processing service. The participant manager manages study participants' subject IDs, which are used by the sensing app to authenticate with the portal server. The portal server receives the data from the sensing app. The portal server stores the sensor data uploaded from participant's phones in the data storage, which is typically a database (e.g., MySQL, MongoDB). The data processing service relies on the data storage server to provide services like data collection monitoring, data transformations (for further analysis), and prediction services (e.g., predicting mental health outcomes).

1.1.2 Behavioral Modeling

The smartphone sensing data capture a user's behavior at the moment. For example, a physically activity inference describes the user's activity at the time the data is

collected. A GPS coordinate describes where a user is at that moment. We need behavioral modeling methods to aggregate and summarize large volumes of sensing data to describe the user's behaviors at a higher level. As a result, the modeling methods generate a number of behavioral features that are the input to our analysis. In what follows, we describe two methods to model behaviors from smartphone sensing data.

Compute daily behavioral features. One intuitive method to model behaviors is to compute daily summaries for each of the sensing data. For example, we compute the non-sedentary duration in a day from the activity data to describe how physically active a user is; we can compute the number of conversations and the duration of all conversations in a day to describe sociability; we compute distance traveled from the location data to describe mobility. We can also partition a day into epochs (e.g., night: 12 am - 6 am, morning: 6am - 12 pm, afternoon: 12 pm - 6 pm, and evening: 6 pm - 12 am) and compute the behavior summaries in each of the epochs. These epochs relate to morning, afternoon, evening, and night. Partitioning a day into epochs gives insight about the structure of the daily behaviors (e.g., more active at night but sedentary in the day), which may be informative to mental health states.

Combining different types of sensor data. The combination of different types of sensor data can produce more context-specific behaviors. For example, the StudentLife study [196, 197] described in Chapter 2 shows we can combine location, activity, and audio data to infer partying, studying, and how focused a student is when studying. Domain knowledge may help in combining different types of sensor data in meaningful ways. In Chapter 3, we discuss designing behavioral features that capture depression symptoms defined in DSM-5 (Diagnostic and Statistical Manual

of Mental Disorders, 5th Edition) [18].

Mining behavioral patterns. Another method to model human behaviors from smartphone data is to compute features that capture people's lifestyle. For example, researchers [21] propose a generalizable solution to model and reason about behavioral routines. Routines can describe people's sleeping, daily mobility, and socializing patterns. Saeb et al [164, 163] find mobility patterns (e.g., circadian rhythm of movement, normalized entropy) correlate with depressive symptom severity. Canzian et al [49] present a number of mobility trace features (e.g., the radius of gyration, the routine index) that predict the trajectory of depression. Abdullah et al [10, 9] investigate assessing circadian rhythms from sensors and self-reports to help improving cognitive and physical performance.

1.1.3 Mental Health Measures and Analysis Methods

Mobile mental health studies aim to predict certain measures using the smartphone data. However, the mental health measures and sample methods varies between studies. There are mainly three types of ground truth: one-time, periodic, and mental health outcome events. In what follows, we discuss our ground truth and analysis methods.

One-time ground truth. There are a number of survey instruments to assess mental health. For example, PHQ-9 [174, 122, 121] is a widely used instrument to screen for depression severity; the Generalized Anxiety Disorder 7 (GAD-7) [175] is a questionnaire for measuring generalized anxiety disorder; and the brief psychiatric rating scale (BPRS) [151] is a rating scale used by clinicians to measure psychiatric symptoms such as depression, hallucinations, and unusual behaviors. The survey

instruments are usually administered at the beginning and/or the end of a study.

We are interested in two types of analysis. First, we would like to understand the relationship between sensing data and ground truth. Second, we would like to predict the ground truth using the sensing data. To understand the relations between sensing data and mental health ground truth, we can apply various standard statistical tests. For example, we can apply correlation analysis and regression analysis if the ground truth is ordinal (i.e., the order of choices,) or interval (i.e., difference between values is meaningful). We can apply t-test and ANOVA if the ground truth is nominal (i.e., categorical values). There are usually a large number of behavioral features derived from sensing data, therefore, we run many separate hypothesis tests. If we use the standard alpha level of 5%, we would expect to find many false discoveries (i.e., the multiple testing problem). For example, if we run 1000 random tests with the standard alpha level of 5%, we would get around 50 significant results, most of which are false discoveries. We could apply various controlling procedures to address the multiple testing problem (e.g., the Bonferroni correction [70], the false discovery rate (FDR) [207]). For example, FDR-controlling procedures are designed to control the expected proportion of false "discoveries" [207].

To predict mental health ground truth using sensing data, we apply various machine learning models. We first partition the dataset into two parts: a training set and a test set. We use the training set to train machine learning models and evaluate the models using the test set. If the ground truth is ordinal or interval, we train regression models, such as linear regression [83], support vector regression [61], and gradient boosting regression tree [84]. If the ground truth is nominal, we train classification models, such as logistic regression [193, 158], support vector machines [61],

and random forest [44, 104]. We use cross-validation methods [146, 117] to evaluate predictive models. Cross-validation may have a number of rounds. A round of cross-validation partitions a sample of data into subsets, trains the model on one subset (i.e., the training set), and validates the model on the other subset (i.e., the testing set).

Periodic ground truth. Other studies aim to track how mental health states change over time. Researchers might administer EMAs periodically (e.g., once a week) to assess participants' mental health. The EMA questions are usually short and quick to respond to. For depression studies, researchers might administer PHQ-4 [123], which consists of 4 short questions. For schizophrenia studies described in Chapter 4, we administer a 10-item EMA to participants every Monday, Wednesday, and Friday. The EMA questions ask participants to self-report their symptom severity (e,g., hearing voices, depression) and overall mental health wellbeing (e.g., feeling hopeful).

Since participants respond to EMAs many times in the study, the data from the same participant are correlated. Therefore, we cannot apply traditional statistical models to understand the relationship between sensing data and ground truth. Instead, we apply models that properly address the within-individual dependencies, such as generalized linear mixed models (GLMM) [128, 139] and generalized estimating equations (GEE) [46]. For predicting periodic mental health ground truth, we need to carefully design the cross-validation method to avoid including correlated data in the training set and the test set. For example, we could apply leave-one-subject-out cross-validation [146, 117] to evaluate prediction performance. If leave-one-subject-out cross-validation is not practical, we could partition the data in a way that there

are significant time gaps between the training data and test data, i.e., any examples in the test data should not be close in time to any examples in the training data.

Mental health outcome events. Other studies aim to predict certain mental health outcomes, for example, hospitalization and relapse (i.e., a recurrence of a mental illness condition). These mental health outcomes usually do not occur frequently, which are very challenging for the analysis. We can frame outcome events detection as an anomaly detection problem, in which we assume mental health outcome events may lead to significant changes in a person's behavior. We can also treat the event detection problem as a binary classification problem, in which we partition the sensing data to fixed-length time windows, and predict whether or not the event happens in a time window. In this case, we can apply analysis methods that are similar to analyzing periodic ground truth. However, we need to apply data augmentation methods (e.g., resampling) to address imbalanced data issues.

1.2 Problem Statement

After discussing smartphone sensing and mobile mental health studies in general, we now present the specific problems we address in this thesis. We explore the following three major problems in assessing mental health using smartphones: 1) building a core sensing system for mobile mental health studies; 2) using the sensing system to reveal college students' mental health, academic performance, and behavioral trends in an academic term (see Chapter 2-3); and 3) applying sensing technology in a more challenging population (people with serious mental illness (SMI)) by tracking schizophrenia patients' symptoms, providing in-time interventions, and predicting relapses (See Chapter 4 - 6). A detailed problem statement is as follows.

1.2.1 Assessing Mental Health and Academic Performance in College Students

Many questions arise when we think about the academic performance and mental well-being of college students. Why do some students do better than others? Under similar conditions, why do some individuals excel while others fail? Why do students burnout, drop classes, even drop out of college? What is the impact of stress, mood, workload, sociability, sleep and mental well-being on educational performance? Consider students at Dartmouth College, an Ivy League college in a small New England college town. Students typically take three classes over a 10-week term and live on campus. Dartmouth classes are generally demanding where student assessment is primarily based on class assignments, projects, midterms and final exams. Students live, work and socialize on a small self-contained campus representing a tightly-knit community. The pace of the 10 week Dartmouth term is fast in comparison to a 15 week semester. The atmosphere among the students on campus seems to visibly change from a relaxed start of term, to an intense midterm and end of term. Typically classes at Dartmouth are small (*e.g.*, 25-50 students), but introductory classes are larger (*e.g.*, 100-170), making it difficult for a faculty to follow the engagement or performance of students on an individual level. Unless students contact a student dean or faculty about problems in their lives, the impact of such challenges on performance remains hidden.

To shine a light on student life we develop the *StudentLife* [196] smartphone app and sensing system to automatically infer human behavior in an energy-efficient manner. We also use EMA to probe students' states (*e.g.*, stress, mood) across the term. We administer a number of well-known pre-post health and behavioral surveys

at the start and end of term. Our goals are the following: 1) build a robust core sensing system to continuously and unobtrusively collect behavioral data from smartphones; 2) test the core sensing sensing system in a study consist of college students; and 3) model students' behaviors using smartphone data and determine the connections between the smartphone data and mental health and academic performance.

1.2.2 Tracking Depression in College Students

Clinical depression or major depressive disorder (MDD) is one of the most common and debilitating health challenges of our time. In 2015, an estimated 6.7% of all U.S. adults had at least one major depressive episode in the past year [166]. Major depressive disorder accounts for a staggeringly high proportion of illness-related burden worldwide [148, 191], and is the second leading cause of years lost to disability in the U.S. [147]. College age young adults 18 to 25 are more likely to have major depressive episodes than any other age groups. In 2015, an estimated 10.3% of young adults had a major depressive episode over the past year with 6.5% reporting the episode resulted in severe impairment [166]. The college years introduce major stressors for young adults that may exacerbate students' propensity for psychopathology, including increased academic pressures, social challenges, unfamiliar living and physical environments, financial pressures, and cultural differences that affect self-worth [97, 112]. Furthermore, students must negotiate loss of familiar support systems and social networks (e.g., high school friends). Consequently, many young adults can feel overwhelmed, struggle to find their place, and become more susceptible to depression or other mood disorders. Surveys at colleges across the U.S. found that 53% of respondents experienced depression at some point after entering college with 9% reporting

suicidal ideation [85]. A recent study of Facebook profiles showed that 25% of students displayed depressive symptoms [144]. In addition, a 2016 study by the American College Health Association found that 38.2% of students at 2- and 4-year institutions reported feeling "so depressed that it was difficult to function" in the past year [14]. At Dartmouth College, a 2016 survey shows that depression and anxiety are the most common health problems, exceeding national averages in the adult population; 19% of Dartmouth students report being diagnosed with depression, and 24% say that depressive symptoms had harmed their academic performance [65]. Importantly, up to 84% of college students who screen positively for depression never seek mental health services [73]. Many students become aware (i.e., insight to illness) of their depression only after experiencing significant functional deterioration [112]. Many colleges offer mental-health services and counseling, but the stigma associated with mental illness is a major barrier to care-seeking [34, 60]. Evidence also suggests that higher education institutions' current approaches are not addressing depression adequately [112]. Depression rates continue to increase [85]. There is a need to understand what is happening on our campus with these increasing depression rates. One thing is clear: the demand for mental services on US campuses is increasing with many institutions not capable of dealing with the rising needs of students. Clinicians, mental health counsellors, teachers, and administrators on our campuses do not understand why this inflection toward higher rising risk has occurred.

Identifying early warning signs of depression (i.e., "red flags") could mitigate or prevent major depression disorder's negative consequences [47, 112]. However, if students do not pay attention to their clinical condition or do not seek care when needed, depression can lead to devastating outcomes, including self-injurious behavior

and suicide [85, 86]. There is a growing realization in academia and industry that everyday mobile phones and wearables (e.g., fitbits, smartwatches), which passively collect and analyze behavioral sensor data 24/7, will complement traditional periodic depression screening methods (e.g., PHQ-9 survey [174, 122, 121]) and visits to mental health specialist – ultimately, if validated at scale mobile sensing has the potential to replace periodic screening questionnaires such as PHQ-9. Recently, researchers have made progress in understanding the relationship between behavioral sensor data from phones and mental health [49, 19, 196, 126, 164, 163]. In addition, there is considerable activity in the startup space in the area of mental health. A number of companies are starting to use mobile technologies to assess and help people living with depression [87, 189, 100, 143]. However, while progress is being made (e.g., significant correlations between mobile sensing data and depression have been found across different studies [196, 163]) to the best of our knowledge there is no mobile passive sensing technology capable of predicting rising risk, impending depressive episodes, or occurrence from a combination of smartphone and wearable passive sensing to date.

We aim to leverage depression symptom domain knowledge to develop symptom features derived from passive sensing data from smartphones that are meaningful and predictive of depression severity.

1.2.3 Assessing Serious Mental Illness

Schizophrenia is a severe and complex psychiatric disorder that develops in approximately 1% of the world's population [190]. 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 [16]. 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 [41, 88]. 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 [145].

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 [115], 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 the FOCUS self management app [28] 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 [30]. A pilot study in the efficacy of tracking patients [33] 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.

The *CrossCheck* study is a randomized control trial (RCT)[50] conducted in collaboration with a large psychiatric hospital in New York City, 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). Participants in the smartphone arm are given a Samsung Galaxy S5 Android phone equipped with the CrossCheck app, which collect a large range of passive sensing data. We aim to find behavioral features derived from the smartphone data that are predictive of schizophrenia symptom severity and ultimately impending relapses. We explore using the symptom predictions to guide informed interventions.

1.3 Protection of Human Subjects

The StudentLife study has been approved by the Committees for the Protection of Human Subjects at Dartmouth College. The CrossCheck 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. We

uses a number of data security techniques to secure the data collected during the StudentLife and CrossCheck studies. All data collected from smartphones (i.e., sensor, survey and EMA data) are stored on secure servers at Dartmouth College. Project data stored on smartphones are transferred periodically to servers using Transport Layer Security (TLS) standard. Servers have a password-protected login. All servers containing project data are located in lockable offices. The research buildings are locked during non-business hours. The data stored on the server are labeled using research numbers instead of names, and any documents containing personally identifiable information needed for project management (e.g., signed consent forms, contact sheets) are stored in locked cabinets in a secure location accessible only to authorized project personnel. Assessment interviews, measures, captured by study staff during the periodic assessments are saved using research numbers rather than names. Access to such data to researchers outside the team is limited as per IRB protocol and user consent. The various StudentLife studies reported in this article is supported the National Institute of Mental Health, grant number 5R01MH059282-12. The CrossCheck study is supported the National Institute of Mental Health, grant number R01MH103148.

1.4 Thesis Contributions

Smartphone sensing opens the door to continuously and unobtrusively tracking people's mental health. The idea of a passive sensing algorithm working efficiently in the background of a mobile phone with the goal of collecting, analyzing and predicting mental health outcomes in a privacy preserving and validated manner with no user burden is potentially game changing for medicine. In this research, we make a

number of contributions toward this vision; more specifically, we make the following contributions:

- We build a core smartphone sensing system that can not only be applied to college students but also to more challenging population: people living with serious mental illness.

- We present the *StudentLife* study, which is the first to use automatic and continuous smartphone sensing to assess mental health, academic performance and behavioral trends of a student body. We collect a large number of smartphone data from 48 Dartmouth students over a 10-week term in 2013. We also collect validated mental health measurements and students' GPAs. We design behavioral features that capture students' life on campus (e.g., partying, studying) by fusing multiple sensor streams. We identify correlations between automatic sensing data and a broad set of well-known mental well-being measures. We propose for the first time a model that can predict a student's cumulative GPA using automatic behavioral sensing data from smartphones. We observe trends in the sensing data, termed the *Dartmouth term lifecycle*, where students start the term with high positive affect and conversation levels, low stress, and healthy sleep and daily activity patterns. As the term progresses and the workload increases, stress appreciably rises while activity, sleep, conversation, positive affect, visits to the gym and class attendance drop.

- We build upon the initial StudentLife study, and propose a set of passive sensor based symptom features derived from phones and wearables that we hypothesize proxy 5 out of the 9 major depressive disorder symptoms defined in DSM-5 [18].

We identify a number of correlations between the symptom features and PHQ-8, and we use ANOVA to compare the means of the symptom features between the *non depressed group* and the *depressed group*, as defined in PHQ-8 [124]. We show that these two groups are clearly identified in our data set. Finally, we show that we can predict PHQ-4 and PHQ-8 using the proposed symptom features.

- We deploy the CrossCheck system (based on the StudentLife sensing core [196]) in a year long randomized control trial, which aims to track symptoms in people with schizophrenia and predict impending relapses. We recruit 61 participants in the smartphone arm, in which participants carry study phones with our sensing app. We collect smartphone passive sensing data, self-report symptom EMAs, clinician assessed BPRS (the brief psychiatric rating scale) [101] symptom scores, and relapse events. We first look to predict participants' self-reported symptoms and we identified 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 as a mean to develop a real-time monitoring and relapse prevention system. We present and evaluate models that can predict participants' aggregated EMA scores that measure several dynamic dimensions of mental health and functioning in people with schizophrenia. We find 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.

1.4 Thesis Contributions

- We build the CrossCheck symptom prediction system, which 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. The CrossCheck symptom prediction system predicts participants' BPRS scores each week. Our research staff use the predictions to determine whether or not a participant is at risk. We discuss anecdotal information associated with three patients in the study. These case studies show that our system can identify patients with rising risk.

- We discuss the design considerations for building relapse prediction system. Specifically, we identify two main challenges: 1) relapse cases are rare, and 2) the CrossCheck relapse dataset is imbalanced. We discuss and evaluate a number of methods to address these challenges. We investigate the efficacy of using passive sensing data and/or self-report EMAs to predict relapses. We present classification performance from using only EMA or sensing data, and a combination of EMA and sensing data. We investigate the best time window to predict relapse. We explore using PCA to transform the feature space and reduce the dimensionality for classification. Finally, we present features that are the most predictive of impending relapse.

There are a growing number of studies that applied similar methodologies described in this thesis to investigate using smartphones to assess mental illnesses, per-

sonality traits, mood, academic performance, and work performance. The Campus-Life Consortium [129] initiated by Abowd et al.in 2015 (and recently broadened into the Community Life Consortium in 2018 by Saeed Abdullah (Penn State), Gabriella Harari (Stanford) and Edison Thomaz (UT Austin)) aims to extend projects like StudentLife to collect data from more campus communities, including Georgia Tech, Cornell University, Penn State, Carnegie Mellon University, Cambridge University, UT Austin, University of Washington, and Dartmouth College. The project collects data from mobile and wearable devices and social media. There is also a growing number of workshops dealing with mobile mental health including the NSF Workshop on Future Technology to Preserve College Student Health and Foster Wellbeing (College Student Health), Northwestern University, Chicago, July 30-31, 2015 (see studenthealth.cs.dartmouth.edu). Finally, we have released the StudentLife dataset [6] to the research community and we will release our StudentLife core sensing system to help future studies in assessing mental health using smartophones.

The thesis is structured as follows. In the first part of the thesis (Chapter 2-3), we present our work with a college student population. We explore modeling students' on campus behaviors, identifying Dartmouth term behavioral trends, correlations between smartphone data and mental health and academic performance, and predicting depression and GPAs. In the second part of the thesis (Chapter 4-6) we present results from applying sensing technology in a rigorously designed clinical study to a population living with serious mental illness. We believe the studies, the methods, and results presented in this thesis open the way to new forms of mental health sensing, symptom tracking, and real time intervention going forward.

Chapter 2

StudentLife: Using Smartphones to Assess Mental Health and Academic Performance of College Students

2.1 Introduction

We rely on students' self-reports to understand college students' life on campus. However, such self-reports lack details and cannot show the dynamics of an academic term. Consider students at Dartmouth College, an Ivy League college in a small New England college town. Students typically take three classes over a 10-week term and live on campus. The pace of the 10 week Dartmouth term is fast in comparison to a 15 week semester. The atmosphere among the students on campus seems to visibly

change from a relaxed start of term, to an intense midterm and end of term. Typically classes at Dartmouth are small (*e.g.*, 25-50 students), but introductory classes are larger (*e.g.*, 100-170), making it difficult for a faculty to follow the engagement or performance of students on an individual level. Unless students contact a student dean or faculty about problems in their lives, the impact of such challenges on performance remains hidden. To shine a light on student life, we develop the *StudentLife* [196] smartphone app and sensing system to automatically infer human behavior in an energy-efficient manner. The StudentLife app integrates MobileEMA, a flexible ecological momentary assessment [171] (EMA) component to probe students' states (*e.g.*, stress, mood) across the term. We administer a number of well-known pre-post health and behavioral surveys at the start and end of term. In this chapter, we present the results from a deployment of StudentLife on Google Nexus 4 Android phones at Dartmouth College in 2013.

StudentLife is the first to use automatic and continuous smartphone sensing to assess mental health, academic performance and behavioral trends of a student body. We identify strong correlation between automatic sensing data and a broad set of well-known mental well-being measures, specifically, PHQ-9 depression, perceived stress (PSS), flourishing, and loneliness scales. Results indicate that automatically sensed conversation, activity, mobility, and sleep have significant correlations with mental well-being outcomes. we propose for the first time a model that can predict a student's cumulative GPA using automatic behavioral sensing data from smartphones. We use the *Lasso* (Least Absolute Shrinkage and Selection Operator) [184] regularized linear regression model as our predictive model. Our prediction model indicates that students with better grades are more conscientious, study more, experience positive

moods across the term but register a drop in positive affect after the midterm point, experience lower levels of stress as the term progresses, are less social in terms of conversations during the evening period, and experience change in their conversation duration patterns later in the term. We observe trends in the sensing data, termed the *Dartmouth term lifecycle*, where students start the term with high positive affect and conversation levels, low stress, and healthy sleep and daily activity patterns. As the term progresses and the workload increases, stress appreciably rises while activity, sleep, conversation, positive affect, visits to the gym and class attendance drop.

2.2 Related Work

There is a growing interest in using smartphone sensing [142, 20, 43, 37, 188, 12, 55] to infer human dynamics and mental health [11, 155, 116, 82, 62, 99, 156, 130, 133]. The StudentLife study is influenced by a number of important behavioral studies: 1) the friends-and-families study [11], which uses Funf [4] to collect data from 130 adult members (*i.e.*, post-docs, university employees) of a young family community to study fitness intervention and social incentives; and 2) the reality mining project [71], which uses sensor data from mobile phones to study human social behavior in a group of students at MIT. The authors show that call records, cellular-tower IDs, and Bluetooth proximity logs accurately detect social networks and daily activity.

In [155] the authors use an early mobile sensing platform device [55] equipped with multiple embedded sensors to track a 8 older adults living in a continuing care retirement community. The authors demonstrate that speech and conversation occurrences extracted from audio data and physical activity infer mental and social well-being. Researchers at Northwestern University [164] find that mobility and phone usage

features extracted from mobile phone data correlates with depressive symptom severity measured by PHQ-9[174, 122, 121] for 40 participants recruited from the general community over a two week period. Results show features from GPS data, including circadian movement, normalized entropy, location variance, and phone usage features, including usage duration and usage frequency, are associated with depressive symptom severity. It is important to note that the research team were able to reproduce the findings from their initial study [163] using our StudentLife dataset [196]. The replication of the Northwestern University researchers' study results [164] using a different dataset indicates that the mobility features could be broadly applicable in depression sensing across different communities; that is, students on a small college campus and people recruited in a metropolitan area.

There seems to be growing evidence that mobility features have significant signal when it comes to depression sensing. Canzian et al [49] develop an expanded set of mobility features over [164, 163] and found that location data correlates with PHQ-9 [174, 122, 121]. 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 [140] report their preliminary results from a 30 day 25 participant study on the association between human-phone interaction features (e.g., interactions with notifications, number of applications launched) and PHQ-8 scores. Demirci et al.[66] conduct an study with 319 university students to investigate the relationship between smartphone usage and sleep quality, depression, and anxiety. They divide the participants into a smartphone non-user group, a low smartphone use group, and a high smartphone use group based on the Smartphone Addiction Scale (SAS)[125]. The authors find the Beck Depression

Inventory (BDI) [25, 26] score is higher in the high use group than the low use group. This study is solely based on self-report and does not have a sensing component but illustrates that interaction usage could be important in depression sensing.

In [78], the authors present a study of 79 college-age participant from October 2015 to May 2016 in which they find a number of mobility features (e.g., home stay duration, normalized entropy) correlate with PHQ-9. The authors report using SVM with RBF kernel [170], they can predict clinical depression diagnoses with the precision of 84%. Wahle et al. [192] recruit 126 adults to detect depression levels from daily behaviors inferred by phone sensing, in addition to exploring intervention. In the study 36 subjects with an adherence of at least 2 weeks are included in the analysis. The authors compute features based on activity, phone usage, and mobility. They report 61.5% accuracy in predicting a binary depression state. In [13], the authors investigate difference in speech styles, eye activity, and head poses between 30 depressed subjects and 30 non-depressed subjects. The authors report 84% average accuracy in predicting the depression state using a combination of the speech style, eye activity, and head pose features. In [153] Place et al., report on a 12 week study with 73 participants who report at least one symptom of post-traumatic stress disorder (PTSD) [206] or depression. The study assess symptoms of depression and PTSD using features extracted from passive sensing, including the sum of outgoing calls, count of unique numbers texted, absolute distance traveled, dynamic variation of the voice, speaking rate, and voice quality. They report area under the ROC curve (AUC) for depressed mood is 0.74. Chow et al.[56] hypothesizes that time spent at home is associated with depression, social anxiety, state affect, and social isolation. The authors use passive sensing from phones to compute a participant's

time spent at home during a day. The study recruits 72 undergraduates and finds participants with higher depression tended to spend more time at home between the hours of 10 am and 6 pm. The DeepMood project [178] develops a recurrent neural network algorithm to predict depression. The authors conduct a study with 2382 self-declared depressed participants using self-reports to collect self-reported mood, behavioral log, and sleeping log. They report their long short-term memory recurrent neural networks (LSTM-RNNs) [105] model predict depression state with AUC-ROC 0.886.

Detecting bipolar disorder is to some degree related to work on depression sensing. The MONARCA project [150, 149, 96, 162] first reported on findings from mobile sensing and bipolar disorder. The authors [150] 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 [8] show the automatic inference of circadian stability as a measure to support effective bipolar management. Maxuni et al. [138] extend these insights by using speech and activity levels to successfully classify stratified levels of bipolar disorder.

There is a considerable interest in studying the health and performance of students. In [185], the authors study the effect of behaviors (*i.e.*, social support, sleep habits, working hours) on grade points based on 200 randomly chosen students living on the campus at a large private university. However, this study uses retrospective survey data manually entered by users to assess health and performance. Watanabe [201, 202] uses a wearable sensor device to investigate the correlation between face-to-face interaction between students during break times and scholastic performance. Previous research [79] aimed at predicting performance has used a neural

network model to predict student's grades from their placement test scores. Various data collected from entering students are used in [137] to predict student academic success using discriminant function analysis. [119] proposes a regression model to predict the student's performance from their demographic information and tutor's records. [160] applies web usage mining in e-learning systems to predict students' grades in the final exam of a course. In [208], the authors propose an approach based on multiple instance learning to predict student's performance in an e-learning environment. Recent work [180] showed that they can predict a student is at risk of getting poor assessment performance using longitudinal data such as previous test performance and course history.

2.3 Study Design

In this section, we discuss how participants were recruited from the student body, and then describe our data collection process. We also discuss compliance and data quality issues in this longitudinal study.

2.3.1 Participants

All participants in the study were voluntarily recruited from the CS65 Smartphone Programming class [1], a computer science programming class at Dartmouth College offered to both undergraduate and graduate students during Spring term in 2013. This study is approved by the Institutional Review Board at Dartmouth College. 75 students enrolled in the class and 60 participants joined the study. As the term progressed, 7 students dropped out of the study and 5 dropped the class. We remove

this data from the dataset analyzed in the Section 2.6. Among the 48 students who complete the study, 30 are undergraduates and 18 graduate students. The class demographics are as follows: 8 seniors, 14 juniors, 6 sophomores, 2 freshmen, 3 Ph.D students, 1 second-year Masters student, and 13 first-year Masters students. In terms of gender, 10 participants are female and 38 are male. In terms of race, 23 participants are Caucasians, 23 Asians and 2 African-Americans. 48 participants finished the pre psychological surveys and 41 participants finished all post psychological surveys.

All students enrolled in the class were offered unlocked Android Nexus 4s to complete assignments and class projects. Many students in the study had their own iPhones or Android phones. We denote the students who use their own Android phones to run the StudentLife sensing system as *primary users* and those who use the Nexus 4s as *secondary users*. Secondary users have the burden of carrying both their own phones and the Nexus 4s during the study. We discuss compliance and data quality of users in Section 2.3.3.

The StudentLife study has been approved by the Committees for the Protection of Human Subjects at Dartmouth College.

2.3.2 Study Procedure

The StudentLife study consists of orientation, data collection and exit stages. In addition, we deployed a number of management scripts and incentive mechanisms to analyze and boost compliance, respectively.

Entry and Exit. During the orientation stage, participants sign the consent form to join the study. Each student is given a one-on-one tutorial of the StudentLife system and study. Prior to signing the consent form, we detail the type of data to be

collected by the phone. Students are trained to use the app. Students do not need to interact with the background sensing or upload functions. They are shown how to respond to the MobileEMA system. A series of entry health and psychological baseline surveys are administered using SurveyMonkey as discussed in Section 2.6 and shown in Table 2.1. As part of the entry survey students provide demographic and information about their spring term classes. All surveys are administered using SurveyMonkey [7]. These surveys are pre measures which cover various aspects of mental and physical health. Outcomes from surveys (*e.g.*, depression scale) are used as ground truth in the analysis. During the exit stage, we administered an exit survey, interview and the same set of behavioral and health surveys given during the orientation stage as post measures.

Data Collection. The data collection phase lasted for 10 weeks for the complete spring term. After the orientation session, students carried the phones with them throughout the day. Automatic sensing data is collected without any user interaction and uploaded to the cloud when the phone is being recharged and under WiFi. During the collection phase, students were asked to respond to various EMA questions as they use their phones. This in-situ probing of students at multiple times during the day provides additional state information such as stress, mood, happiness, current events, etc. The EMA reports were provided by a medical doctor and a number of psychologists on the research team. The number of EMAs fired each day varied but on average 8 EMAs per day were administered. For example, on days around assignment deadlines, we scheduled multiple stress EMAs. We set up EMA schedules on a week-by-week basis. On some days we administer the same EMA (*e.g.*, PAM and stress) multiple times per day. On average, we administer 3-13 EMA questions

per day (e.g., stress). The specific EMAs are discussed in Section 2.5.

Data Collection Monitoring. StudentLife includes a number of management scripts that automatically produce statistics on compliance. Each time we notice students' phones not uploading daily data (*e.g.*, students left phones in their dorms during the day), or gaps in weekly data (*e.g.*, phones powered down at night), or no response to EMAs, we sent emails to students to get them back on track.

Incentives. To promote compliance and data quality, we offer a number of incentives across the term. First, all students receive a StudentLife T-shirt. Students could win prizes during the study. At the end of week 3, we gave away 5 Jawbone UPs to the 5 top student collectors randomly selected from the top 15 collectors. We repeated this at week 6. We defined the top collectors as those providing the most automatic sensing and EMA data during the specific period. At the end of the study, we gave 10 Google Nexus 4 phones to 10 collectors who were randomly selected from the top 30 collectors over the complete study period.

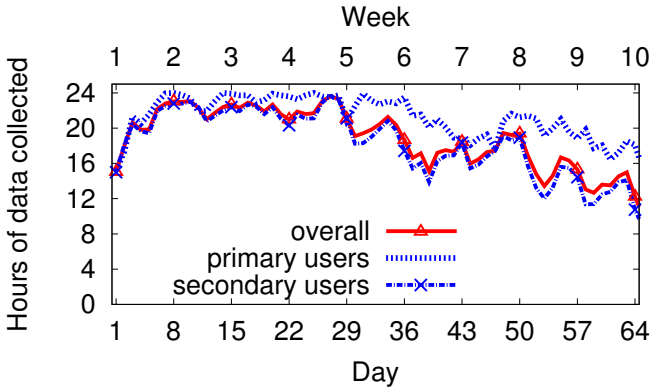
Privacy considerations. Participants' privacy is a major concern of our study. In order to protect participants' personal information, we fully anonymize each participant's identity with a random user id and kept the user id map separate from all other project data so that the data cannot be traced back to individuals. Call logs and SMS logs are one-way hashed so that no one can get phone numbers or messages from the data. Participants' data is uploaded using encrypted SSL connections to ensure that their data cannot be intercepted by third-parties. Data is stored on secured servers. When people left the study their data was removed.

2.3.3 Compliance and Data Quality

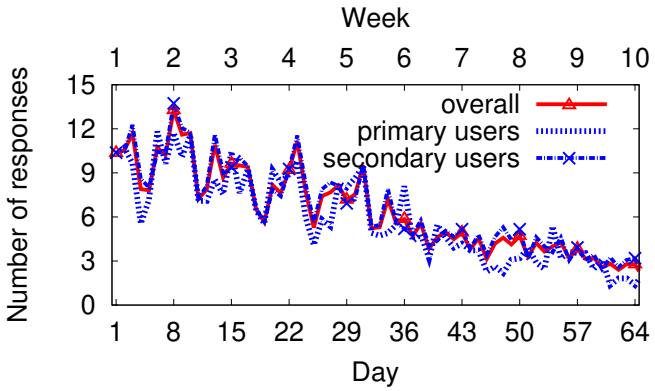
The StudentLife app does not provide students any feedback by design. We do not want to influence student behavior by feedback, rather, we aim to unobtrusively capture student life. Longitudinal studies such as StudentLife suffer from a drop in student engagement and data quality. While automatic sensor data collection does not introduce any burden other than carrying a phone, collecting EMA data can be a considerable burden. Students typically are compliant in responding to survey questions at the start of a study, but as the novelty effect wears off, student compliance drops.

There is a 60/40 split of iPhone/Android users in the study group. Of the 48 students who completed the study, 11 are primary phone users and 37 secondary users. One concern is that the burden of carrying two phones for 10 weeks would result in poorer data quality from the secondary users compared to the primary users. Figure 2.1(a) shows the average hours of sensor data we have collected from each participant during the term. As expected, we observe that primary users are better data sources, but there is no significant difference. We can clearly see the trend of data dropping off as the term winds down. Achieving the best data quality requires 24 hours of continuous sensing each day. This means that users carry their phones and power their phones at night. If we detect that a student leaves their phone at the dorm during the day, or it is powered down, then we remove that data from the dataset. The overall compliance of collecting automatic sensing data from primary and secondary users over the term is 87% and 81%, respectively.

Figure 2.1(b) shows the average number of EMA responses per day for primary and secondary users. The figure does not capture compliance per se, but it shows



(a) Automatic sensing data quality over the term



(b) EMA data quality over the term

Figure 2.1: Compliance and quality of StudentLife data collected across the term.

that secondary users are slightly more responsive to EMAs than primary users. On average we receive 5.8 and 5.4 EMAs per day per student across the whole term from secondary and primary users, respectively. As the term progresses there is a drop in both administered EMAs and responses. However, even at the end of term, we still receive over 2 EMAs per day per student. Surprisingly, secondary users (72%) have better EMA compliance than primary users (65%). During the exit survey, students favored short PAM-style EMAs (see Figure 2.3a), complained about the longer EMAs, and discarded repetitive EMAs as the novelty wore off. By design, there is no notification when an EMA is fired. Participants need to actively check

their phone to answer scheduled EMA questions. The EMA compliance data (see Figure 2.1(b)) shows that there are no significant differences between primary and secondary phone users. It indicates that secondary phone users also used the study phone when they were taking the phone with them. Therefore, the study phone can capture the participants' daily behavior even it was not their primary phone.

In summary, Figure 2.1 shows the cost of collecting continuous and EMA data across a 10-week study. There is a small difference between primary and secondary collectors for continuous sensing and EMA data, but the compliance reported above is promising and gives confidence in the analysis discussed in Section 2.6.

2.4 StudentLife App and Sensing System

In what follows, we describe the design of the StudentLife app and sensing system, as shown in Figure 2.2.

2.4.1 Automatic and Continuous Sensing

We build on our prior work on the BeWell App [127] to provide a framework for automatic sensing in StudentLife. The StudentLife 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 accelerometer, proximity, audio, light sensor readings, location, colocation, and application usage. The inferences and other sensor data are temporarily stored on the phone and are efficiently uploaded to the StudentLife cloud when users recharge their phones under WiFi. In what follows, we discuss the physical activity, sociabil-

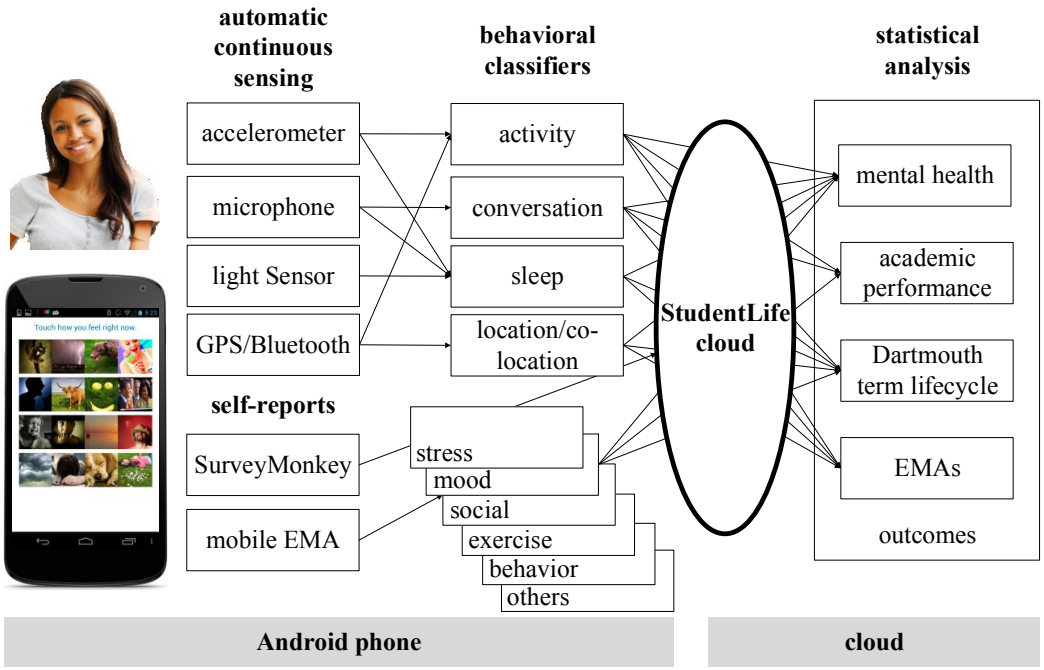


Figure 2.2: StudentLife app, sensing and analytics system architecture.

ity/conversation and sleep inferences computed on the phone which represent important heath well-being indicators [127].

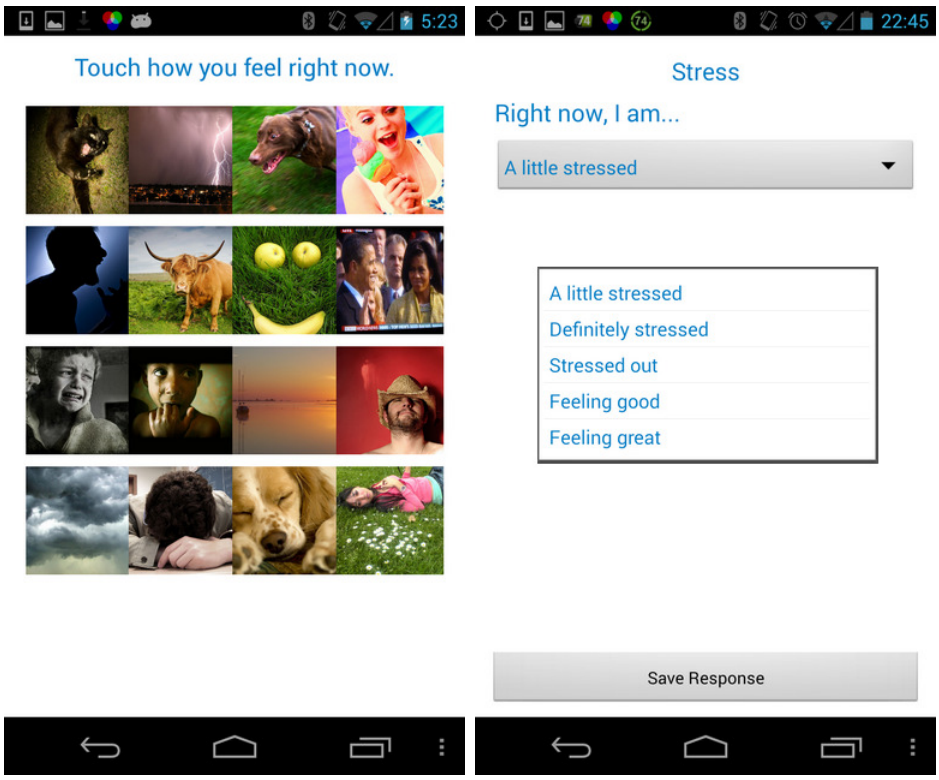
Activity Detection. We use the physical activity classifier from our prior work [127, 131] to infer stationary, walking, running, driving and cycling based on features extracted from accelerometer streams. The activity classifier extracts features from the preprocessed accelerometer stream, then applies a decision tree to infer the activity using the features. The activity classifier achieves overall 94% of accuracy [131]. (Note, we conducted our study before Google announced the availability of an activity recognition service for Android phones). We extend our prior work to compute a daily activity duration, and indoor and outdoor mobility measures, discussed as follows. The activity classifier generates an activity label every 2 seconds. We are only interested in determining whether a participant is moving. For each 10-min

period, we calculate the ratio of non-stationary inferences. If the ratio is greater than a threshold, we consider this period active, meaning that the user is moving. We add up all the 10-min active periods as the daily activity duration. Typically, students leave their dorms in the morning to go to various buildings on campus during the day. Students spend a considerable amount of time in buildings (*e.g.*, cafes, lecture rooms, gym). We consider the overall mobility of a student consists of indoor and outdoor mobility. We compute the outdoor mobility (*aka* traveled distance) as the distance a student travels around campus during the day using periodic GPS samples. Indoor mobility is computed as the distance a student travels inside buildings during the day using WiFi scan logs. Dartmouth College has WiFi coverage across all campus buildings. As part of the study, we collect the locations of all APs in the network, and the Wi-Fi scan logs including all encountered BSSIDs, SSIDs, and their signal strength values. We use the BSSIDs and signal strength to determine if a student is in a specific building. If so, we use the output of activity classifier's walk inference to compute the activity duration as a measure of indoor mobility. Note, that Dartmouth's network operations provided access to a complete AP map of the campus wireless network as part of the IRB.

Conversation Detection. StudentLife implements two classifiers on the phone for audio and speech/conversation detection: an audio classifier to infer human voice, and a conversation classifier to detect conversation. We process audio on the fly to extract and record features. We use the privacy-sensitive audio and conversation classifiers developed in our prior work [155, 127]. Note, the audio classification pipeline never records conversation nor analyses content. We first segment the audio stream into 15-ms frames. The audio classifier then extracts audio features, and uses a two-state

hidden Markov model (HMM) to infer speech segments. Our classifier does not implement speaker identification. It simply infers that the user is "around conversation" using the output of the audio classifier as an input to a conservation classifier. The output of the classification pipeline captures the number of independent conversations and their duration. We consider the frequency and duration of conversations around a participant as a measure of sociability. Because not all conservations are social, such as lectures and x-hours (*i.e.*, class meetings outside lectures), we extend our conservation pipeline in the cloud to remove conversations associated with lectures and x-hours. We use student location to determine if they attend lectures and automatically remove the conservation data correspondingly from the dataset discussed in Section 2.5. We also keep track of class attendance for all students across all classes, as discussed in Section 2.6.

Sleep Detection. We implement a sleep classifier based on our previous work [53, 127]. The phone unobtrusively infers sleep duration without any special interaction with the phone (e.g., the user does not have to sleep with the device). The StudentLife sleep classifier extracts four types of features: light features, phone usage features including the phone lock state, activity features (*e.g.*, stationary), and sound features from the microphone. Any of these features alone is a weak classifier for sleep duration because of the wide variety of phone usage patterns. Our sleep model combines these features to form a more accurate sleep model and predictor. Specifically, the sleep model assumes that sleep duration (Sl) is a linear combination of these four factors: $Sl = \sum_{i=1}^{4} \alpha_i \cdot F_i,\ \alpha_i \geq 0$ where α_i is the weight of the corresponding factor. We train the model using the method described in [53] with an accuracy of +/- 32 mins to the ground truth. We extend this method to identify the sleep onset time by looking at



(a) PAM EMA (b) Stress EMA

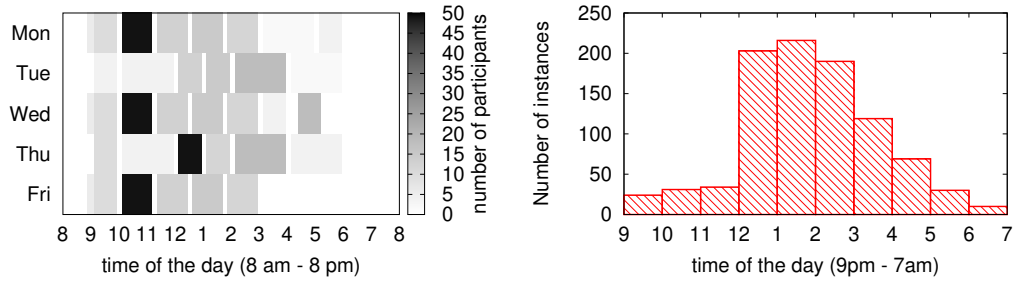
Figure 2.3: MobileEMA: First the PAM popup fires followed by one of the StudentLife EMAs – in this example the single item stress EMA.

when the user is sedentary in term of activity, audio, and phone usage. We compare the inferred sleep onset time from a group of 10 students who use the Jawbone UP during the study to collect sleep data. Our method predicts bedtime where 95% of the inferences have an accuracy of +/- 25 mins of the ground truth. The output of our extended sleep classifier is the onset of sleep (i.e., bedtime), sleep duration and wake up time.

2.4.2 MobileEMA

We use in-situ ecological momentary assessment (EMA) [171] to capture additional human behavior beyond what the surveys and automatic sensing provide. The user is prompted by a short survey (*e.g.*, the single item [182] stress survey as shown in Figure 2.3b) scheduled at some point during their day. We integrate an EMA component into the StudentLife app based on extensions to Google PACO [5]. PACO is an extensible framework for quantified self experiments based on EMA. We extend PACO to incorporate:

- *photographic affect meter (PAM)* [154] to capture participant's instantaneous mood;
- *pop-up EMAs* to automatically present a short survey to the user when they unlock or use the phone; and,
- *EMA schedule and sync* feature to automatically push a new EMA schedule to all participants and synchronize the new schedule with StudentLife cloud.



(a) Meeting time for all classes over the term

(b) Sleep onset time distribution for all students over the term

Figure 2.4: Statistics on class meeting times and sleep onset time (i.e., bedtime).

PACO is a self-contained and complex backend app and service. We extend and

remove features and integrate the EMA component into the StudentLife app and cloud. We set up EMA questions and schedules using the PACO server-side code [5]. The cloud pushes new EMA questions to the phones. The StudentLife app sets up an alarm for each EMA in the list and fires it by pushing it to the users' phone screen as a pop-up. We implement PAM [154] on the Nexus 4 as part of the EMA component. PAM presents the user with a randomized grid of 16 pictures from a library of 48 photos. The user selects the picture that best fits their mood. Figure 2.3(a) shows the PAM pop-up asking the user to select one of the presented pictures. PAM measures affect using a simple visual interface. PAM is well suited to mobile usage because users can quickly click on a picture and move on. Each picture represents a 1-16 score, mapping to the Positive and Negative Affect Schedule (PANAS) [203]. PAM is strongly correlated with PANAS ($r = 0.71, p < 0.001$) for positive affect. StudentLife schedules multiple EMAs per day. We took the novel approach of firing PAM before showing one of the scheduled EMAs (*e.g.*, stress survey). Figure 2.3b shows an EMA test after the PAM pop-up. We are interested in how students' mood changes during the day. By always preceding any EMA with PAM, we guarantee a large amount of affect data during the term.

2.5 StudentLife Dataset

Using the StudentLife system described in Section 2.4, we collect a dataset from all subjects including automatic sensor data, behavioral interferences, and self-reported EMA data. Our ground truth data includes behavioral and mental health outcomes computed from survey instruments detailed in Table 2.1, and academic performance from spring term and cumulative GPA scores provided by the registrar. We discuss

three *epochs* that are evident in the StudentLife dataset. We uses these epochs (*i.e., day* 9am–6pm, *evening* 6pm–12am, *night* 12am–9am) as a means to analyze some of the data, as discussed in Section 2.6. The StudentLife dataset is publicly available [6].

Automatic Sensing Data. We collect a total of 52.6 GB of sensing inference data from smartphones over 10 weeks. The data consist of: 1) activity data, including activity duration (total time duration the user moves per day), indoor mobility and the total traveled distance (*i.e.,* outdoor mobility) per day; 2) conversation data, including conversation duration and frequency per day; 3) sleep data, including sleep duration, sleep onset and waking time; and finally 4) location data, including GPS, inferred buildings when the participant is indoors, and the number of co-located Bluetooth devices.

Epochs. Students engage in different activities during the day and night. As one would expect, sleep and taking classes dominate a student's week. Figure 2.4(a) shows the collective timetable of class meetings for all the classes taken by the students in the study. The darker the slot, the greater proportion of students taking classes in the slot. We can observe that Monday, Wednesday, Friday slots from 10:00-11:05 am and the x-period on Thursday 12:00-12:50 pm are dominant across the week; this is the teaching time for the CS65 Smartphone Programming class which all students in the study are enrolled in. Figure 2.4(a) clearly indicates that the timetable of all classes ranges from 9am to 6pm – we label this as the *day epoch*. Students are not taking classes for the complete period. Many class, social, sports, and other activities take place during the day epoch but class is dominant. The next dominant activity is sleep. Students go to bed at different times. Figure 2.4(b) shows the distribution of bedtime for all students across the term. We see that most students

go to bed between 12am and 4am but the switch from evening to night starts at 12am, as shown in Figure 2.4(b). We label the period between 12am and 9am as the *night epoch*, when most students are working, socializing or sleeping – but sleep is the dominant activity. We consider the remaining period between the end of classes (6pm) and sleep (12am) as the *evening epoch*. We hypothesize that this is the main study and socialization period during weekdays. We define these three epochs as a means to analyze data. We acknowledge that weekdays are different from weekends but consider epochs uniformly across the complete week. We also look for correlations in complete days (e.g., Monday) and across epochs (i.e., Monday day, evening and night).

EMA Data. Students respond to psychological and behavioral EMAs on their smartphones that are scheduled, managed, and synchronized using the MobileEMA component integrated into StudentLife app. We collect a total of 35,295 EMA and PAM responses from 48 students over 10 weeks. EMA and PAM data are automatically uploaded to the cloud when students recharge their phones under WiFi. Students respond to a number of scheduled EMAs including stress (stress EMA), mood (mood EMA), sleep duration (sleep EMA)(which we use to confirm the performance of our sleep classifier), the number of people students encountered per day (social EMA), physical exercise (exercise EMA), time spent on different activities (activity EMA), and short personality item (behavior EMA). All EMAs were either existing validated EMAs (e.g., single item stress measure [182]) found in the literature, or provided by psychologist on the team (e.g., mood EMA).

Survey Instrument Data. Table 2.1 shows the set of surveys for measuring behavioral and mental well-being and personality traits we administer as part of our pre-post

Table 2.1: Mental well-being and personality surveys.

survey	measure
patient health questionnaire (PHQ-9) [122]	depression level
perceived stress scale (PSS)[58]	stress level
flourishing scale [67]	flourishing level
UCLA loneliness scale [161]	loneliness level
big five inventory (BFI) [111]	personality traits

Table 2.2: PHQ-9 depression scale interpretation and pre-post class outcomes.

depression severity	**minimal**	**minor**	**moderate**	**moderately severe**	**severe**
score	1-4	5-9	10-14	15-19	20-27
number of students (pre-survey)	17	15	6	1	1
number of students (post-survey)	19	12	3	2	2

stages, as discussed in Section 2.3. These questionnaires provide an assessment of students' depression, perceived stress, flourishing (i.e., self-perceived success), loneliness, and personality. Students complete surveys using SurveyMonkey [7] one day prior to study commencement, and complete them again one day after the study. Surveys are administered on the phone and stored in the StudentLife cloud (Figure 2.2). In what follows, we overview each instrument. The Patient Health Questionnaire (PHQ-9) [122] is a depression module that scores each of the 9 DSM-IV criteria as 0 (not at all) to 3 (nearly every day). It is validated for use in primary care. Table 2.2 shows the interpretation of the scale and the number of students that fall into each category for pre-post assessment. The perceived stress scale (PSS) [58] measures the degree to which situations in a person's life are stressful. Psychological stress is the extent

to which a person perceives the demands on them exceed their ability to cope [58]. Perceived stress is scored between 0 (least stressed) to 40 (most stressed). The flourishing scale [67] is an 8-item summary measure of a person's self-perceived success in important areas such as relationships, self-esteem, purpose, and optimism. The scale provides a single psychological well-being score. Flourishing is scored between 8 (lowest) to 56 (highest). A high score represents a person with many psychological resources and strengths. The UCLA loneliness (version 3) [161] scale scores between 20 (least lonely) to 80 (most lonely). The loneliness scale is a 20-item scale designed to measure a person's subjective feelings of loneliness as well as feelings of social isolation. Low scores are considered a normal experience of loneliness. Higher scores indicate a person is experiencing severe loneliness. Table 2.3 shows the pre-post measures (i.e., mean and standard deviation) for each scored survey for all students. We discuss these assessments in Section 2.6.

Table 2.3: Statistics of mental well-being surveys.

survey	**pre-study**			**post-study**		
outcomes	participants	mean	std	participants	mean	std
depression	40	5.8	4.9	38	6.3	5.8
flourishing	40	42.6	7.9	37	42.8	8.9
stress	41	18.4	6.8	39	18.9	7.1
loneliness	40	40.5	10.9	37	40.9	10.5

Academic Data. We have access to transcripts from the registrar's office for all participants as a means to evaluate their academic performance. We use spring and cumulative GPA scores as ground truth outcomes. Undergraduates can receive an A–E grade or I (incomplete). Students who get an Incomplete must agree to complete the course by a specific date. GPA ranges from 0 to 4. For the CS65 smartphone programming class we had all the assignment and project deadlines – no midterms or

finals are given in this class. Students provide deadlines of their other classes at the exit interview from their calendars or returned assignments or exams.

2.6 Results

In what follows, we discuss the main results from the StudentLife study. We identify a number of significant correlations between objective sensor data from smartphones and mental well-being. We present results using a subset of the StudentLife dataset to analyze and predict academic performance. We also identify a Dartmouth term lifecycle that captures the impact of the term on behavioral measures representing an aggregate term signature experienced by all students.

2.6.1 Correlation with Mental Health

We first consider correlations between automatic and objective sensing data from smartphones and mental well-being. We also discuss results from correlations between EMA data. Specifically, we report on a number of significant correlations between sensor and EMA data and pre-post survey ground truth outcomes for depression (PHQ-9), flourishing, perceived stress, and loneliness scales, as discussed in Section 2.5 and shown in Table 2.3. We calculate the degree of correlation between sensing/EMA data and outcomes using the Pearson correlation [57] where r ($-1 \leq r \leq 1$) indicates the strength and direction of the correlation, and p the significance of the finding.

PHQ-9 Depression Scale. Table 2.2 shows the pre-post PHQ-9 depression severity for the group of students in the study. The majority of students experience minimal or minor depression for pre-post measures. However, 6 students experience moderate

Table 2.4: Correlations between automatic sensor data and PHQ-9 depression scale.

automatic sensing data	r	p-value
sleep duration (pre)	-0.360	0.025
sleep duration (post)	-0.382	0.020
conversation frequency during day (pre)	-0.403	0.010
conversation frequency during day (post)	-0.387	0.016
conversation frequency during evening (post)	-0.345	0.034
conversation duration during day (post)	-0.328	0.044
number of co-locations (post)	-0.362	0.025

depression and 2 students are moderately severe or severely depressed at the start of term. At the end of term 4 students experience either moderately severe or severely depressed symptoms. We find a number of significant correlations ($p \leq 0.05$) between sleep duration, conversation frequency and duration, colocation (i.e., number of Bluetooth encounters) and PHQ-9 depression, as shown Table 2.4. An inability to sleep is one of the key signs of clinical depression [3]. We find a significant negative correlation between sleep duration and pre ($r = -0.360, p = 0.025$) and post ($r = -0.382, p = 0.020$) depression; that is, students that sleep less are more likely to be depressed. There is a known link between lack of sleep and depression. One of the common signs of depression is insomnia or an inability to sleep [3]. Our findings are inline with these studies on depression [3]. However, we are the first to use automatic sensor data from smartphones to confirm these findings. We also find a significant negative association between conversation frequency during the day epoch and pre ($r = -0.403, p = 0.010$) and post ($r = -0.387, p = 0.016$) depression. This also holds for the evening epoch where we find a strong relationship ($r = -0.345, p = 0.034$) between conversation frequency and depression score. These results indicate that students that have fewer conversational interactions are more likely to be depressed. For

conversation duration, we find a negative association ($r = -0.328, p = 0.044$) during the day epoch with depression. This suggests students who interact less during the day period when they are typically social and studying are more likely to experience depressive symptoms. In addition, students that have fewer co-locations with other people are more likely ($r = -0.362, p = 0.025$) to have a higher PHQ-9 score. Finally, we find a significant positive correlation ($r = 0.412, p = 0.010$) between the validated single item stress EMA [182] and the post PHQ-9 scale. This indicates that people that are stressed are also more likely to experience depressive symptoms, as shown in Table 2.7.

Table 2.5: Correlations between automatic sensor data and flourishing scale.

automatic sensing data	**r**	**p-value**
conversation duration (pre)	0.294	0.066
conversation duration during evening (pre)	0.362	0.022
number of co-locations (post)	0.324	0.050

Flourishing Scale. There are no literal interpretation of flourishing scale, perceived stress scale (PSS) and UCLA loneliness scale instruments, as discussed in Section 2.5. Simply put, however, the higher the score the more flourishing, stressed and lonely a person is. We find a small set of correlations (see Table 2.5) between sensor data and flourishing. Conversation duration has a weak positive association ($r = 0.294, p = 0.066$) during the 24 hour day with flourishing. With regard to conversation during the evening epoch we find a significant positive association ($r = 0.362, p = 0.022$) with flourishing. We also find that students with more co-locations ($r = 0.324, p = 0.050$) are more flourishing. These results suggest that students that are more social and around people are more flourishing. Finally, positive affect computed from the PAM

self-report has significant positive correlation ($r = 0.470, p = 0.002$) with flourishing, as shown in Table 2.7. This is as we would imagine. People who have good positive affect flourish.

Table 2.6: Correlations between automatic sensor data and perceived stress scale (PSS).

automatic sensing data	**r**	**p-value**
conversation duration (post)	-0.357	0.026
conversation frequency (post)	-0.394	0.013
conversation duration during day (post)	-0.401	0.011
conversation frequency during day (pre)	-0.524	0.001
conversation frequency during evening (pre)	-0.386	0.015
sleep duration (pre)	-0.355	0.024

Perceived Stress Scale. Table 2.6 shows the correlations between sensor data and perceived stress scale (PSS). Conversation frequency ($r = -0.394, p = 0.013$) and duration ($r = -0.357, p = 0.026$) show significantly negative correlation with post perceived stress. In addiction, we see more significant negative associations if we just look at the day epoch. Here, conversation frequency ($r = -0.524, p = 0.001$) and duration ($r = -0.401, p = 0.011$) exhibit significant and strong negative correlations with pre and post measure of perceived stress, respectively. This suggests students in the proximity of more frequent and longer conversations during the day epoch are less likely to feel stressed. We cannot distinguish between social and work study conversation, however. We hypothesize that students work collaborative in study groups. And these students make more progress and are less stressed. There is also strong evidence that students that are around more conversations in the evening epoch are less stressed too. Specifically, there is strong negative relationship ($r = -0.386, p = 0.015$) between conversation frequency in the evening epoch and stress.

Table 2.7: Correlations between EMA data and mental well-being outcomes.

mental health outcomes	EMA	r	p-value
flourishing scale (pre)	positive affect	0.470	0.002
loneliness (post)	positive affect	-0.390	0.020
loneliness (post)	stress	0.344	0.037
PHQ-9 (post)	stress	0.412	0.010
perceived stress scale (pre)	positive affect	-0.387	0.012
perceived stress scale (post)	positive affect	-0.373	0.019
perceived stress scale (pre)	stress	0.458	0.003
perceived stress scale (post)	stress	0.412	0.009

Table 2.8: Correlations between automatic sensor data and loneliness scale.

automatic sensing data	r	p-value
activity duration (post)	-0.388	0.018
activity duration for day (post)	-0.326	0.049
activity duration for evening (post)	-0.464	0.004
traveled distance (post)	-0.338	0.044
traveled distance for day (post)	-0.336	0.042
indoor mobility for day (post)	-0.332	0.045

There is also a link between sleep duration and stress. Our results show that there is a strong negative association ($r = -0.355, p = 0.024$) between sleep duration and perceived stress. Students that are getting more sleep experience less stress. Finally, we find significant positive ($r = 0.458, p = 0.003$) and negative correlations ($r = -0.387, p = 0.012$) between self-reported stress levels and positive affect (i.e., PAM), respectively, and the perceived stress scale. There is a strong connection between daily reports of stress over the term and the pre-post perceived stress scale, as shown in Table 2.7. Similarly, students that report higher positive affect tend to be less stressed.

Loneliness Scale. We find a number of links between activity duration, distance travelled, indoor mobility and the loneliness scale, as shown in Table 2.8. All our results relate to correlations with post measures. Activity duration during a 24 hour day has a significant negative association ($r = -0.388, p = 0.018$) with loneliness. We can look at the day and evening epochs and find correlations. There is a negative correlation ($r = -0.464, p = 0.004$) between activity duration in the evening epoch and loneliness. Distance traveled during the complete day ($r = -0.338, p = 0.044$) and the day epoch ($r = -0.336, p = 0.042$) show trends with being lonely. Indoor mobility during the day epoch has strong negative links ($r = -0.332, p = 0.045$) to loneliness. Indoor mobility is a measure of how much a student is moving in buildings during the day epoch. Students that are less active and therefore less mobile are more likely to be lonely. It is difficult to speculate about cause and effect. Maybe these students move around less are more isolated (e.g., stay in their dorm) because they have less opportunity to meet other students outside of class. These students could feel lonely and therefore more resigned not to seek out the company of others. There is also no evidence that people who interact with others regularly do not experience loneliness. This supports our lack of findings between conversation and loneliness. The PAM EMA data (positive affect) has a strong negative association (r=$-0.390, p = 0.020$) with positive affect. In addition, stress self-reports positively correlate ($r = 0.344, p = 0.037$) with loneliness. Students who report higher positive affect and less stress tend to report less loneliness, as shown in Table 2.7.

2.6.2 Predicting Academic Performance

We uses a subset of the StudentLife dataset to analyze and predict academic performance. We only use undergraduate students' (N=30) data because only undergraduates have GPAs. In contrast, Dartmouth graduate students do not have GPAs and only receive High Pass, Pass, Low Pass or No Credit in their classes. We propose new methods to automatically infer *study* (i.e., study duration and focus) and *social* (i.e., partying) *behaviors* using passive sensing from smartphones [198]. We use time series analysis of behavioral states to predict cumulative GPA. We use linear regression with lasso regularization to identify non-redundant predictors among a large number of input features and use these features to predict students' cumulative GPA.

Assessing study and social behavior. The StudentLife dataset provides a number of low-level behaviors (e.g., physical activity, sleep duration, and sociability based on face-to-face conversational data) but offers no higher level data related to study and social behaviors, which are likely to impact academic performance. We attribute meanings or semantics to locations – called behavioral spaces [198] as a basis to better understand study and social behaviors. That is, we extract high level behaviors, such as studying (e.g., study duration and focus) and social (e.g., partying) behaviors by fusing multiple sensor streams with behavioral spaces.

We use behavioral space information to determine study behavior [198]. Each student takes three classes, which are scheduled at specific periods during the week [2]. Students' transcripts indicate what classes they took. The registrar office has the schedule and location for each class. We use location, date (i.e., weekday M-F) and time to automatically determine if a student attends a class or not, checking the dwell time at the location at least equals 90% of the scheduled period (e.g., 110 minutes).

Using this approach the phone can automatically determine the classes a student is taking and their attendance rates.

We heuristically determine if a student's dwell time at a study areas (e.g., library, labs, study rooms, cafes where student primarily work) is at least 20 minutes. We consider periods shorter that 20 minutes are less likely to be real study periods. In addition to dwell time, we use activity and audio attributes to determine a student's level of focus at a study area. The value of activity indicates how often the phone moves – the person is either moving around in the study area or stationary but using the phone. We consider a number of scenarios. If a student is in a study (e.g., a library) and moves around we consider this contributes to a lack of focus. If the phone is mostly stationary in a study area, we consider this contributes to focus. We also use the audio attribute to determine the level of ambient noise in study areas. We consider quiet environments may contribute to study focus and noisy environments do not. In term of focus, a higher activity value indicates that the student moves around less and thus is more focused and a higher audio value indicates that the student is in a quieter environment which is more conducive to being focused. We do not combine these values but use them as independent variables in the analysis section.

We consider behavioral spaces (e.g., Greek houses, dorms) and their attributes to infer if a student is partying [198]. If a student is in a party we assume that they will be moving and around acoustic sound of conversation or music. We also consider the day of the week as being significant for the fraternity and sorority parties (i.e., Wednesday, Friday and Saturday). We discard dwell times under 30 minutes at partying locations.

We partition each Greek house dwell periods (i.e., visit or stay) into 10-minute windows and calculate audio and activity attributes. We hypothesize that the audio and the activity attributes should be significantly different when the student is partying or not partying. We use k-means clustering [205] to find the partying thresholds for both the audio (e.g., music or being surrounded by a large group of people) and activity (e.g., dancing) attributes.

Capturing behavioral change. We extract behavioral change features from the low-level automatic sensing (e.g., sleep duration) and EMA data (e.g., stress) and high-level study and social behaviors discussed in the previous section. We create time series of each behavior for each student. The behavior time series samples each behavior each day. Each time series summarizes a different behavior (e.g., physical activity, conversation frequency and duration, sleep, social behavior, and study behaviors). In order to understand behavior changes across the term, we propose two features [198]: *behavioral slope*, which captures the magnitude of change (e.g., increase or decrease in sleep) over the complete term as well as the first and second half of the term for all students – from the start of term to the midterm point, and then from the midterm point to the end of term; and *behavioral breakpoints*, which capture the specific points in the term where individual behavior change occurs – the number of breakpoints a student experiences indicates the rate of change that occurs. The method to extract these behavioral change features are described in detail in [198].

Predicting cumulative GPA. Predicting GPA is a regression problem; that is, predicting an outcome variable (i.e., GPA) from a set of input predictors (i.e., features). We evaluate various regression models such as regularized linear regression, regression trees, and support vector regression using cross-validation. We select the

Lasso (Least Absolute Shrinkage and Selection Operator) [184] regularized linear regression model as our predictive model. Lasso is a method used in linear regression; that is, Lasso minimizes the sum of squared errors, with a bound on the sum of the absolute values of the coefficients. Considering we have a large number of features, collinearity needs to be addressed. There are two categories of methods that address collinearity: feature selection and feature transformation. Lasso regularization is one of the feature selection methods. *Lasso* solves the following optimization problem:

$$\min_{\beta_0,\beta}(\frac{1}{2N}\sum_{i=1}^{N}(y_i - \beta_0 - x_i^T\beta)^2 + \lambda\sum_{j=1}^{p}|\beta_j|)$$

where N is the number of observations; y_i is the ground truth of observation i; x_i is the p degree feature vector at observation i; λ is a nonnegative regularization parameter, which controls the number of nonzero components of β (i.e., number of the selected features); β_0 is the intercept; and β is the weight vector. The regularization parameter λ is selected using cross-validation. The optimization problem is essentially to minimize the mean square error $\frac{1}{2N}\sum_{i=1}^{N}(y_i - \beta_0 - x_i^T\beta)^2$ of fitting while keeping the model as simple as possible (i.e., select a minimal number of features to avoid overfitting). Thus, *Lasso* automatically selects more relevant features (i.e., predictors) and discards redundant features to avoid overfitting.

We use the mean absolute errors (MAE), the coefficient of determination (R^2) [48], and Pearson correlation to measure the performance of outcome prediction. MAE measures how close predictions are to the outcomes. The mean absolute error is given by MAE $= \frac{1}{n}\sum_{i=1}^{N}|y_i - \beta_0 - x_i^T\beta|$. Smaller MAE is preferred because it indicates that the predictions are closer to the ground truth. R^2 is a another statistic that measures the goodness of fit of a model and indicates how much of the variance our

model explains. R^2 ranges from 0 to 1, where 1 indicates that the model perfectly fits the data. R^2 can be seen to be related to the unexplained variance where $R^2 = 0$ if the feature vector X tells us nothing about the outcome. We use Pearson correlation to measure the linear relations between the ground truth and the predictive outcome.

We apply leave-one-subject-out cross validation [117] to determine the parameters for *Lasso* and the weights for each feature. In order to make the weight regularization work properly, each feature is scaled within the range $[0, 1]$. Selected features have non-zero weights. The MAE of our predicted cumulative GPA is 0.179, indicating that the predictions are within ± 0.179 of the groundtruth. The R^2 is 0.559, which indicates that the features can explain 55.9% of the GPA variance. The predicted GPA strongly correlates with the ground truth with $r = 0.81$ and $p < 0.001$, which further indicates that our predictions can capture outcome differences using the given features.

Table 2.9 shows the selected features to predict the cumulative GPAs and their weights. Interestingly, *lasso* selects a single long term measure (i.e., conscientious personality trait), two self-report time series features (i.e., affect and stress), and three automatic sensing data behaviors (i.e., conversational and study behavior). The weights indicate the strength of the predictors. Students who have better GPAs are more conscientious, study more, experience positive moods (e.g., joy, interest, alertness) across the term but register a drop in positive affect after the midterm point, experience lower levels of stress as the term progresses, are less social in terms of conversations during the evening period between 6-12 pm, and experience later change (i.e., a behavioral breakpoint) in their conversation duration pattern.

Table 2.9: Lasso Selected GPA Predictors and Weights.

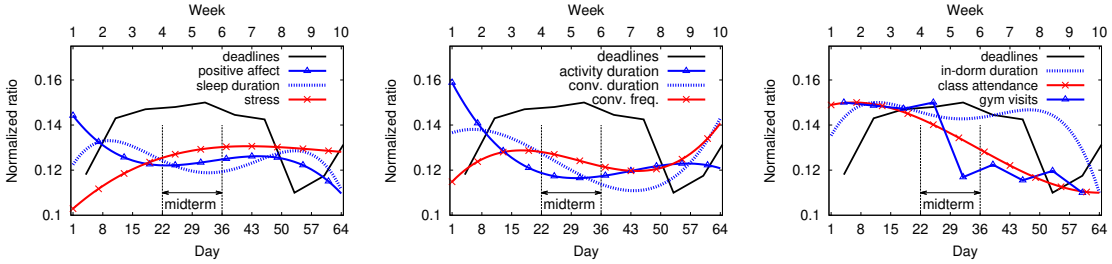
	features	weight
sensing	conversation duration night breakpoint	0.3467
	conversation duration evening term-slope	-0.6100
	study duration	0.0728
EMA	positive affect	0.0930
	positive affect post-slope	-0.1215
	stress term-slope	-2.6832
survey	conscientiousness	0.0449

2.6.3 Dartmouth Term Lifecycle

We analyze the Dartmouth term lifecycle using both sensing data and self-reported EMA data. Figure 2.5(a-c) shows key behavioral measures and activities over the complete term. Figure 2.5(a) shows EMA data for stress and positive affect (PA), and automatic sensing data for sleep duration. Figure 2.5(b) shows continuous sensing trends specifically activity duration, and conversation duration and frequency. Finally, Figure 2.5(c) shows location based data from GPS and WiFi, specifically, attendance across all classes, the amount of time students spent in their dorms or at home, and visits to the gym. We hypothesize that these sensing, EMA and location based curves collectively represent a "Dartmouth term lifecycle". Whether these trends could be observed across a different set of students at Dartmouth or more interestingly at a different institution is future work. In what follow we discuss workload across the term, mental well-being using EMA data (i.e., stress and positive affect) and automatic sensing data measures.

Academic Workload. We use the number of assignment deadlines as a measure of the academic workload of students. We collect class deadlines during exit interviews

and validate them against students' calendars and returned assignments dates. Figure 2.5 shows the average number of deadlines for all student across each week of the term. The number of deadlines peaks during the mid-term period in weeks 4 and 5. Interestingly, many classes taken by the students do not have assignment deadlines during week 8. Final projects and assignments are due in the last week of term before finals, as shown in Figure 2.5(a). As discussed before, all study participants take the same CS65 Smartphone Programming class, for which they share the same deadlines. Among all CS65's lab assignment, Lab 4 is considered to be the most challenging programming assignment. In the last week of term the students need to give final presentations and live demos of group projects for the smartphone programming class. The students are told that app developed for the demo day has to work to be graded. The demo is worth 30% of their overall grade.



(a) EMA and sleep data (b) Automatic sensing data (c) Location-based data

Figure 2.5: Dartmouth term lifecycle: collective behavioral trends for all students over the term.

Self Reported Stress and Mood. Figure 2.5(a) shows the average daily stress level and positive affect over the term for all subjects as polynomial curves. Students are more stressed during the mid-term (days 22-36) and finals periods. The positive affect results show a similar trend. Students start the term with high positive affect, which then gradually drops as the term progresses. During the last week of term,

students may be stressed because of finals and class projects, with positive affect dropping to its lowest point in the term. Overall, the results indicate that the 10-week term is stressful for students as workload increases. Figure 2.5(a) clearly shows that students return to Dartmouth after spring break feeling the most positive about themselves, the least stressed, the most social in terms of conversation duration and the most active (as shown in Figure 2.5(b)). As the term progresses toward mid-term week, positive affect and activity duration plunge and remain at low levels until the final weeks where positive affect drops to its lowest point.

Automatic Sensing Data. We also study behavioral patterns over the term by analyzing automatic sensing data. We plot the polynomial fitting curves for sleep duration, activity duration, conversation duration, conversation frequency, as shown Figure 2.5(b), and location visiting patterns in Figure 2.5(c). Our key findings are as follows. We observe from Figure 2.5(a) that sleep peaks at the end of the first week and then drops off and is at its lowest during the mid-term exam weeks. Sleep then improves until the last week of term when it plummets to its lowest point in the cycle. As shown in Figure 2.5(b) students start the term with larger activity duration, which gradually drops as they become busier with course work and other term activities. Finally, the activity duration increases a little toward the end of term. Activity duration reaches its lowest point on day 36 when students are focused on completing the Lab 4 assignment – considered the most demanding assignment in the smartphone programming class.

The student's level of face-to-face sociability starts high at the start of term, then we observe an interesting conservation pattern, as shown in Figure 2.5(b). As the term intensifies, conversation duration drops almost linearly until week 8, and

then rebounds to its highest point at the end of term. Conversely, the frequency of conservation increases from the start of term until the start of midterms, and then it drops and recovers toward the end of term. We speculate that sociability changes from long social/study related interactions at the start of term to more business-like interactions during midterms when students have shorter conservations. At the end of term, students are having more frequent, longer conversations.

Figure 2.5(c) provides a number of interesting insights based on location based data. As the term progresses and deadlines mount the time students spend at the significant places in their lives radically changes. Visits to the gym plummet during midterm and never rebound. The time students spend in their dorm is low at the start of term perhaps due to socializing then remains stable but drops during midterm. At week 8 time spent in dorms drops off and remains low until the end of term. The most interesting curve is class attendance. We use location data to determine if students attend classes. We consider 100% attendance when all students attend all classes and x-hours (if they exist). The term starts with 75% attendances and starts dropping at week 3. It steadily declines to a point at the end of term were only 25% of the class are attending all their classes. Interestingly, we find no correlation between class attendance and academic performance. We speculate that students increasingly start missing classes as the term progresses and the work load rises. However, absence does not positively or negatively impact their grades. We put this down to their self learning ability but plan to study this further as part of future work.

It is difficult in this study to be concrete about the cause and effect of this lifecycle. For example, stress or positive affect could have nothing to do with workload and everything to do with hardship of some sort (*e.g.*, campus adjustment, roommate

conflicts, health issues). We speculate the intensive workload compressed into a 10 week term puts considerable demands on students. Those that excel academically develop skills to effectively manage workload, social life and stress levels.

2.7 Discussion

2.7.1 Compliance

Participants report feeling boredom with EMAs questions during the exit interview. As a result, EMA response rate keeps dropping as shown in Figure 2.1. We get most EMA responses from the simplest EMA questions. Stress EMA is a single question EMA pops up multiple times a day, but students are willing to answer them. Students like PAM because it is simple and fun. In contrast, we only get a few responses from more complicated Behavior EMA, which has over 5 questions. Therefore, in future study, the EMA design should be as simple as possible and require little user effort.

The biggest challenge for continuous sensing is energy consumption. Although the battery life with continuous sensing is around 14 to 16 hours according to our tests, a few students report that they have to charge their phone in the middle of day. This is because these students are heavy phone users.

2.7.2 Academic Performance

Intuitively, class attendance should be positively correlate to academic performance. However, we do not find such correlation. After inspecting the data, we find that for students whose Sprint term GPA is greater than 3.7, 7 students' attendance rates are greater than 60%, while 5 students' attendance rates are below 40%. Our observation

is that for some high performers, they do not need to attend classes to excel in classes.

2.7.3 Analyzing multiple factors

We only analyze linear correlation between single behavior with multiple mental health or academic performance outcomes. However, one specific outcome might be influenced by multiple factors. For example, computer major students may need less time than non-computer science major students to understand the content of a computer science class. Therefore, their effort alone might not be sufficient to predict performance. We need other data modeling technique to learn how different factors act together to influence one particular outcome. Also, this data modeling technique should be transparent so that we can not only predict the outcome, but also learn the relations between the multiple factors and the outcome.

2.7.4 Extracting high level activities

One challenge of applying smartphone sensing technique to study students' life (or other social group's life) is to identify activities beyond simple physical activity and conversation. For example, in order to understand a student's life, we would like to learn their class attendances, when and where they eat, how often they go to party etc. None of these behaviors can be inferred from single sensor. However, by combining different sensor streams and necessary context information, we can infer these high level activities. For example, we can infer if a student attended a class by looking at their location history, class schedule and class location.

2.8 Conclusion

In this chapter, we presented the StudentLife sensing system and results from a 10-week deployment. We discussed a number of insights into behavioral trends, and importantly, correlations between objective sensor data from smartphones and mental well-being and predicting undergraduate students' cumulative GPA for a set of students at Dartmouth College. Our results showed that it is promising to use the StudentLife sensing system to collect behavioral data and in-situ self-report EMAs from college students and use the data to infer college students' mental health and academic performance.

There are a number of limitations in this work. First, the number of participants is small and all participants were recruited from the same Computer Science class, therefore, the participants cannot represent the wider Dartmouth student population. Second, many of our participants used the study phones as their secondary phones. It was burdensome for them to carry two phones all the time and not practical for large scale studies. The StudentLife sensing app only works on Android phones, whereas many students use iPhones. Lastly, the StudentLife study is not entirely focused on predicting students' mental health outcomes. We aim to tackle these limitations in Chapter 3. We will discuss applying the smartphone sensing technology in a more challenging environment (i.e., people with serious mental illness) in Chapter 4-6.

Chapter 3

Tracking Depression Dynamics in College Students Using Mobile Phone and Wearable Sensing

3.1 Introduction

The StudentLife study described in Chapter 2 has shown great potentials in use smartphones to assess college students' mental health and academic performance. We have found many correlations between the behavioral features derived from the smartphone passive sensing data and a number of well-known pre-post health and behavioral surveys (e.g., PHQ-9, Perceived Stress Scale). In this chapter, we focus on using smartphone data to understand the connections between students' behaviors and depression and predict students' depression states. To address the limitations described in Chapter 2.8, we improve the StudentLife sensing app to support iPhones and wearables (i.e., Microsoft Band 2). We recruit more participants (i.e., 83 students)

from a wider student population over two academic terms in 2016. To ensure study compliance and data quality, participants install the updated StudentLife app on their own phones. We take a deep dive into the connections between human behaviors and depression symptoms defined in the standard mental disorders diagnostic manual (DSM-5 [18]) and propose symptom features that are predictive of depression in college students.

The standard mental disorders diagnostic manual (DSM-5 [18]) defines 9 common symptoms associated with major depression disorders: depressed mood, sleep changes, weight change, fatigue or loss of energy, restlessness or feeling slow, diminished interest or pleasure in activities, diminished ability to concentrate, feelings of worthlessness, and thoughts of death and suicide. Existing work [49, 19, 196, 126, 164, 163] have found relationships between depression and generic behavioral features from passive sensing. However, they do not discuss how these behavioral features are associated with the well-defined depression symptoms. In this chapter, we take a different approach and propose a set of depression sensing symptom features (called *symptom features* for short) derived from phone and wearable passive sensor data that represent proxies for the DSM-5 depression symptoms in college students; that is, we design a set of behavioral features to capture the characteristics of the depression symptoms that take into account lifestyles of students (e.g., going to class, working in study areas, socializing on campus). Specifically, we hypothesize: (1) the *sleep change symptom* can be measured by sleep duration, start time, and end time inferred from passive sensor data from phones [54, 196, 195]; (2) the *diminished ability to concentrate symptom* can be associated with excessive smartphone use [66, 125], specifically when measured in study spaces across campus (e.g., libraries,

study rooms, quiet working areas associated with cafes, dorms, etc.) where students typically focus on their course work; (3) the *loss of interest or pleasure in activities symptom* may cause changes in activity and social engagement patterns [18], thus can be associated with changes in activity, conversation, and mobility patterns inferred from mobile sensing data; and finally (4) the *depressed mood symptom* and *fatigue or loss of energy symptom* relate to changes in physiology, thus, may be associated with changes in heart rate data passively measured by wearables (e.g., prior work found heart rate data is associated with depressed mood [109, 113] and fatigue [169]). To test our hypothesis, we conduct a study of 83 undergraduate students at Dartmouth College across two 9-week terms during the winter and spring terms in 2016. The study has been approved by the Committees for the Protection of Human Subjects at Dartmouth College. The research is supported the National Institute of Mental Health, grant number 5R01MH059282-12. Each student installed an updated StudentLife app on their own Android or Apple phones and were given a Microsoft Band 2 [141] for the duration of a 9-week term. The updated StudentLife app continuously collects behavioral passive sensing data from smartphones and physiological sensing data from Microsoft Band 2 [141]. We compute the symptom features from the passive sensing data and conduct correlation analysis between the symptom features and PHQ-8/PHQ-4 depression outcomes. We look at the correlations between the symptom features and the PHQ-8 [124] and PHQ-4 [123] depression groundtruth. We further look to predict PHQ-4 depression subscale states (i.e., *non depressed* and *depressed*) [123] using the proposed symptom features.

The contributions of this chapter are: (i) we propose a set of passive sensor based symptom features derived from phones and wearables that we hypothesize proxy 5 out

of the 9 major depressive disorder symptoms defined in DSM-5; (ii) we find a number of correlations between the proposed symptom features and PHQ-8 item scores. The findings evaluate the efficacy of the symptom features to capture depression symptoms; (iii) we identify a number of correlations between the symptom features and PHQ-8; (iv) we use ANOVA to compare the means of the symptom features between the *non depressed group* and the *depressed group*, as defined in PHQ-8 [124]. We show that these two groups are clearly identified in our data set; and (v) we show that we can predict PHQ-4 and PHQ-8 using the proposed symptom features.

3.2 Depression Sensing using Symptom Features

How we assess depression has not changed in 30 years. Mental health specialists rely on depression screening tools (e.g., Patient Health Questionnaire (PHQ) [174, 122, 121], Major Depression Inventory (MDI) [24], Beck Depression Inventory (BDI) [25], Hamilton Depression Rating Scale (HDRS) [92]) and clinical interviews to assess the severity of the symptoms of major depressive disorder as defined in the DSM-5 [18]. These questionnaires rate the frequency of the symptoms that factor into scoring a severity index called a "depression score". These tools, however, rely on periodic subjective self-reports. A person is said to suffer from a depressive episode if they experiencing at least 5 of the 9 depression symptoms during the same 2-week period, including either depressed mood or loss of interest or pleasure in activities, the symptoms cause clinically significant distress or impairment in social, occupational, or other important areas of functioning [18], and the symptoms are not attributable to other medical conditions [18]. Existing work on predicting depression using mobile phone sensing focuses on using generic behavioral features to predict or correlate the

depression scale. *We take a different approach and hypothesize that mobile sensor data from phones and wearables represent proxies for the DSM-5 depression symptoms in college students; that is, we design a set of behavioral features to capture the characteristics of the depression symptoms that take into account lifestyles of students (e.g., going to class, working in study areas, socializing on campus).* In what follows, we describe the depression symptom features that represent 5 of the 9 major depressive disorder symptoms. The symptom feature implementation details are described in Section 3.4.1.

Sleep changes may result in students experiencing difficulty sleeping (insomnia) or sleeping too much (hypersomnia) and changes in sleep schedules. Many times because of the demands of the term (e.g., assignment due dates, exams, social life, sports, etc.) students experience changes in the regular sleep patterns as demands on the term increase. We infer students' sleep time, wake up time, and sleep duration using passive sensing from phones [54]. We use phone sensing to determine if depressed students sleep more or less than non-depressed students and if depressed students have more irregular sleep schedules; that is, more variation in the time they go to bed and wake up. As a result we can accurately infer the sleep changes symptom.

Diminished ability to concentrate may cause students to appear distracted, unfocused, and unable to perform well. We use phone usage to measure if a student is more likely to be distracted. Previous work [66] shows that smartphone overuse may lead to increased risk of depression and/or anxiety. Specifically, we measure the number of unlocks of the phone and associated usage duration across the day and at specific locations; for example, dorm, in study areas and in the classroom, etc. When a student is at the classroom or study areas, they are supposed to focus on

work at hand or studying. In such locations we assume the more phone usage (i.e., spend more time on their phones) may indicate they are having difficulty focusing on their work. We hypothesize that phone usage in the classroom and study places is a potential indicator of a student's diminished ability to concentrate in comparison to regular phone use in social spaces, dorm, gym, or walking around campus. By using phone location data and contextual labeling of the campus [197] we can differentiate these different use cases accurately.

Diminished interest or pleasure in activities may cause changes in students' activity and social engagement patterns that are not easily explained by external forces like deadlines or exams. We can observe changes in a student's physical activity (i.e., more/less active) and mobility (e.g., visit more/fewer places on campus). Specifically, we look at the time spent at different types of places. On campus buildings and spaces in buildings are usually associated with a primary function (e.g., study area, classroom, library, gym, social, cafes, etc). Most Dartmouth undergraduate students study in a large number of shared study spaces across campus including libraries and typically dormitory buildings are used to rest, sleep and socialize and rarely used to study. We compute time spent in all areas but specifically in study places and dorms; specifically, we compute a number of behavioral features including dwell time at locations, phone usage (unlock frequency and duration), and the number and duration of conversation students are around in these spaces. All the location based behavioral features are normalized by the dwelling duration. We use the fusion of these features to proxy a student's diminished interest or pleasure in activities.

Depressed mood and **Fatigue or loss of energy** are difficult to detect from passive behavioral sensing from phones. Instead we consider heart rate from wear-

ables. Previous work has found that heart rate data is associated with depressed mood [109, 113] and fatigue [169]. We determine if depressed students' heart rate is different from non-depressed students. We consider this might be a potential signal and proxy for depressed mood or fatigue or loss of energy. In addition, we also determine if more depressed students visit on-campus health facilities more.

3.3 Data Collection

We collect a smartphone and wearable sensing dataset from 83 Dartmouth College undergraduate students across two 9-week terms during winter (56 students) and spring (27 students) terms in 2016. The average age of the participants is 20.13 (std=2.31) and 40 are male and 43 are female (26 are Asian, 5 are African American, 24 are Caucasian, 1 is multiracial, and 26 not specified). This study is approved by the Institutional Review Board at Dartmouth College. Figure 3.1 shows the sensing system, symptom feature mappings, and the depression ground truth. In what follows, we discuss the sensing system, the study design, and the dataset.

3.3.1 Mobile Sensing System for Phones and Wearables

We update the StudentLife sensing app described in Chapter 2 to support iOS. We replace our in-house activity classifier with the build-in Android and iOS activity recognition API. The updated StudentLife app continuously infers and record participants' physical activities (e.g., stationary, in a vehicle, walking, running, cycling), sleep (duration, bed time, and rise time) based on our prior work on sleep inference using phones [54], and sociability (i.e., the number of independent conversations a

3.3 Data Collection

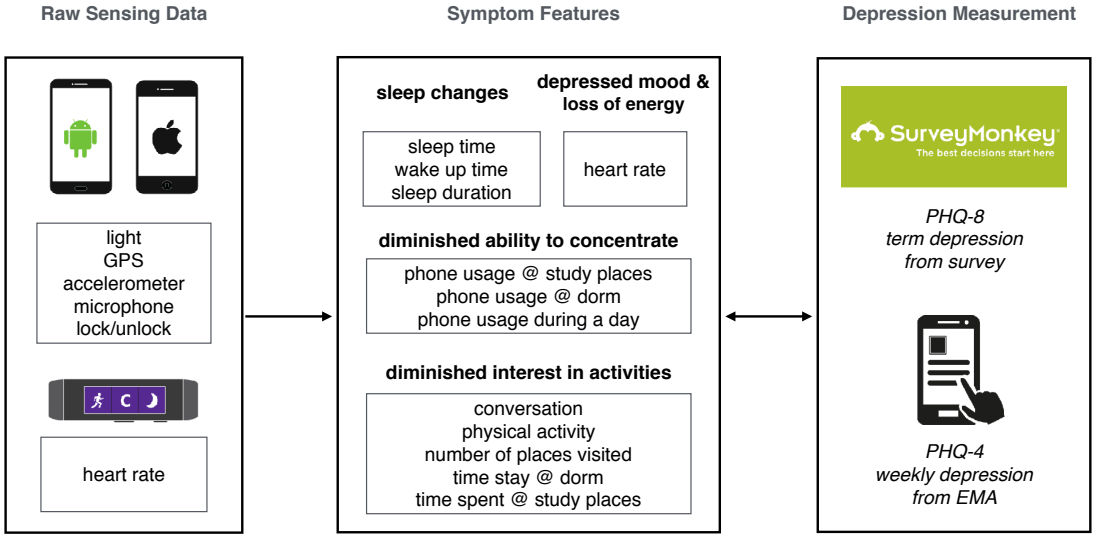


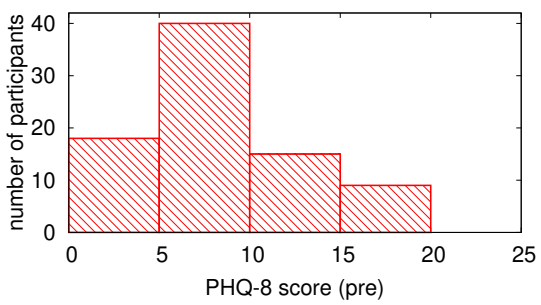
Figure 3.1: We continuously collect behavioral passive sensing data from Android and Apple iOS smartphones and physiological sensing data from Microsoft Band 2. We compute the symptom features from the passive sensing data. The symptom features map smartphone and wearable passive sensing to 5 depression symptoms defined in DSM-5: sleep changes, diminished ability to concentrate, diminished interest in activities, depressed mood, and fatigue or loss of energy. We look for associations between the symptom features and the PHQ8/PHQ4 depression outcomes.

participant is around and their duration). The app also collects audio amplitude, location coordinates, and phone lock/unlock events. A built-in MobileEMA component is used to administer self-reported PHQ-4 [123] EMAs. The app uploads the data to the secured server when the user is charging their phones and under WiFi. StudentLife is extended to collect data from wearables; specifically, we collect physiological data from Microsoft Band 2 [141] given to each of the students in our study. The StudentLife app collects the heart rate, galvanic skin response (GSR), skin temperature, and activity data from the band in real time. Band data is uploaded to the StudentLife app over Bluetooth and then uploaded to our servers as described

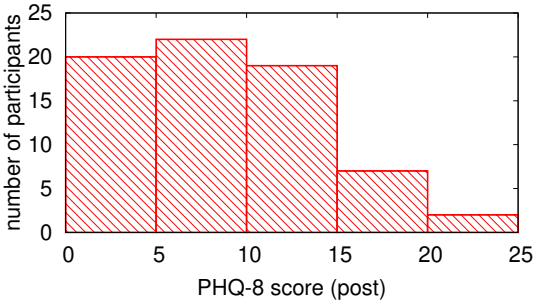
above. Note, during data modeling and analysis we found poor data quality issues associated with GSR and skin temperature data from the band. First, the GSR sample rate provided by the Microsoft Band SDK [141] is too low (0.2 HZ) to be useful in analysis. Such a low sample rate limited extracting useful GSR features. We also found that the skin temperature sensor reading is affected by the ambient environment temperature. The temperature differences between indoor and outdoor during New Hampshire winter can be as large as 70 degree Fahrenheit. We observed significant drops in skin temperature when participants are outside during the winter term. For these reasons we collected but did not uses GSR and skin temperature data in our modeling. While the StudentLife app infers sleep data from the phone only within +/- 25 mins of error [54] the band has much better sleep measurements. However, because the band only lasted one day due to limited battery and the demands of continuous sensing most students wore the band during the day and recharged it at night. The result is that we have limited sleep data from the band. We therefore only use sleep measurements from our phone data. Even though we collect GSR, sleep, skin temperature we end up only using heart rate and activity data from the band.

3.3.2 Depression Groundtruth

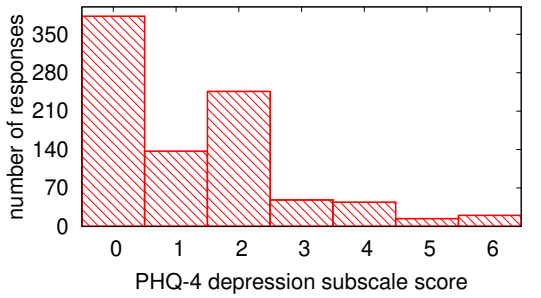
We use the self-reported PHQ-8 [124] and PHQ-4 [123] as groundtruth for depression outcomes in our study. This is a widely used measure with excellent validity. PHQ-8 is administered at the beginning and the end of the study period as a pre-post depression measures. PHQ-4 is administered once a week and used to capture depression dynamics across the term. The PHQ-8 scores 8 of the 9 major depressive disorder symptoms over the past two weeks where each item (i.e., question) is scored by the



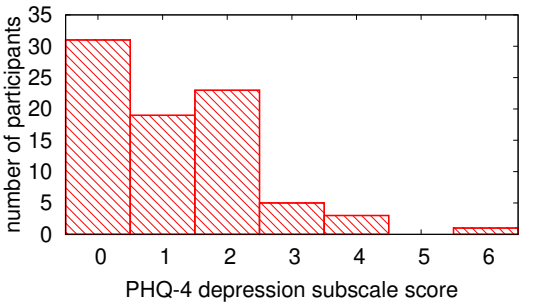
(a) The distribution of pre PHQ-8 scores



(b) The distribution of post PHQ-8 scores



(c) The distribution of PHQ-4 depression subscale scores



(d) The distribution of mean PHQ-4 depression subscale scores

Figure 3.2: The distribution of the PHQ-8 and PHQ-4 responses. (a) The mean score for the pre PHQ-8 is 6.09 ($N = 82, \text{std} = 4.33$), where 16 students are in the depressed group (PHQ-8 $\geq$ 10). (b) The mean score for the post PHQ-8 is 6.69 ($N = 71, \text{std} = 5.46$), where 17 students are in the depressed group. (c) The mean score of the PHQ-4 depression subscale is 1.34 ($N = 707, \text{std} = 1.50$), where 108 responses are above the depressed cutoff (≥ 3). (d) The mean per-participant PHQ-4 depression subscale score is 1.31 (std = 1.17), where 4 participants' mean PHQ-4 depression subscale score is above the depressed cutoff (≥ 3).

user from 0 (not at all) to 3 (nearly every day). The PHQ-8 does not score the *thoughts of death and suicide symptom* nor do we consider this in our study. The resulting PHQ-8 depression score ranges from 0 to 24 indicating five levels of depression: (1) none to minimal depression (range 0-4); (2) mild depression (range 5-9); (3) moderate depression (range 10-14); (4) moderately severe depression (range 15-19); and finally (5) severe depression (20 to 24). The score can also be interpreted as *no depression*

(range 0-9) and *current depression (range 10-24)* according to [124]. Figure 3.2(a-b) shows the distribution of the pre-post PHQ-8 responses. The mean score for the pre PHQ-8 is 6.09 (std = 4.33), where 16 out of 82 students are in the *depressed group* (PHQ-8 $\geq$ 10). The mean score for the post PHQ-8 is 6.69 (std = 5.46), where 17 out of 71 students are in the depressed group. We receive fewer post PHQ-8 surveys because some participants do not complete the survey.

The PHQ-4 is an ultra-brief tool for screening both anxiety and depression disorders over the past two weeks. It uses a depression subscale to score depression and an anxiety subscale to score anxiety. We only consider the depression subscale in our study on depression. The depression subscale comprises 2 questions from the PHQ-4 that score the *diminished interest or pleasure in activities symptom* and the *depressed mood symptom*. The score of the depression subscale ranges from 0-6, where a score of 3 or greater is considered depressed according to [123]. We collected in total 707 PHQ-4 responses across the terms. Figure 3.2(c) shows the distribution of the PHQ-4 depression subscale scores from all responses. The mean score of the PHQ-4 depression subscale is 1.34 (std = 1.50), where 108 responses are above the depressed cutoff ($\geq$ 3). Figure 3.2(d) shows the distribution of each student's PHQ-4 depression subscale. The mean per-student PHQ-4 depression subscale score is 1.27 (std = 1.15), where 5 students' mean PHQ-4 depression subscale score is above the depressed cutoff ($\geq$ 3).

3.4 Methods

In what follows, we present our symptom features and methods to evaluate the efficacy of using the symptom features to predict depression severity in college students.

3.4.1 Symptom Features

We present our symptom features listed in Table 3.1 that capture the 5 out of 9 major depressive symptoms based on data collected using phone and wearable passive mobile sensing data.

Table 3.1: Depression symptom features.

DSM symptom	Symptom features
sleep changes	sleep duration sleep start sleep end
diminished ability to concentrate	unlock duration unlock duration at dorm unlock duration at study places
diminished interest or pleasure in activities	stationary time conversation duration number of places visited time at dorm time at study places
depressed mood and fatigue or loss of energy	heart rate

Sleep. We compute three sleep features to measure the *sleep change symptom*: sleep duration, sleep onset time, and wake time. We assume students experiencing this symptom may sleep significantly more or less than normal, or experience irregular sleep schedules (i.e., more variations in sleep onset time or wake time). The sleep inferences are based on four phone sensors: ambient light, audio amplitude, activity, and screen on/off [54]. The sleep classifier does not infer naps. It simply computes the longest period of inferred sleep. The sleep classifier approximates sleep duration within +/- 30 minutes and has been used in a number of other studies [8, 196, 195].

Physical activity. Students who experience the *diminished interest or pleasure in activities symptom* may change their activity pattern (e.g., being less mobile and more stationary) [18]. We compute the stationary duration during a day to measure students' sedentary levels. The app continuously infers physical activities using the Android activity recognition API [89, 195] and iOS Core Motion [15]. Activity recognition infers whether a user is on foot, stationary, in vehicle, on bicycle, tilting, or doing an unknown activity. We compute the non-physically active duration (i.e., the stationary duration) using the still label from the classifier. Both Android and iOS activity recognition API detects the stationary state with high accuracy.

Speech and conversational interaction. Students who experience the *diminished interest or pleasure in activities symptom* may experience social withdrawal and change their social engagement patterns [18]. We compute the number of independent conversations and their duration everyday as a proxy for social interaction. The StudentLife app infers the amount of speech and conversation a participant is around [195, 196]. In [196, 195] we discuss the detailed in design of the conversation 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. The conversation classifier does not identify speakers. Therefore, we do not know that if the participant is actively involved in the conservation or not (e.g., they could be sitting at a table in a cafe where others around them are speaking). However, we have validated this in a number of studies and it is capable of capturing levels of social interaction.

Location and mobility. There is evidence that people's mobility patterns are related to depression. We use mobility and location features as a proxy for the *diminished interest or pleasure in activities symptom*. Prior work indicates [78, 49] that people with this symptom avoid leaving their homes. We compute students' distance traveled, the number of places visited and time spend at dorms and study areas across campus based on location data [49, 196, 164] and a semantic understanding of locations across Dartmouth campus. We use DBSCAN [136] to cluster GPS coordinates collected during the day to find significant places that students dwell at. The DBSCAN algorithm groups location/GPS coordinates that are close together as a significant place where students visit. We compute the number of places visited as a feature. We label every on-campus building and spaces in buildings (e.g., classrooms, study area, dorms, libraries, cafes, social spaces, gyms, etc.) according to their primary function; for example, we label each student's dorm as the place where they dwell between 2-6 am. In addition, we also determine the number of times a student visits the on-campus health center as a contextual mobility feature.

Phone usage. We use phone usage to measure the *diminished ability to concentrate symptom*. Phone overuse has been linked to depression in college students [66, 125]. We compute the number of phone lock/unlock events and the duration that the phone is unlocked during a day, when a student is at their dorms and study areas. To avoid the impact of stay duration on the location based phone usage features (e.g., a student tends to record higher phone usage when they stay at a place longer), we normalize the phone usage features for location based usage data by duration of their stay. Excessive smartphone usage at study places or in the classroom may indicate students are experiencing difficulty in concentrating on the work at hand.

Heart rate. We use students' physiological signals to detect if they are experiencing *depressed mood* or *fatigue or loss of energy symptoms*. Previous work has found that heart rate variability is associated with depressed mood [109, 113] and fatigue [169]. The average of beat-to-beat or NN intervals (AVNN) is one of the heart rate variability measures [134]. The heart rate (HR) is the inverse of the AVNN in milliseconds: HR = 60000/AVNN. The StudentLife app collects heart rate data from Microsoft Band 2 in real time. The accuracy of wrist heart rate monitors depends on many factors. The heart rate measured by Microsoft Band 2 is accurate when the user wears the band correctly (i.e., not too loose nor too tight) and is relatively stationary during measurement periods. Based on our testing we found that the heart rate error is within 2 beats for the band. However, the accuracy suffers if moving their arms because of motion artifacts in the signal. In order to get an accurate measure of daily heart rate, we compute the median heart rate during each day. The median heart rate is a more robust measure of daily heart rate than the mean, maximum, and minimum heart rate because median heart rate is less likely to be influenced by outliers.

3.4.2 Feature Set Construction

We construct a PHQ-8 dataset and a PHQ-4 dataset to look at association between the symptom features and the depression groundtruth.

The PHQ-8 dataset uses students' pre-post PHQ-8 responses as the ground truth. We look to find correlations between the PHQ-8 scores and the symptom features. In addition to the PHQ-8 scores, we include the PHQ-8 depression group assignment (i.e., *non depressed group (range 0-9)* and *depressed group (range 10-24)* as defined in [124]) as students' binary pre-post depression state. We compute the term mean,

standard deviation, and the slope of each symptom features described in Section 3.4.1. The term mean features describes the average of the daily symptom features over the 9-week term. For example, the mean time spend at study places is the average time a student spends at study places every day during the term. The term standard deviation describes the variations in the daily symptom features. For example, a higher standard deviation in sleep start time and end time indicates that the student has an irregular sleep schedule. The term slope features describe how the daily symptom features change over the term. We fit the daily 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 change. For example, a positive slope in conversation duration indicates the student is around more and more conversations as the term progresses, whereas a negative slope indicates the surrounding conversations decrease over time. The absolute value of the slope shows how fast the conversation duration changes. In addition to the symptom features, we also include the mean PHQ-4 score from each student in the dataset. The correlations between the mean PHQ-4 and PHQ-8 show the validity of the EMA administered PHQ-4.

The PHQ-4 dataset uses students' weekly self-administered PHQ-4 depression subscale scores as the groundtruth. We include the PHQ-4 depression group assignment (i.e., non depressed group (range 0-2) and depressed group (range 3-6) as defined in [123] as students' binary weekly depression state. For each PHQ-4 response, we compute the mean symptom features using data from the past 2 weeks. This is because PHQ-4 asks for participants' symptoms during the last two weeks.

3.4.3 PHQ-8 Association and Prediction Analysis

We test our hypothesis that the mobile sensing derived depression features represent proxies for the depression symptoms for college students by first running Pearson correlation analysis to assess the relations between the term symptoms features and the pre-post PHQ-8 item scores. Each of the PHQ-8 items maps to one of the major depression disorder symptoms defined in DSM 5 except the suicidal ideation symptom. The correlations between the PHQ-8 item scores and symptom features may give us preliminary insight about whether or not the symptom features are likely to be associated with individual symptoms.

We then run Pearson correlation analysis to assess the relations between the term symptoms features and the pre-post PHQ-8 scores. PHQ-8 scores are validated depression severity measures. The correlations suggest how the symptom features are related to the overall depression severity. We report the correlation coefficients and the p values. We also apply the False Discovery Rate (FDR) [207] to address the multiple testing problem [84, 70].

In addition to the correlation analysis, we use ANOVA [167] to test whether or not the mean of the symptom features are significantly different between the non depressed group and the depressed group. The PHQ-8 defines a non depressed group (< 10) and a depressed group (≥ 10) [124]. We use ANOVA [167] to test whether or not the mean of the symptom features are significantly different between the non depressed group and the depressed group. Analysis of variance (ANOVA) is a statistical model that is widely used to analyze the differences among group means [167]. We report the F statistics and the p value from ANOVA. The F statistics indicate the ratio of between-group variability and the within-group variability. The p value

indicates whether or not the group means are significantly different.

Finally, we use lasso regularized linear regression [184] to predict pre-post PHQ-8 scores. The lasso regularization selects features that are predictive of the outcome by penalizing irrelevant features' weights to zeros [184]. Before training the model, we normalize each feature to have zeros mean and one standard deviation. Feature normalization avoids the different feature scales adversely affecting regularization. We use 10-fold cross-validation to select the regularization hyperparameter, which controls the penalizing strength of non-zero-weight features. We choose the hyperparameter that minimizes the mean squared error (MSE). We report the mean absolute error and correlations between the predicted PHQ-8 scores and the groundtruth. We use the population PHQ-8 mean as the prediction baseline to compare the prediction performance.

3.4.4 PHQ-4 Regression and Prediction Analysis

We further test our hypothesis by identifying a number of associations between 2-week symptom features and the PHQ-4 depression subscale scores using regression analysis. We use the 2-week symptom features to predict if a student is considered depressed according the PHQ-4 depression subscale. The periodic PHQ-4 responses are longitudinal across the term where every student provides multiple responses. Ordinary linear regression and correlation cannot be applied to analyze longitudinal data because the responses from the same individual are likely correlated. We run bivariate regression analysis using the generalized linear mixed model (GLMM) [139] to understand associations between the 2-week symptom features and the PHQ-4 depression subscale scores. GLMM is a widely used model to analyze longitudinal

data. It describes the relationship between two variables using coefficients that can vary with respect to one or more grouping variables. In our case, the group variable is student. GLMM better explains the intra-individual differences. We normalize each symptom feature to have zero mean and one standard deviation. The regression coefficient of a normalized feature b can be interpreted as a unit increase in the feature value is associated with b increases in the associated PHQ-4 depression subscale value. A positive coefficient indicates that a greater feature value is associated with a greater depression score, whereas a negative coefficient indicates that a greater feature value is associated with a smaller depression score. PHQ-4 are collected at a weekly rate. The 2-week symptom features may artificially create dependency among consecutive PHQ-4 scores. To address the dependency problem, we first run GLMM on all PHQ-4 data, then we remove consecutive PHQ-4 responses by skipping a week's PHQ-4 response. The skip-a-week dataset is half of the PHQ-4 dataset in size. We compare the regression results from the complete PHQ-4 dataset and the skip-a-week dataset. Similar results from the two datasets would suggest dependency does not have an impact on the analysis.

We then use lasso regularized logistic regression [184] to predict whether or not a student is depressed week by week (i.e., the reported PHQ-4 depression subscale is ≥ 3). We use the PHQ-4 data to train a logistic regression model to predict each student's PHQ-4 for a given week. We first use 10-fold cross-validation to select the regularization hyperparameter, which controls the penalizing strength of non-zero-weight features. We then choose the hyperparameter that maximizes the regression deviance to train a generic PHQ-4 prediction model. We report the prediction performance from the 10-fold cross-validation.

3.5 PHQ-8 Results: assessing depression across the term

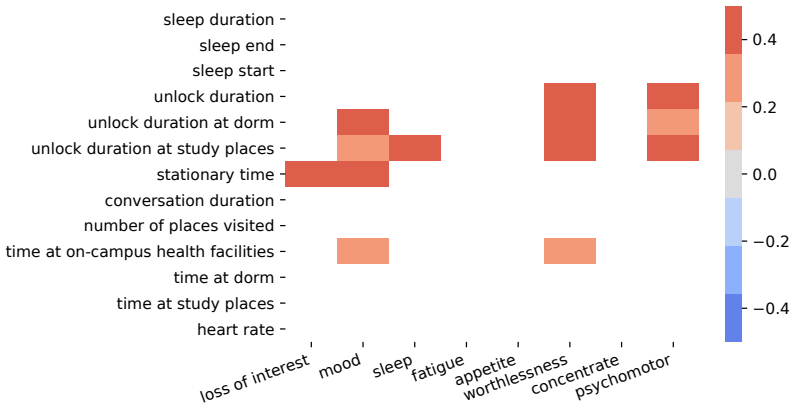
In what follows, we first report the correlations between the symptom features and PHQ-8 item scores to show whether or not the symptom features are likely to be associated with individual symptoms. We then report the correlations between the symptom features and PHQ-8 scores. Finally, we report ANOVA results to show whether or not the mean of the symptom features are significantly different between the non depressed group and the depressed group.

3.5.1 Correlations Between Symptom Features and PHQ-8 Item Scores

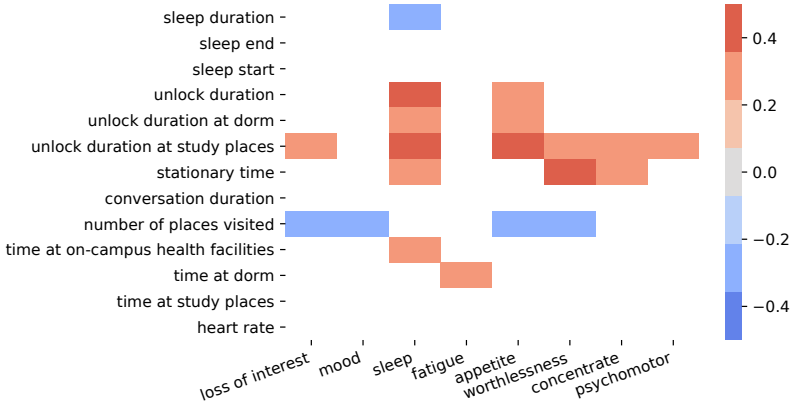
In what follows, we discuss the relationship between the specific PHQ-8 item scores and the proposed symptom features. Figure 3.3 shows the correlation matrices of the term mean symptom features and eight pre-post PHQ-8 item scores. We omit correlations with $p > 0.05$. In what follows, we discuss our findings.

Higher sleep changes (more irregular sleep patterns) item score is associated with shorter sleep duration (in line with our hypothesis), longer unlock duration during the day at dorm and study places, more time being stationary, and spending more time at on-campus health facilities. The sleep start time and end time, however, are not correlated with the sleep changes item score.

Higher mood item score is associated with longer unlock duration at dorm and study places, more time being stationary, and spending more time at on-campus health facilities. The heart rate is not correlated with the mood item score.



(a) PHQ-8 pre



(b) PHQ-8 post

Figure 3.3: The correlation matrix of proposed symptom features and PHQ-8 pre-post item scores. Correlations with $p > 0.05$ are omitted.

Higher loss of energy (fatigue) item score is associated with spending more time at dorm. The heart rate is not a predictor of this PHQ-8 item.

Higher diminished ability to concentrate item score is associated with longer unlock duration at study places (in line with our hypothesis) and more time being stationary.

3.5 PHQ-8 Results: assessing depression across the term

Higher diminished interest in activities item score is associated with longer unlock duration at study places, more time being stationary (in line with our hypothesis), and visiting fewer places a day (in line with our hypothesis).

Higher feeling worthless item score is associated with longer unlock duration during the day at dorm and study places, spending more time at on-campus health facilities, more time being stationary, and visiting fewer places a day.

Higher psychomotor retardation/agitation item score is associated with longer unlock duration during the day, at dorm, and at study places.

Higher appetite changes item score is associated with visiting fewer places a day, longer unlock duration during the day, at dorm, and at study places.

We find more statistically significant correlations between the symptom features and the post PHQ-8 item scores. We believe this is because post PHQ-8 scores better capture students depression states during the term whereas pre PHQ-8 scores capture the depression states when students started the academic term. The symptom features computed from the data collected during the term better capture students' depression symptoms during the term. The results show there are indeed association between the proposed symptom features and symptoms scores. We also find that some symptom features correlate with symptoms that are considered relevant. For example, phone use data (e.g., unlock duration) is associated with sleep changes.

Table 3.2: Pearson correlations between the term symptom features and pre-post PHQ-8 scores

		PHQ-8 pre		PHQ-8 post	
		r	p	r	p
sleep duration	all		> 0.05		> 0.05
sleep start	std	0.236	0.059	0.301	0.024
sleep end	std	0.183	0.145	0.271	0.043
unlock duration	mean	0.282	0.010*	0.268	0.024
unlock duration at dorm	mean	0.245	0.027	0.206	0.085
unlock duration at dorm	std	0.270	0.014*	0.222	0.062
unlock duration at study places	mean	0.391	< 0.001**	0.322	0.006*
unlock duration at study places	std	0.260	0.018	0.120	0.319
stationary time	mean	0.256	0.040	0.347	0.009*
conversation duration	slope	0.467	< 0.001**	0.223	0.062
number of places visited	mean	-0.066	0.556	-0.269	0.023
time at on-campus health facilities	mean	0.210	0.059	0.029	0.812
time at dorm	all		> 0.05		> 0.05
time at study places	all		> 0.05		> 0.05
heart rate	all		> 0.05		> 0.05
PHQ-4 depression subscale	mean	0.743	< 0.001**	0.849	< 0.001**
PHQ-4 depression subscale	std	0.328	0.003*	0.521	< 0.001**
PHQ-4 depression subscale	slope	0.045	0.688	0.438	< 0.001**

*FDR < 0.1, **FDR < 0.05

3.5.2 Correlations Between Symptom Features and PHQ-8 Depression Scores

The correlation results are presented in Table 3.2. In what follows, we discuss the correlation results in detail. Sleep duration, sleep start time, and sleep end time are proxies to measure the sleep changes symptom. We do not find correlations between sleep duration and PHQ-8 score. However, we find students who have more variations in their bed time schedule report higher pre ($r = 0.236, p = 0.059$) and post ($r = 0.301, p = 0.024$) PHQ-8 scores. Students who have more variations in their

wake up time report higher post PHQ-8 scores ($r = 0.271, p = 0.043$). The results show that students with more irregular sleep patterns tend to be more depressed.

Phone unlock duration during the day, at their dorms, and in study places are proxies to measure the diminished ability to concentrate symptom. We find all 3 unlock duration features correlate with the PHQ-8 scores. In general, students who use their phones more during the day report higher pre PHQ-8 score ($r = 0.282, p = 0.010$) and post PHQ-8 scores ($r = 0.268, p = 0.024$). When students are at their dorms, those who use their phone more report higher pre PHQ-8 score ($r = 0.245, p = 0.027$). We find strong correlations when students are at study places where the typically goal is to focus on school work. Students who spend more time using their phones in study areas report higher pre PHQ-8 scores ($r = 0.391, p < 0.001$) and higher post PHQ-8 scores ($r = 0.322, p = 0.006$). The results show that context aware device usage can be used to detect distractions and measure the ability of students to concentrate.

Stationary time during the day, conversation duration, number of places visited, and time spent at dorms and study places are proxies to measure the diminished interest or pleasure in activities symptom. Students who report higher post PHQ-8 scores are likely to spend more time being stationary ($r = 0.256, p = 0.040$ for the pre survey and $r = 0.374, p = 0.009$ for the post survey) and visit fewer places during the day ($r = -0.269, p = 0.023$ for the post survey). Students who report higher pre PHQ-8 scores see an increase in conversation duration (i.e., conversation duration slope) as the term progresses ($r = 0.467, p < 0.001$). However, there is no correlation between the mean conversation duration and the post PHQ-8 scores. We speculate this is because students who are depressed at the beginning of the term are not social

at the beginning of the term. However, as the term progresses, students become more socially engaged. This may because some students in the cohort seek help at on campus health facilities.

The daily median heart rate is a proxy for the depressed mood symptom and the fatigue or loss of energy symptom. However, it does not correlate with the PHQ-8 score. Students who report higher pre PHQ-8 scores at the beginning of the term spend more time at the campus health center ($r = 0.210, p = 0.059$). However, the correlation does not hold for the post PHQ-8 scores. The results show that although students who are more depressed at the beginning of the term actively seek medical help, students who become depressed during the term may not seek out help at campus health center. When talking to the clinicians and mental health counselors at the Dartmouth campus health center they indicate that the peek demand time they see students is during midterm weeks, during once per term social festivals, and at the end of term (week before, during and after the final exam period). Clearly, there seems to be a barrier for depressed students to visit the campus health center. Many issues might stop a student reaching out including not understanding they are experiencing depression or stigma associated with mental illness. The result is consistent with what other colleges experience [73].

We administer weekly PHQ-4 EMAs to track how students' depression changes as the term progresses. We compute the PHQ-4 depression subscale term average for each student and compare the subscale scores with the PHQ-8 scores. The PHQ-4 depression subscale scores strongly correlate with both pre ($r = 0.743, p < 0.001$) and post ($r = 0.849, p < 0.001$) PHQ-8 scores. The result show that PHQ-4 responses are consistent with PHQ-8, which gives us confidence to use the PHQ-4 depression

subscale to track depression changes during the term. The PHQ-4 depression subscale standard deviation correlate with both pre ($r = 0.328, p = 0.003$) and post ($r = 0.521, p < 0.001$) PHQ-8 scores. It suggests students who are more depressed have more variations in depression severity over the term. The PHQ-4 depression subscale slope correlate with post ($r = 0.438, p < 0.001$) PHQ-8 scores but not with the pre scores, which suggests the symptom severity may increase over the term for students who are more depressed by the end of the term.

3.5.3 Depression Groups Mean Comparison

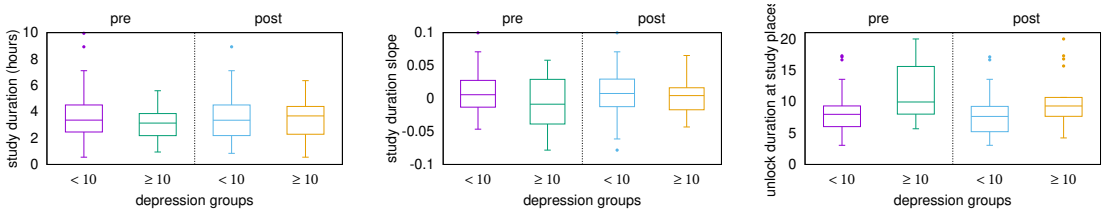
We show the ANOVA group comparison results in Table 3.3. In what follows, we discuss the group differences in sleep, conversation, and study behaviors.

Table 3.3: ANOVA significance of mean term symptom feature differences between the non depressed and depressed group

		PHQ-8 pre		PHQ-8 post	
		F	p	F	p
unlock duration	mean	5.179	0.026	5.733	0.019
unlock duration at study places	mean	11.599	0.001	6.084	0.016
unlock duration at study places	std	5.694	0.019	1.426	0.237
unlock duration at dorm	mean	4.748	0.032	6.121	0.016
unlock duration at dorm	std	5.042	0.027	5.443	0.023
conversation duration	slope	13.46	< 0.001	0.379	0.540
time at study places	slope	4.199	0.044	0.546	0.462
PHQ-4 depression subscale	mean	22.240	< 0.001	57.256	< 0.001
PHQ-4 depression subscale	std	0.312	0.578	11.207	0.001
PHQ-4 depression subscale	slope	0.319	0.574	6.716	0.012

Figure 3.4 shows the depression groups' distribution of the time spent at study places, the slope of the time spent at study places over the term, and the unlock duration at study places. Figure 3.4(a) shows that there is no significant differences in time

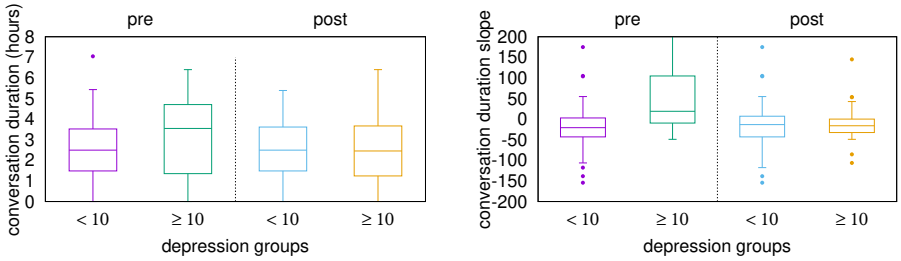
3.5 PHQ-8 Results: assessing depression across the term



(a) time at study places (b) time at study places slope (c) unlock duration at study places

Figure 3.4: The distribution of the time at study places, the slope of the time at study places over the term, and the unlock duration at study places of the pre-post PHQ-8 non depressed group and depressed group. Students from the depressed group

spend at study places between non depressed and depressed groups. Figure 3.4(b) shows that the pre PHQ-8 depressed group students spend a decreasing amount of time at study places whereas the non depressed group students spend same amount of time at study places across the term ($F = 4.199, p = 0.044$). Figure 3.4(c) shows that the PHQ-8 pre and post depressed group students spend more time using their phones at study places. The differences are significant with $F = 11.599, p = 0.001$ for the pre PHQ-8 groups and $F = 6.084, p = 0.016$ for the post PHQ-8 groups. Figure 3.5

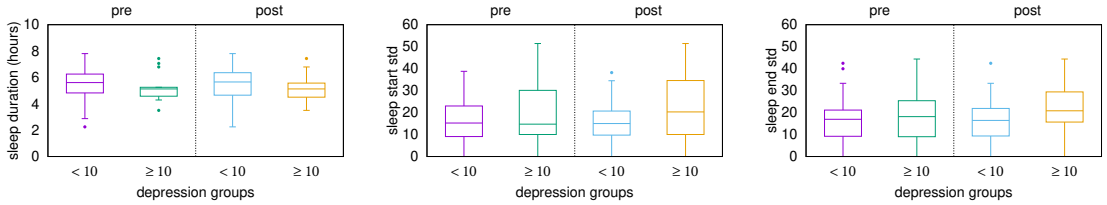


(a) conversation duration (b) conversation duration slope

Figure 3.5: The distribution of the conversation duration and the conversation duration slope of the pre-post PHQ-8 non depressed group and depressed group.

shows the distribution of the conversation duration and the conversation duration slope of the pre-post PHQ-8 depression groups. Figure 3.5(a) shows that there is no

significant differences in the mean conversation duration of the pre and post PHQ-8 depressed groups. However, the pre PHQ-8 depressed group shows a large in-group variation in conversation duration. Figure 3.5(b) shows that the pre PHQ-8 depressed group's conversation duration slope is positive whereas the non depressed group has a slight negative slope. The difference is significant with $F = 13.46, p < 0.001$. The result shows that students in the pre PHQ-8 depressed group are around an increasing amount of conversations as the term progresses whereas students in the pre PHQ-8 non depressed group are around decreasing amount of conversations. The difference in conversation duration slope does not exist in post PHQ-8 groups. The result is consistent with the correlation analysis.



(a) sleep duration (b) sleep start standard deviation (c) sleep end standard deviation

Figure 3.6: The distribution of sleep duration, sleep start time standard deviation, and sleep end time standard deviation for the pre-post PHQ-8 non depressed group and depressed group. The group differences are not statistically significant according to ANOVA.

Figure 3.6 show the distribution of sleep duration, sleep start time standard deviation, and sleep end time standard deviation of the non depression group and the depression group for pre-post PHQ-8. Figure 3.6(a) shows that the sleep duration of the depressed group is shorter than the non depressed. Figure 3.6(b-c) show that students in the depressed group have more variations in sleep start and end times. The differences, however, are not statistically significant ($p > 0.05$) according to ANOVA.

3.5.4 Predicting PHQ-8 Scores

Pre PHQ-8 scores. The lasso regularization selects 10 features to predict pre PHQ-8 scores. Specifically, it selects phone usage at study places, the stationary time, time spend at on-campus health facilities, sleep start time standard deviation, unlock duration at dorm standard deviation, unlock duration at study places slope, conversation duration slope, number of places visited slope, time at on-campus health facilities slope, and heart rate slope. The MAE of the baseline model, where the mean PHQ-8 score is used as the predicted PHQ-8 score, is 3.44. Our prediction model predict the pre PHQ-8 scores with MAE of 2.40, which is 1.04 lower than the baseline. The predicted PHQ-8 score strongly correlate with the groundtruth with $r = 0.741, p < 0.001$.

Post PHQ-8 scores. The lasso regularization selects 5 features to predict post PHQ-8 scores. Specifically, it selects phone usage at study places, the stationary time, number of places visited, sleep start time standard deviation, and conversation duration slope. The MAE of the baseline model is 4.29. Our prediction model predict the pre PHQ-8 scores with MAE of 3.60, which is 0.59 lower than the baseline. The predicted PHQ-8 score strongly correlate with the groundtruth with $r = 0.578, p < 0.001$.

Most of the selected features have shown significant linear correlations with the PHQ-8 outcomes, as shown in previous sections. Interesting enough, the lasso regularization selects the heart rate term slope to predict the pre PHQ-8 scores, whereas heart rate term slope does not show correlations with the PHQ-8 scores. We suspect the heart rate data may provide extra information in predicting PHQ-8 scores when combined with other symptom features.

3.6 PHQ-4 Results: tracking depression weekly dynamics

In what follows, we further test our hypothesis by identifying a number of associations between 2-week symptom features and the PHQ-4 depression subscale scores using regression analysis. We use the 2-week symptom features to predict if a student is considered depressed according the PHQ-4 depression subscale.

3.6.1 Regression Analysis

Table 3.4 shows the bivariate generalized linear mixed model (GLMM) [139] regression results. We report the coefficient b and the p value associated with the variable (i.e., the 2-week symptom feature) from the bivariate regression between the 2-week symptom features and the PHQ-4 depression subscale scores. The value of the coefficient indicates the direction and strength of the association between symptom features and PHQ-4 depression subscale scores. The p-value indicates the probability that the coefficient is equal to zero. A low p-value (i.e., < 0.05) indicates that the coefficient is not equal to zero and likely to be the learned value.

The results suggest students who are around fewer conversations per day ($p = 0.002$), sleep less ($p = 0.024$), and visit fewer places ($p = 0.003$) are likely to be more depressed. Students who go to sleep late ($p = 0.027$) and wake up late ($p = 0.001$) are likely to be more depressed. The regression coefficients and p-values are similar between the full PHQ-4 dataset and the skip-a-week dataset, which suggests the dependency does not have an impact on the analysis.

Table 3.4: Associations between symptom features and PHQ-4 depression subscale score

	all		skip a week	
	coefficient	p	coefficient	p
number of conversations	-0.269	0.002	-0.222	0.037
sleep duration	-0.156	0.024	-0.222	0.025
sleep end	0.151	0.027	0.236	0.012
sleep start	0.223	0.001	0.317	0.001
number of visited places	-0.211	0.003	-0.224	0.016

3.6.2 Prediction Analysis

The regularization selects 9 features to make the prediction as shown in Table 3.5. Specifically, it selects the stationary time, the number of conversations, heart rate, sleep end time, time spend at a dorm, time spend at study places, phone usage at study places, and the number of places visited. Similar to predicting PHQ-8 scores, the regularization selects the heart rate feature as a predictor. It further suggests the heart rate feature helps predicting depression outcomes.

Table 3.5: Selected features to predict PHQ-4 depression subscale non depressed and depressed group

lasso selected features
stationary time, number of conversations, heart rate, sleep end, time at dorm, time at study places, unlock duration at study places, unlock number at study places, number of places visited

Figure 3.7 shows the receiver operating characteristic (ROC) curve [94] of the logistic regression model obtained from the 10-fold crossvalidation. The ROC curve show the different true positive rate and false positive rate using a different threshold to determine if the logistic regression output is a positive case (i.e., depressed) or not.

The area under the ROC curve (AUC) [94] is a widely used metric to evaluate a binary classifier. AUC ranges from 0.5 to 1. Higher score indicates better performance. Our PHQ-4 state model's AUC is 0.809, which indicate good prediction performance. The model archives 81.5% of the recall (i.e., 81.5% of the depressed cases are correctly identified) and 69.1% of the precision (i.e., 69.1% of the inferred depressed cases are correct).

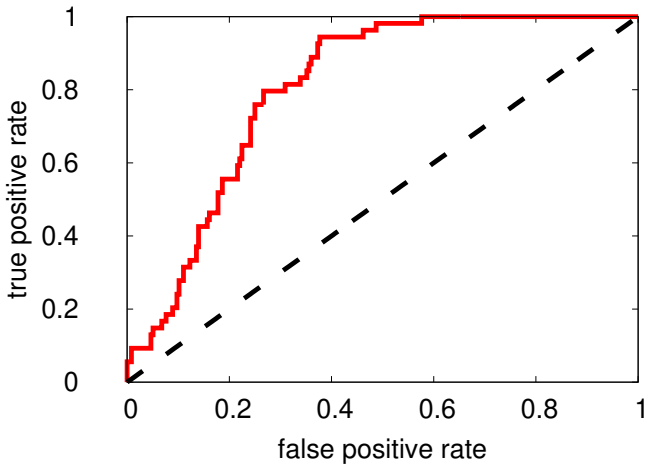


Figure 3.7: The ROC curve of using lasso logistic regression to predict PHQ-4 depression states. The area under the ROC curve (AUC) is 0.809.

3.6.3 Case Study Showing Depression Dynamics

Many of the students in this study have interesting behavioral curves and depression dynamics. Here we highlight one anecdotal case study. Figure 3.8 shows the depression dynamics of a student's PHQ-4 depression subscale score, number of conversations they are around, sleep duration, bed time, wake time, and number of places visited over a 9-week term. We do not want to identify this student by detailing their academic or demographic information. The curves show the student starts the term

in a non depressed state but their PHQ-4 depression subscale score deteriorates as the term progresses and peaks during week 4 (anecdotal this is the mid term week but we have no evidence that this is causal). The student's depression subscale score drops after week 4 and the student reports a non depressed state in week 6 before dropping to 0 in week 8. If we now compare the students behavioral data from the StudentLife app we can observe some interesting trends in the sensor data. Comparing the student's sleep data, the number of places visited every day and number of conversations they are around before and after their PHQ-4 depression subscale score peaks in week 4, we can observe that this student is around fewer conversations, sleep less, goes to bed later at night and wakes up earlier in the morning, and visit fewer places. This all seems indicative of a busy student who might be experiencing increased stress and anxiety because of the increasing academic demands of the term if we only looks at the behavioral curves from sensing data. However, seeing the depression dynamics from the PHQ-4 confirms that this student is struggling with the elevated levels of depression. The student has coping skills that are unknown to us and recovers from this increased risk without (from our data) any visits to the campus health center. As the term ends the student recovers showing resilience their behavioral sensing curves sleeping earlier, getting up later and therefore sleeping longer, visiting more locations on campus during the day, and being around more conversation indicating more engagement with fellow students and less isolation. All around a healthier person at the end of term. We do not present this example as some common set of curves we analyzed but an interesting one that gives insights into this student's academic term.

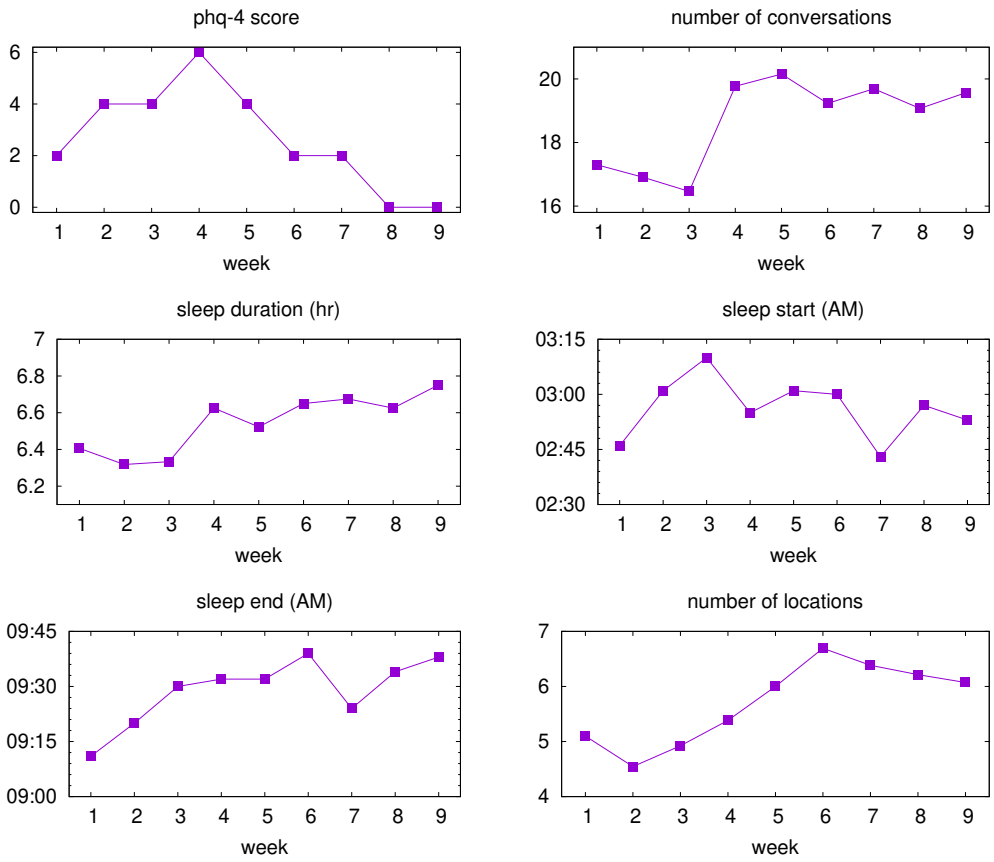


Figure 3.8: The dynamics of a student's PHQ-4 depression subscale score, number of conversations around, sleep duration, bed time, wake time, and number of places visited over a 9-week term. The student starts the term in a non depressed state but their PHQ-4 depression subscale score deteriorates as the term progresses and peaks during week 4 and drops to 0 in week 8. The student is around fewer conversations, sleep less, goes to bed later at night and wakes up earlier in the morning, and visit fewer places before week 4. As the term ends the student recovers showing resilience and their behavioral sensing curves sleeping earlier, getting up later and therefore sleeping longer, visiting more locations on campus during the day, and being around more conversation.

3.7 Discussion

Existing research on using passive sensing to predict depression do not design features that explicitly map to depression symptoms defined in DSM-5. For example, prior

work propose location features [49, 164, 163] based on the assumption that depressed persons would travel less and have more irregular mobility patterns. In this chapter, we propose to use well-known symptoms to guide us design passive sensing features that are more likely to be associated with depression. Incorporating the symptom domain knowledge and the behavioral characteristics of the population (e.g., college students), we can come up with novel behavioral features that leverage multiple sensor streams. For example, with the knowledge of college students' daily routine, we can leverage multiple sensors on the smartphone to assess students are likely to be distracted when they should be studying. However, the proposed symptoms features might not be able to generalize to other populations. For example, people with other occupations (i.e., non students) do not usually go to classes. Generic features, such as distance traveled and conversation duration, might not generalize well. We would expect salespersons would be around more conversations than people working in quite offices. Our method shows that we need to tailor behavioral features to different populations.

While the current study provides evidence that passive sensing may help tracking and predicting real-world changes in mental health, specifically depression, there are a number of limitations to our study. The sample size of our study dataset is relatively small relative to the number of features and quantity of data acquired for each of those features. As such, the results here should be considered relatively exploratory and preliminary. There is a need for the community interested in depression sensing to take the next step and conduct a large scale, longitudinal study with a diverse cohort well beyond students. While our results show associations between symptom features and depression groundtruth, associations cannot tell us if changes in behavior would

benefit or worsen depression severity. Future studies, need to be designed with the aim to find causalities between the proposed symptom features and changes in depression as discussed in [187]. Another limitation is that all of the subjects were Dartmouth undergraduates. While our results are statistically significant and encouraging they are limited to students at Dartmouth College. There are a number of other on-going studies looking into a wide-variety of student health issues as part of the CampusLife Consortium [129]. Results from these studies may shine a light on differences across different campuses (e.g., small Ivy in small town, large research university in city).

We evaluate the symptom features using the PHQ-8 item scores. The item scores, however, might not be a good indicator of symptoms because individual PHQ-8 items have not been validated against depression symptoms defined in DSM-5. Future studies may need to collaborate with clinicians to get better individual symptom measures.

Finally, the use of early wearables (in our case the Microsoft Band 2) present problems for longitudinal studies. The band could only hold a charge for approximately 14 hours while continuously sensing. This meant students would have to charge their bands before bed to get any data from the band during the night. Students had to take the band off while showering. Neglecting to put the band back on after showering would also cause losing data. Newer bands such as the Gamin Vivosmart 3 can run for 4 days and are waterproof potentially increasing compliance for wearables and making them more useful.

3.8 Conclusion

We proposed depression symptom features derived from phone and wearable passive sensor data that proxy 5 out of the 9 major depressive disorder symptoms defined in the DSM-5 for college students. We found students who report higher PHQ-8 scores (i.e., are more depressed) are more likely to use their phones more particularly at study places ($r = 0.391, p < 0.001$) in comparison with all day phone usage ($r = 0.282, p = 0.010$); have irregular sleep schedules (i.e., more variations in bed time ($r = 0.301, p = 0.024$) and wake time ($r = 0.271, p = 0.043$); spend more time being stationary ($r = 0.374, p = 0.009$) and visit fewer places during the day ($r = -0.269, p = 0.023$). We identified a number of symptom features capturing depression dynamics during the term associated with the PHQ-4 depression subscale groundtruth. Specifically, students who report higher PHQ-4 depression subscale scores (i.e., are more depressed) are around fewer conversations ($p = 0.002$), sleep for shorter periods ($p = 0.024$), go to sleep later ($p = 0.001$), wake up later ($p = 0.027$), and visit fewer places ($p = 0.003$) during the last two week period. We showed that the symptom features can predict whether or not if a student is depressed each week with 81.5% recall and 69.1% precision. We believe the methods and results presented in this chapter open the way for new forms of depression sensing going forward.

Although the results are encouraging, we still face many challenges to move forward with mental health sensing. First, we have tested the sensing technology in college students, who are tech-savvy and usually do not suffer from serious mental illnesses. We need to test the effectiveness of mental health sensing in people with serious mental illnesses. Second, we have used the smartphone data to correlate and predict mental health, however, we have not used the information for intervention.

I.e., how can we use mental health sensing to help clinicians adjust treatment when necessary. Finally, we need more rigorous study designs. For example, the study should have a clear inclusion and exclusion criteria for participation and account for different demographic factors and social economics. To test the effectiveness of intervention based on mental health sensing, we would need a randomized controlled trial to reduce bias.

In the rest of this thesis, we discuss applying smartphone sensing technology in a more challenging population: people with serious mental illness (i.e., schizophrenia). We discuss the CrossCheck RCT, which split the participants into two groups: a smartphone group, and a treatment-as-usual group. We focus on the smartphone group and discuss using the smartphone data to track how schizophrenia symptoms changes and use the smartphone data to provide interventions. We would also discuss using the smartphone data to predict whether or not a patient is going to relapse.

Chapter 4

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

4.1 Introduction

In the first part of this thesis, we discuss using smartphone sensing to assess college students' mental health (especially depression) and academic performance. The promising results show there are great potentials in applying mental health sensing from smartphones in people with serious mental illnesses and provide interventions. In this chapter, we discuss the CrossCheck randomized control trial and analyze preliminary data from the trial. CrossCheck app is a smartphone sensing system

deployed to outpatients with schizophrenia. It 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 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 chapter, 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: (i) 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; (ii) 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; and (iii) 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.

4.2 Related Work

There is 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 [40, 72, 176]. Early work in mobile mental health for schizophrenia conducted by Ben-Zeev et al. [33] first study the feasibility and acceptability of using mobile devices for behavioral sensing among individuals with schizophrenia. In [33] 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. Kerz et al. [114] 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. A recent pilot study [23] collect smartphone data from 17 patients with schizophrenia in Boston and find increased rate of behavioral anomalies detected in the 2 weeks prior to relapse. the subjects, can provide an unprecedented and detailed view into patient behavior outside the clinic.

4.3 CrossCheck Study Design

The CrossCheck study is a randomized control trial (RCT)[50] conducted in collaboration with a large psychiatric hospital in New York City, 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 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.

4.3.1 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 [204]. Individuals with a legal guardian are excluded.

4.3.2 Recruiting Participants

The staff at the recruitment hospital first screened candidates based on criteria described in 4.3.1. 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

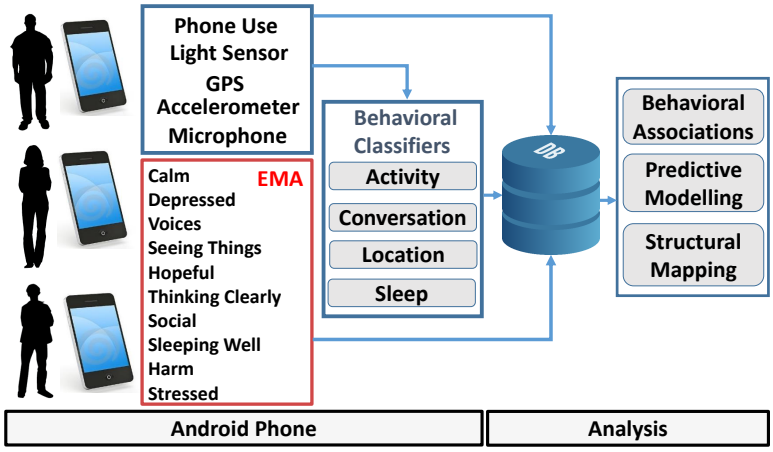


Figure 4.1: CrossCheck sensing and analysis system.

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).

4.3.3 CrossCheck System

The CrossCheck sensing system is built based on StudentLife described in Chapter 2 that uses smartphone sensing and self-report tools. Compared with the StudentLife sensing system, 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 [196] to administer EMAs [33]. During the collection phase, participants are asked to respond to EMA questions every Monday, Wednesday, and Friday (see Section 4.4). This chapter 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 4.1 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 non-compliant 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.

4.4 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 [150, 138, 155, 196, 8, 127, 49, 164]. We use self-reported EMA data as mental health state indicators of schizophrenia patients.

4.4.1 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.

4.4.2 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 [155, 196, 127] 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 [41]. 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[54, 196].

Location. Prior studies have shown that a user's mobility patterns from geo-location traces are associated with mental health and wellness [49, 196, 164]. 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 [49, 164]. 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 [136] 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 [54]. We use the mean illumination over a day and within the four epochs.

4.4.3 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 [18]. 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[90]. The set of EMA questions we use in CrossCheck are based on self-reported dimensions defined in previous schizophrenia research [31]. 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 4.1). 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 4.1: 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.

4.5 Analysis and Results

We identify a number of important associations between phone-based behavioral features described in Section 4.4 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.

4.5.1 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 Section 4.5.3. We then run prediction analysis using gradient boosted regression trees (GBRT) [84, 152] to evaluate the feasibility of predicting EMA sum scores, which is discussed in Section 4.5.4. Finally, we generate person specific models using random forest (RF) [44] 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 4.2 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.

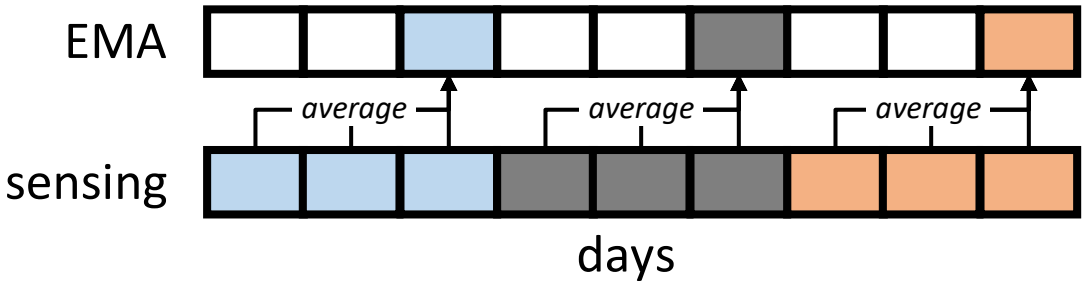


Figure 4.2: Feature/EMA preparation

4.5.2 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) [132] method. The t-SNE [132] 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

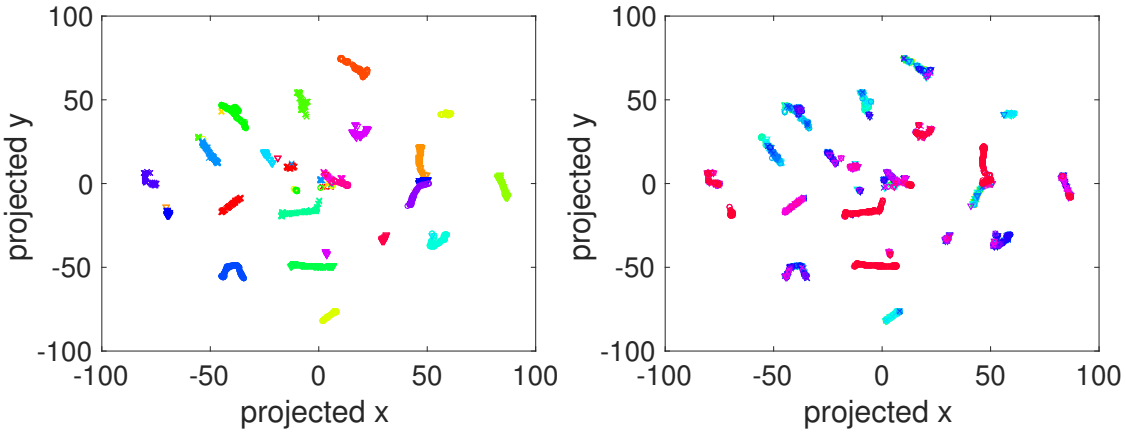
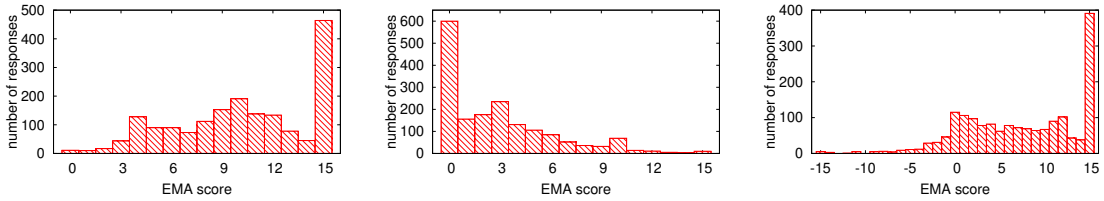


Figure 4.3: Feature visualization using t-SNE. (a) Data is color coded by user ID. Individual subject's data clusters together. (b) Data is color coded by EMA sum scores. Data with same score tend to cluster within subject.

dissimilar data points are projected to distant points. The feature visualization is shown in Figure 4.3.

Figure 4.3(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 is interesting because it shows that our features captures behavioral difference between different individuals and that the data is highly person dependent. Figure 4.3(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 4.3(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.



(a) Positive score distribution. (b) Negative score distribution. (c) Sum scores distribution.

Figure 4.4: EMA aggregated score distributions

4.5.3 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) [128, 46, 209, 68] – 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 inter-person 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.

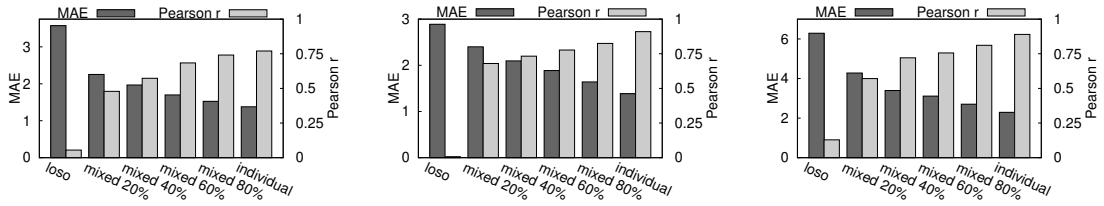
We apply a bivariate regression using GEE to all 610 combinations of the 61 features and 10 EMA questions. We apply the Benjamini-Hochberg procedure (BH) proposed in [35, 36] to inform the false discovery rate (FDR) in our exploratory regression analysis. The BH procedure finds a threshold for the p 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 $p < 0.0016, \text{FDR} < 0.1$, and 7 regressions with $p < 0.00025, \text{FDR} < 0.05$.

Positive Questions. Table 4.2 shows features that are associated with the five positively worded questions (*viz.* *calm*, *social*, *thinking clearly*, *sleeping well* and

Table 4.2: 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***.	

hopeful). A higher 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



(a) Predicting positive scores. (b) Predicting negative scores. (c) Predicting sum scores.

Figure 4.5: 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.

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 4.3 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.

4.5.4 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 4.4.

Personalized EMA Predictions

Predicting the aggregate EMA scores is a regression task. We use gradient boosted regression trees (GBRT) [84, 152] 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 [152]: $F(x) = \sum_{m=1}^{M} \gamma_m h_m(x)$, 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 [108, 84] also GBRT is less sensitive to outliers [108].

4.5 Analysis and Results

Ideally, an EMA score prediction system should be able to predict a new user's scores accurately. However, the visualization of participants' data (Figure 4.3) 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 different 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 [38, 45, 173, 93] 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 [45], 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 [38]. 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 4.5 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 4.3 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 4.5b), 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 con-*

fidence 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 4.4.

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.

Predicting Underlying EMA Trends

In this section, we investigate the prediction of underlying trends in the EMA score specific to each participant. Figure 4.6 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) [44] 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.

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

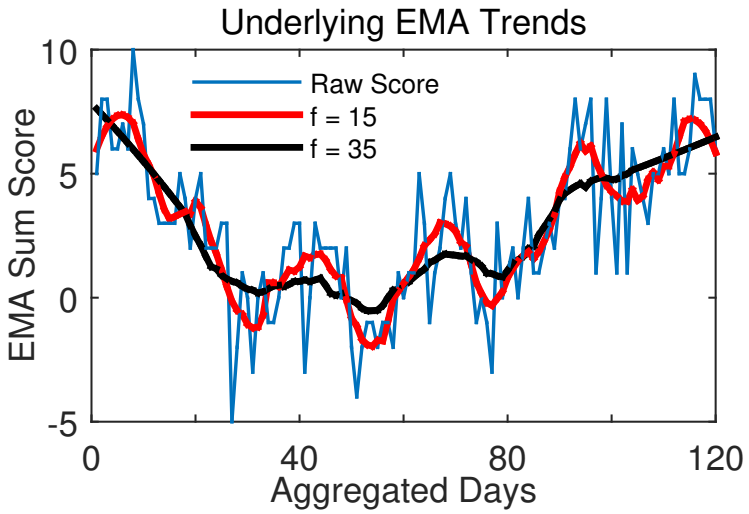


Figure 4.6: Examples of smoothing on EMA sum score from one participant where f is the frame size of the Savitzky-Golay filter.

performance using Mean Squared Error (MSE). Figure 4.7 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 4.7 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

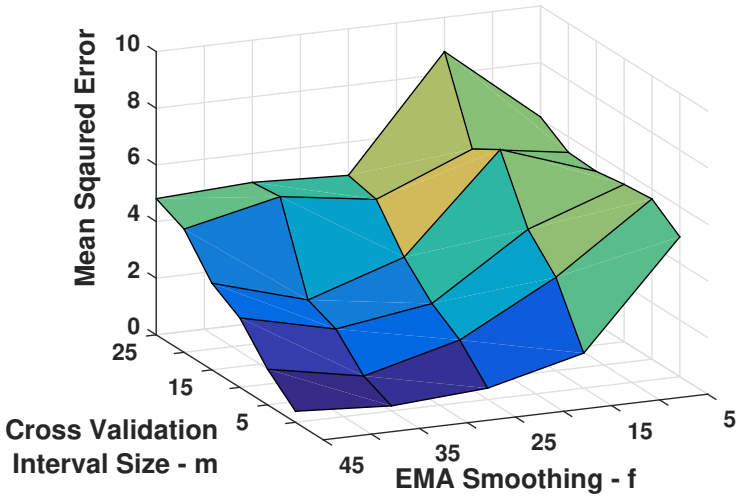


Figure 4.7: Mean Squared Error from Leave-One-Interval-Out validation for interval sizes versus smoothing level.

is relevant and predictive by 10-15 days.

4.6 Conclusion

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 [81]. 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 [49, 164]. 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 [159] 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. In the next chapters, we discuss leveraging the smartphone data to predict clinician evaluated symptom scores and use the predictions to provide interventions. In Chapter 6, we discuss using the smartphone data to predict schizophrenia relapses.

Table 4.3: 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***.	

Table 4.4: 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

Chapter 5

Predicting Symptom Trajectories of Schizophrenia using Mobile Sensing

5.1 Introduction

In Chapter 4, 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. The results show that we can track schizophrenia patients' symptoms. 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. Participant recruitment and the consent procedure are described in Section 4.3. In this chapter,

we turn to predict the 7-item BPRS, which is administered by a trained clinician. The predictions are used by our research staff to determine at-risk patients and alert their their clinicians.

At the CrossCheck study [195] 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 [98, 69, 179, 118] 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 [101]. 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.

We present 7-item BPRS prediction results from CrossCheck 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: (i) 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; (ii) 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; (iii) we use leave-one-record-out and leave-one-subject-out cross validations [117] 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; and (iv) we discuss anecdotal information associated with three patients in the study. These case studies show that our system can identify patients with rising risk.

5.2 CrossCheck Symptom Prediction System

The CrossCheck symptom prediction system comprises the CrossCheck app running on Android phones and the CrossCheck data analytics service running in the cloud. The CrossCheck app collects participants' passive sensing data and self-reported EMA data [195] 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 5.1 summarises the systems components, workflow, and outreach. The CrossCheck app is described

in Chapter 4 in detail. In this section, we focus on the CrossCheck data analytics system.

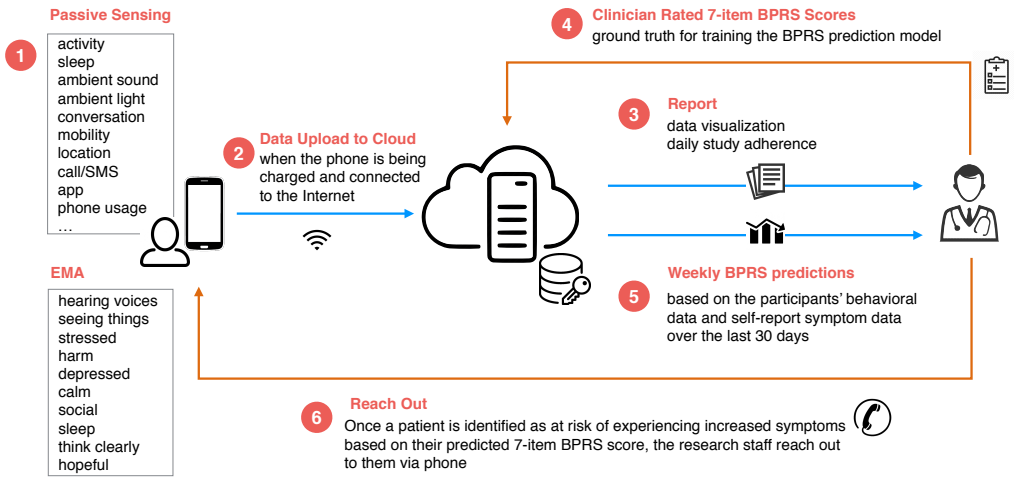


Figure 5.1: System overview of the CrossCheck symptom prediction system

5.2.1 CrossCheck App and Sensing System

The CrossCheck app [195] is built based on StudentLife core sensing system describe in Chapter 2. 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 [196] to administer self-reported EMAs [33]. To protect participants' privacy, the app does not collect phone numbers, content of text messages or any conversation content [196]. We remotely

erase the CrossCheck data on the phone and reset it if 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. We administer a 10-item CrossCheck EMA every Monday, Wednesday, and Friday. The EMA asks participants to score themselves on been feeling calm, social, bothered by voices, seeing things other people can't see, feeling stressed, worried about people trying to harm them, sleeping well, able to think clearly, depressed, and hopeful about the future. The detailed questions are list in Table 4.1.

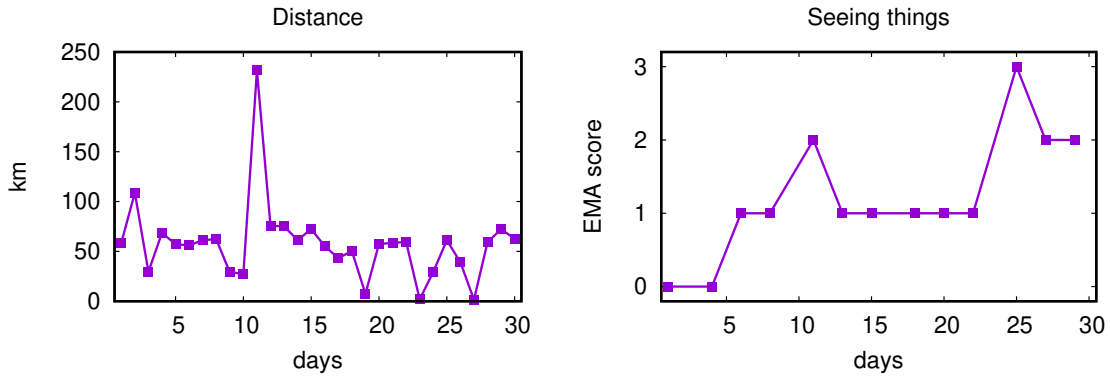
5.2.2 CrossCheck Data Analytics System

The CrossCheck data analytics system receive and process the data from the CrossCheck 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. 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. Our study staff input the clinician rated 7-item BPRS scores to the CrossCheck data analytics 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, Figure5.2 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

Figure 5.2: 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. 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. 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 5.4.

Daily study adherence reports. To monitor patients' study adherence and detect if any participants are experiencing technical issues that prevent the app from up-

loading 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.

5.3 CrossCheck BPRS Dataset

The CrossCheck symptom prediction system is deployed in a randomized controlled trial [50] conducted in collaboration with a large psychiatric hospital, Zucker Hillside Hospital, in New York City [195]. In what follows, we discuss the CrossCheck dataset in detail.

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.

5.3.1 The 7-item Brief Psychiatric Rating Scale

The BPRS [151, 98, 69, 179, 118] 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 [101]. 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 5.1 lists all the BPRS items [151]. 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.

As mentioned only 7 out of 24 items (marked in bold in Table 5.1) are evaluated

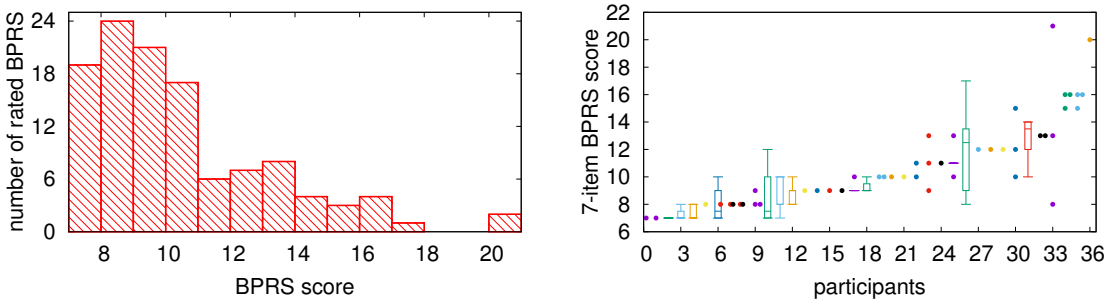
Table 5.1: Brief Psychiatric Rating Scale Items

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.

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 [42] and the theory behind the BPRS see [101, 151]. 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 5.3(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 5.3(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

Figure 5.3: 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).

.

5.3.2 Feature Set Constructions

The dataset has two sets of features: passive sensing features and self-report EMA features, which are described in Chapter 4 in detail. 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.

5.3.3 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.

5.4 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.

5.4.1 Predicting BPRS Scores

We use Gradient Boosted Regression Trees (GBRT) [84, 152] 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 [108] and robust to overfitting [77]. 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) [128, 46, 209, 68] 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 [128]. 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 5.2: 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*

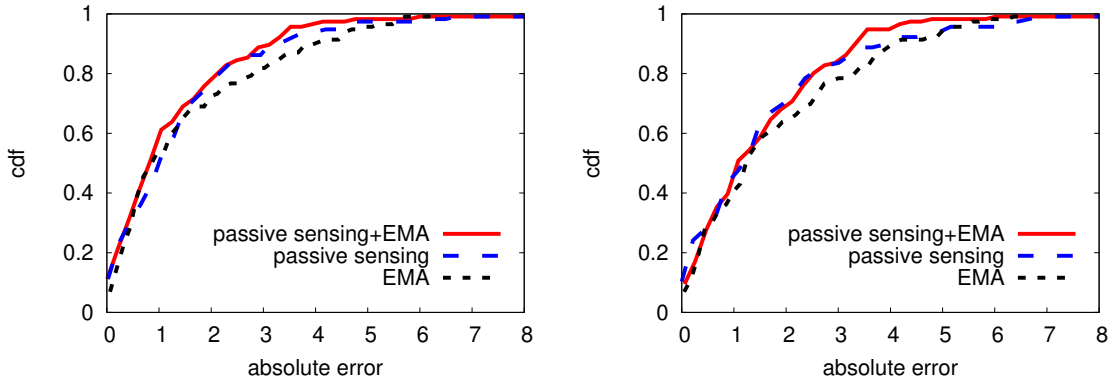
* $p < 0.0001$

Table 5.2 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 5.4 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.

Within-individual prediction errors. Figure 5.5 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 5.3(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



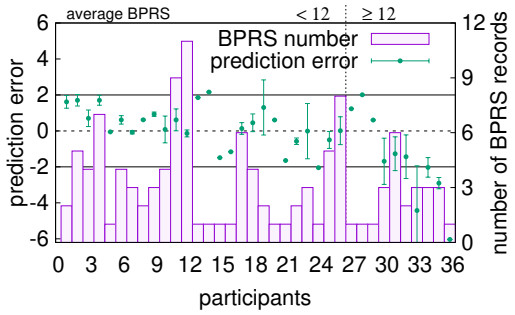
(a) leave-one-record-out. (b) leave-one-subject-out.

Figure 5.4: 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.

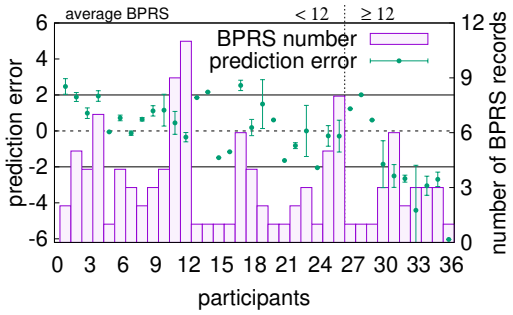
underestimate high-BPRS-score participants' scores (i.e., participants with average BPRS $\geq$ 12). The prediction models need more high-BPRS-score participants' data to improve the prediction performance. The impact of prediction errors on clinical practice is discussed in Section 5.4.3.

5.4.2 Interpreting Selected Features

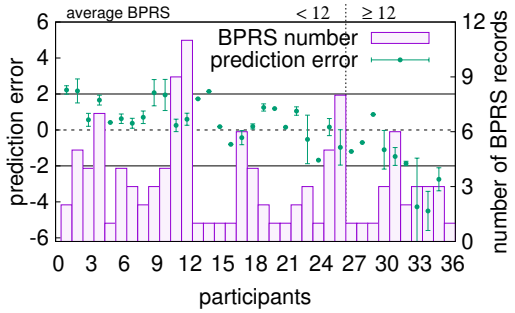
We use bivariate regression analysis to understand the linear relationship between the features selected by GBRT feature importance measures and the 7-item BPRS scores. Considering the longitudinal nature of our dataset; that is, because data from the same subject is likely correlated we apply generalized estimating equations (GEE) [128, 46, 209, 68] 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



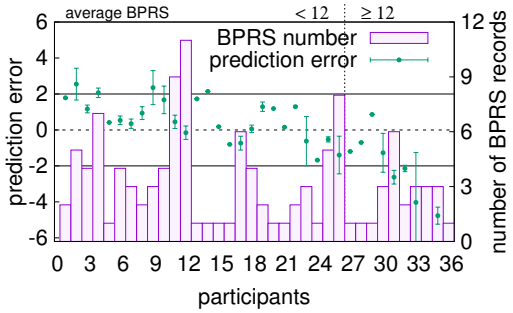
(a) passive sensing + EMA leave-one-record-out.



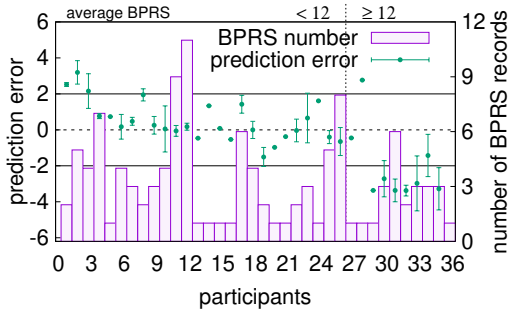
(b) passive sensing + EMA leave-one-subject-out.



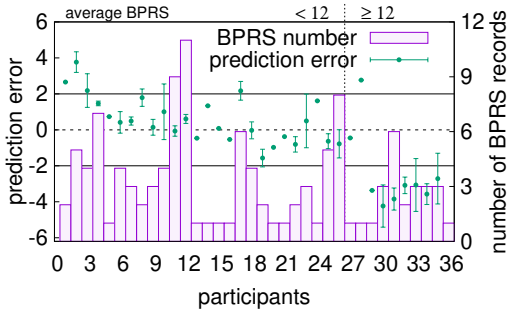
(c) passive sensing leave-one-record-out.



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



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



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

Figure 5.5: 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.

the coefficients are within a reasonable range. Table 5.3 shows the 13 selected features from the model using both passive sensing and EMA features based on the

feature selection criteria described in Section 5.4.1. 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 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 5.4 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

Table 5.3: 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

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 5.5 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.

Table 5.4: 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

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 [91] has found features with little power at predicting outcomes when combined with other features can provide significant performance improvements.

Table 5.5: 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

5.4.3 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 from clinicians. Figure 5.6 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

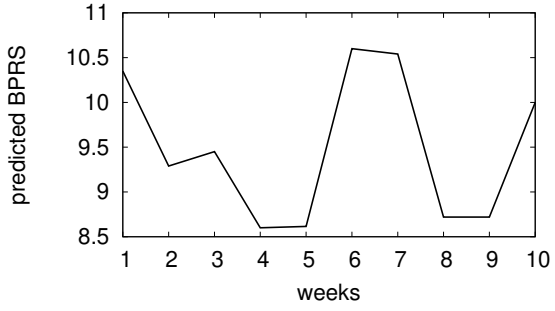
clinician's scores and the predicted scores may due to the different days the patient was rated and the predictions were made. The predictions, however, show that the participants' BPRS scores are slightly higher before the first evaluation, between two evaluations, and after the second evaluation. The 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 5.5(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 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.



111

Figure 5.6: 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.

5.5 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.

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 chapter 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.

5.6 Conclusion

The CrossCheck system discussed in this chapter 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.

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 chapter. 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 [52], 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 5.3 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.

Ultimately, we aim to develop a system capable of not just indicating rising risk, but rather, to develop a model and associated systems to accurately predict relapse and through intervention and treatment adjustment keep patients healthy and out of hospital. In the next Chapter, we evaluate different methods and design considerations to use the smartphone data to predict whether or not a patient is going to relapse.

Chapter 6

Predicting Relapses in Schizophrenia using Mobile Sensing

6.1 Introduction

In Chapter 4 and Chapter 5, we discussed using passive sensing data from smartphones to assess schizophrenia patients' symptom. In this chapter, we take one step forward predicting whether or not a patient is going to relapse using smartphone data from the CrossCheck study. A participant relapses if one of the following 7 events happens [63]: 1) psychiatric hospitalization, 2) increased frequency or intensity of services, 3) increased dosage / additional medication and 25% increase in BPRS (the brief psychiatric rating scale) [151] from baseline/last assessment, 4) suicidal ideation, 5) homicidal ideation, 6) self-injury, and 7) violent behavior resulting in damage to property or person. Relapses are determined by trained clinical assessors using partic-

ipants' clinical records. A relapse date and reason for relapse are recorded by trained clinical assessors. In this chapter, we focus on the relapse date and predict whether or not a participant is going to relapse the next day.

There are two main challenges of predicting schizophrenia relapses in the CrossCheck study participants: 1) the relapse cases are rare: out of 61 participants over a year, we identify 27 relapses from 20 participants; and 2) the CrossCheck relapse dataset is imbalanced. Considering predicting whether or not a patient is going to relapse every 30 days with our dataset, there are in total 430 predictions (i.e., prediction windows) to be made in our dataset. Only 19 of the predictions (i.e., 4.42% of all predictions) are labeled by clinical assessors as relapse. The definition of prediction windows is defined in Section 6.2.3. Most standard classification algorithms assume a relatively balanced class distribution and equal misclassification costs. An imbalanced dataset violates such an assumption, which leads to poor classification performance. We apply various techniques to address the challenges. First, we applying resampling methods combining oversampling relapse examples and undersample non-relapse examples to the training dataset such that the number of relapse and non-relapse examples are the same. Second, we impute missing sensing data to make sure we have enough data to train the classifiers. Third, to avoid over-fitting, we apply feature selection (e.g., L1 regularization) and feature transformation (i.e., PCA) to reduce the feature dimensionality. Finally, we present a new 2-level 3-fold stratified cross validation to incorporate training data resampling in relapse prediction evaluation.

In this chapter, we investigate the efficacy of using passive sensing data and/or self-report EMAs to predict relapses. We present classification performance from using only EMA or sensing data, and a combination of EMA and sensing data. We

investigate what would be the best time window to predict relapse. A relapse prediction time window is the number of days' data we use to predict relapse. We explore using PCA to transform the feature space and reduce the dimensionality for classification. Reducing the dimensionality helps training the prediction model more efficiently and removes multicollinearity (i.e., predictors are correlated). We present features that are the most predictive of impending relapse.

We find the best relapse prediction result (i.e., the highest F1 score) using the first 100 principle components (PCs) from both passive sensing and EMA with 30-day prediction windows (precision=26.8%, recall=28.4%). If we demand the recall to be greater than 50%, we find the best result using 25 PCs from both passive sensing and EMA with 30-day prediction windows (precision=15.4%, recall=51.6%). In what follows, we discuss our results in detail.

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. Participant recruitment and the consent procedure are described in Section 4.3.

6.2 Method

We aim to predict whether or not a CrossCheck participant relapses during the span of the study using the smartphone passive sensing data and self-report EMAs. In what follows, we discuss the relapse dataset, data preprocessing, behavioral features computed from the passive sensing data, and prediction models in detail.

6.2.1 Dataset

By the end of the CrossCheck study, we recruited 61 participants in the smartphone arm. 26 of the 61 participants are female and 35 male. There are 24 African American, 5 Asian, 2 Multiracial, 29 Caucasian and 1 Unknown. The average number of days a participant is in the study is 322 days (SD = 93, min=22, median=361, max=522). We identify 27 relapses from 20 participants, in which 16 participant relapse once, 1 participant relapse twice, and 3 participants relapse three times each. We collect a wide range of passive sensing data from smartphones and self-report EMAs, which are discussed in detail in Chapter 4 and 5.

Specifically, we collect physical activities, locations, ambient sound levels, voice/noise labels, number of calls and text messages, application usage, screen lock/unlock, and ambient light intensity. We compute features from the passive sensing data on a daily basis, which describe participant's behaviors (e.g., duration of different physical activities in a day, conversation duration and frequency, different types of places visited, app usage). We do not collect the content of any phone calls, text messages or applications, and we do not record any raw audio data.

We administer a 10-item EMA every Monday, Wednesday, and Friday. The EMA asks participants to score themselves on been feeling calm, social, bothered by voices, seeing things other people can't see, feeling stressed, worried about people trying to harm them, sleeping well, able to think clearly, depressed, and hopeful about the future. The detailed questions are list in Table 4.1.

6.2.2 Behavioral Features

We incorporate passive sensing features proposed in previous work [195, 199] in which the features are predictive of self-reported and clinician-administered symptoms among schizophrenia patients. The features are computed on a daily basis and also broken down into four epochs of the day: *morning (6am-12pm)*, *afternoon (12pm-6pm)*, *evening (6pm-12am)* and *night (12am-6am)*. These epoch features allow us to model people's behaviors during different parts of the day (e.g., walking in the morning, sleeping in the afternoon, not socially engaged in the evening, using the phone a lot during the night period).

Specifically, we compute the following features. To measure *physical activities*, we compute duration for different activities (e.g., on foot, still, in vehicle, and on bicycle), and in order to measure *mobility*, we compute the number of locations visited and distance traveled. To measure *sleep patterns*, we infer sleep duration, sleep start, and end time from sensing data, and to measure *ambient environmental context*, we compute the amplitude of ambient sound and ambient light. We also compute *face-to-face conversations* features which consist of conversation frequency and duration, and *smartphone-usage* features including the number of phone calls, SMS, and lock/unlock frequency and duration.

Semantic location. We aim to assign semantics to places where participants visit. Specifically, we consider the following places: home, food, travel, art&entertainment, nightlife, education, parks&outdoors, library, shop, gym, medical and residence. We compute the time spent at these places every day.

We first identify significant locations where a participant dwells for a significant amount of time of the day. We find significant locations by clustering the GPS coor-

dinates collected in a day using density-based spatial clustering of applications with noise (DBSCAN) [136]. The centroid of each cluster is considered a significant location. We assume participants are usually at their homes sleeping between 2am to 6am. Therefore, we label a significant location as home where a participant spends most of the time between this period of the night. We then use the Foursquare API [80] to label the other significant locations. The Foursquare API takes a GPS coordinate and a radius as input and returns a list of location entities. Each location entity is associated with name, coordinates, and categories (e.g., food, art&entertainment). We heuristically set the radius to 50 meters. A location may be associated with multiple different categories. We compute the dwell duration at a location for all associated location categories. For instance, if the API returns "food" and "art&entertainment" categories for a given significant location. We include the dwell duration at this significant location to both "food" and "art&entertainment" places. This is an approximation given that there is error in location coordinates. We do not simply select a single returned location entity closest to the significant location coordinate because as mentioned GPS data is noisy [74] and different types of location may reside in the same building (e.g., food, office).

6.2.3 Data preprocessing

In what follows, we describe our data preprocessing procedure, which include aggregating daily features in different relapse prediction time windows, data cleaning, missing data imputation, and feature space transformation and dimensionality reduction.

Relapse prediction time window. We define the relapse prediction window as the number of days before the day identified as the *start of a relapse*. Studies find that most patients with schizophrenia experience symptoms 30 days before relapse [40, 103, 102, 181]. Therefore, we evaluate relapse prediction using four different time windows: 7 days, 14 days, 21 days, and 30 days. We summarize the daily features within the prediction window as the average value of each of the features. The prediction time window construction is illustrated in Fig. 6.1. Specifically, suppose the prediction window size is 7 days, we first identify the date of the first relapse, then we group the 7 days before the relapse day into a 7-day block and label the block as 1 (relapse). We compute the average value of every feature within the 7-day block. Then we groups 7 days before the first day of the relapse block into a 7-day block and label the block as 0 (non-relapse). We repeat until reaches the beginning of the study. If the last block is shorter than 7 days, we discard the block. We discard 30 days of data after each relapse because the participants may be hospitalized and do not have phones with them while in hospital. We repeat the above steps to group and label prediction windows for the rest of the data.

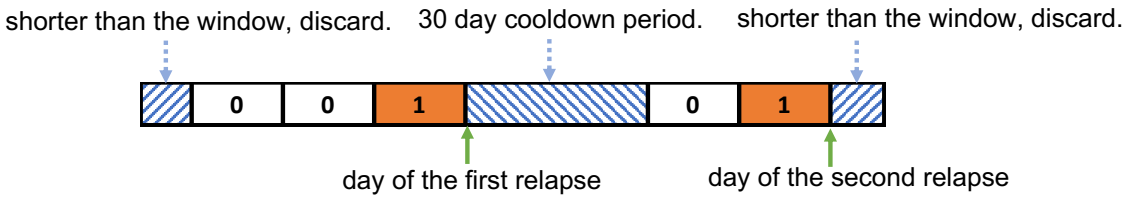


Figure 6.1: Prediction window construction. Each window is labeled as 0 (non-relapse) or 1 (relapse in the following day). We introduce a 30 day cooldown period after a relapse (shaded area), which we exclude from the prediction. The cooldown period is when a patient experiencing relapse (e.g., hospitalized), which should not be used to predict future relapses. We exclude window with fewer days than the target window length (shaded area).

Data cleaning and imputation. We compute behavioral features on a daily basis. Poor daily data quality may skew the behavioral features, therefore, we exclude poor quality data from our analysis. We define poor quality daily data as days in which more than 5 hours of data are missing. We use this threshold across this thesis because we find lowering the threshold (e.g., missing 10 hours of data) does not include significantly more data whereas the data quality is poorer. Specifically, we compute the number of hours of data we have received for each passive sensing data. We label the sensing data as missing if less than 19 hours of data are collected in that day. We also control the data quality for aggregated time window features. We label the average feature values as missing if the feature misses over 70% of the days in the time window. We exclude time windows with more than 70% of feature values that are missing from our analysis. We heuristically pick the threshold to balance the data quality and make sure we have enough data for our analysis. The number of non-relapse and relapse cases are presented in Table 6.1. To handle missing values, we use a Singular Value Decomposition (SVD) based method SVDimpute[186] to impute missing values. SVDimpute is a robust and sensitive method for missing value estimation surpassing the commonly used row average method [186].

Per-participant standardization. We use per-participant standardization to remove between-individual differences from the behavioral features. We hypothesize that different people may have different behavioral baselines. For example, a construction worker might be more physically active than an office worker whereas they have the same mental health outcomes (e.g., relapse). However, the within-individual differences in behaviors might be more inductive of changes in mental health. Per-participant standardization removes the between-individual behavioral

differences and keeps only within-individual behavioral differences. We test our hypothesis in Chapter 6.3.

Per-participant standardization transforms a participant's passive sensing features and EMA responses according to their first 30 days' data. Specifically, we first compute the mean μ_{30} and standard deviation σ_{30} for each of the features in the first 30 days, then we transform the feature as follows: $v_t = (v - \mu_{30})/\sigma_{30}$, where v is the original feature vector and v_t is the transformed feature vector. We apply per-participant standardization before aggregating features into prediction windows. We evaluate relapse prediction performance with or without per-participant standardization.

Feature space transformation and dimensionality reduction. We use principle components analysis (PCA) [168] to transform the feature space and reduce the feature dimensionality. PCA transforms a set of observations of possibly correlated variables (i.e., features) into a set of values of linearly uncorrelated variables called principal components. The principle components are defined in a way that the first principal component accounts for as much of the variability in the data as possible, and each succeeding component in turn account for as much of the rest of the variability in the data as possible. The resulting principle components are an uncorrelated orthogonal basis set. The original observation (i.e., the feature values in a prediction window) can be reconstructed by a linear combination of the principle components. We use the weights of each PC as transformed features. We can use a smaller number of PCs to reconstruct the original observation, which leads to a smaller number of features (i.e., reduce the feature space dimensionality). Each PC can be interpreted as a behavioral pattern. For example, if a PC has large positive weight in the component for phone unlock duration and phone call duration features, and large negative

weight for still duration, we would interpret this PC represents a high phone usage and high sedentary behavioral pattern.

We evaluate relapse prediction for different PCA setups. We first use the raw feature values to predict relapses. Then, we experiment with using different number of PCs to predict relapse. Specifically, we test using the first 1, 2, 5, 10, 25, 50, and 100 PCs, which explain 28.9%, 45.1%, 69.5%, 80.1%, 90.2%, 96.9%, and 99.9% of the variance in sensing and EMA data combined, to predict relapses.

6.2.4 Relapse Prediction as Binary Classification

Relapse prediction is a binary classification problem, i.e., we classify a n-day time window as non-relapse or relapse. We evaluate four popular classifiers: logistic regression, SVM with linear kernel [61], SVM with radial basis function kernel (RBF kernel) [51], and random forest [44, 104]. The classifiers include linear classifiers (i.e., logistic regression, linear SVM), non-linear classifiers (i.e., RBF SVM, random forest), and non-parametric classifier (i.e., random forest). We apply elastic net regularization [210] on logistic regression and linear SVM to avoid over-fitting. The elastic net linearly combines the L1 and L2 penalties of the lasso and ridge methods. We use grid search to find the best model hyper-parameters. We aim to find how different types of classifier perform in predicting relapses.

There are two major challenges in predicting relapses. First, we do not have a large amount of data to train a classification model. We have the most examples with 7-day prediction window, in which 1641 windows are non-relapses and 16 are relapses. Second and more importantly, our participants do not relapse frequently, therefore, our dataset is imbalanced. For example, only 0.97% of the 7-day predic-

tion windows are labeled as relapse. Training the classifiers without augmenting the dataset results in 0% of recall. Prediction in an imbalanced dataset require a large dataset. For example, credit card fraud detection is a similar prediction problem: fraud transactions are only a tiny part of all transactions. A real-world credit card transaction dataset [64] contains a subset of online transactions that occurred in two days, where 492 out of the 284,807 transactions are fraud. Although the dataset is extremely imbalanced, we can still achieve good prediction performance without any data preprocessing. A future work could collect a much larger dataset over many years to help building a reliable relapse predictor. In what follows, we discuss our method to address the challenges in detail.

Resample the training data. To reduce the data bias in the dataset (i.e., more non-relapses than relapses), we apply data resampling techniques to balance the dataset. Resampling techniques are widely used to address the bias in an imbalanced datasets (i.e., majority cases have higher weight than minority cases). We use Synthetic Minority Over-sampling Technique (SMOTE) [52] to balance the training set by over-sampling the minority class (i.e., relapse) and under-sampling the majority class (i.e., non-relapse). Instead of over-sample the minority classes by replication, SMOTE creates "synthetic" minority examples. The synthetic minority examples are generated from k-nearest neighbors of the existing minority examples [52]. We use 5-nearest neighbors to generate synthetic minority examples. SMOTE has shown to be more effective than simple under-sampling and over-sampling methods.

2-level 3-fold stratified cross validation and model selection. We design a new 2-level 3-fold stratified cross validation (CV) to evaluate the relapse prediction performance. The top level CV evaluates the prediction performance and the second

level CV selects model hyper-parameters.

In the top level CV, we first randomly partition the dataset into three folds, each fold has the same non-relapse to relapse ratio. We iteratively select one fold for testing and the other two folds for training the model. We use the second level CV on the training data to identify the optimal set of hyper-parameters that maximize the prediction performance metric - Precision Recall area under curve (PR AUC) [135]. PR AUC is a more informative metric than the ROC AUC when evaluating binary classifiers on imbalanced datasets [165].

In the second level CV, we use another 3-fold stratified cross validation on the training data from the top level CV to grid search the model hyper-parameter space. We resample the training data from the second level CV before training the model. We evaluate the trained model on the test data from the second level CV that have not been resampled and compute the PR AUC. Once a set of optimal hyper-parameters is identified, we apply the optimal hyper-parameters to train the classifier using the resampled training data from the top level CV and evaluate the prediction performance using the test data from the top level CV.

The 2-level 3-fold stratified cross validation has the following advantages. First, each fold in the cross validation has the same relapse to non-relapse ratios. Second, it separates test data from the model training and selection process. The prediction performance metrics are obtained in the top level CV with the original test data (i.e., not been resampled), whereas the model is trained with the resampled training data. Finally, the proposed cross validation method incorporated hyperparameter grid search in the model selection process.

We report the average precision, recall, F1 score, and PR AUC from each fold.

In order to avoid selecting a random seed that lead to impractically high or low prediction performance, we repeat the 2-level 3-fold stratified cross validation 5 times and report the average prediction performance metrics.

6.3 Results

In what follows, we discuss our relapse prediction results in detail. We first define our relapse prediction baseline, followed by the best results from each of the four classifiers (i.e., logistic regression, SVM with linear kernel, SVM with RBF kernel, and random forest). We then discuss how different classifier design considerations (i.e., using raw feature or standardized features, different data types, prediction window length, and PCA) affect prediction performance. Finally, we present features that are important to predict relapses.

6.3.1 Relapse prediction baseline

Because there is no prior work on using passive sensing to predict relapses in schizophrenia patients, we use random guessing as our prediction baseline. Specifically, we randomly label a case with either relapse or non-relapse with the same probability. Other simple prediction baselines (e.g., assign the same label to all examples) produce either 100% or 0% in recall, which is not informative than random guessing in this case because precision and recall are both important metrics in predicting relapses. We then compute precision (i.e., $\text{tp}/(\text{tp} + \text{fp})$, where tp is the number of true positives and fp is the number of false positives), recall (i.e., $\text{tp}/(\text{tp} + \text{fn})$, where fn is the number of false negatives), and F1 score (i.e., $2 \cdot (\text{precision} \cdot \text{recall})/(\text{precision} + \text{recall})$) for the

random labeled cases. The baseline performance is presented in Table 6.1.

Table 6.1: Relapse prediction baseline according to random guessing for a classification.

window length	number of non-relapse	number of relapse	precision	recall	F1
7	1641	16	0.010	0.500	0.019
15	861	18	0.020	0.500	0.039
21	578	20	0.033	0.500	0.063
30	411	19	0.044	0.500	0.081

6.3.2 Results overview

In what follows, we present the best prediction results (i.e., highest F1 score) from the four classifiers obtained using grid search. Table 6.2 shows corresponding the prediction windows length, the number of principle components, precision, recall, F1 score, and PR AUC that are associated with best prediction performance from each of the classifiers. Interestingly enough, all classifier achieve best F1 score using non-standardized data with 30-day prediction time window. We suspect behavioral patterns over a longer period (e.g., 30 days) are more indicative of future relapses. SVM with RBF kernel achieves the best F1 score amongst the four classifier using the first 100 principle components obtained from both sensing and EMA data. The precision is 26.8% and the recall is 27.4%. To put these numbers into perspective, there are 411 cases in the 30-day dataset, 19 of which are relapses. The classifier predicts 19 cases that are relapses, 5 of which are correct and 14 are incorrectly identified as relapse. 14 relapses are misclassified as non-relapse. Logistic regression and SVM with linear kernel achieve slightly worse F1 scores but higher recall. The logistic regression model achieves 35.8% of recall and 21.4% of precision. The SVM

with linear kernel achieves 32.6% of recall and 23.3% of precision. The random forest model achieves the worst F1 score, with 18.9% of recall and 28.1% of precision. All the four classifiers beat the baseline in term of the F1 score and precision. However, the recall is worse than the baseline.

Table 6.2: Best prediction results according to the F1 score

data type	**classifier**	**window length**	**number of PCs**	**precision**	**recall**	**F1**	**PR AUC**
sensing+ema	svm rbf	30	100	0.268	0.284	0.274	0.192
sensing+ema	logistic regression	30	50	0.214	0.358	0.265	0.224
sensing+ema	svm linear	30	50	0.233	0.326	0.262	0.225
sensing	random forest	30	25	0.281	0.189	0.223	0.178

In summary, SVM with RBF kernel, SVM with linear kernel, and logistic regression achieves similar relapse prediction performance whereas random forest achieves the worst performance. 30 days is the best time window to predict relapse. Combining passive sensing data from smartphones and self-report EMA responses helps predicting relapses. Standardizing every participant's data does not help improving the prediction performance. On the contrary, we get worse performance with standardized data. Using PCA to combine features and reduce the feature dimensionality helps improving the performance. We will discuss how using different data as predictors, prediction window, and PCA affects prediction performance in the following sections.

6.3.3 Prioritizing the recall

In the previous section, we present the best prediction results in term of the F1 score. The F1 score is the harmonic average of the precision and recall, which gives the same weight to precision and recall. However, misclassifying relapse as non-relapse may have severe consequences compared with misclassifying non-relapse as relapse. Misclassifying a relapse as non-relapse may lead to non-action (e.g., fail to deliver intervention) and miss the best opportunity to treat the patient, whereas misclassifying a non-relapse as relapse may lead to unnecessary clinical visits thus increased cost. In what follows, we present prediction results with the constraint that recall $\geq 50\%$. The results are presents in Table 6.3.

Table 6.3: Best prediction results according to the F1 score with recall $\geq 50\%$

data type	**classifier**	**window length**	**number of PCs**	**precision**	**recall**	**F1**	**PR AUC**
sensing+ema	svm linear	30	25	0.154	0.516	0.236	0.194
sensing+ema	svm rbf	30	50	0.140	0.537	0.208	0.184
sensing+ema	logistic regression	30	2	0.068	0.505	0.118	0.093
ema	random forest	21	1	0.055	0.562	0.100	0.059

SVM with linear and RBF kernels, and logistic regression achieve best F1 scores when recall $\geq$ 50% using both sensing and EMA data with 30-day time window whereas random forest achieves best result with 21-day window EMA data. All models beat the baseline in term of precision, recall, and F1 score. The random forest model achieves best F1 score using only EMA data as predictors with 21-day prediction window. However, the performance is only slightly better than the 21-day

window baseline.

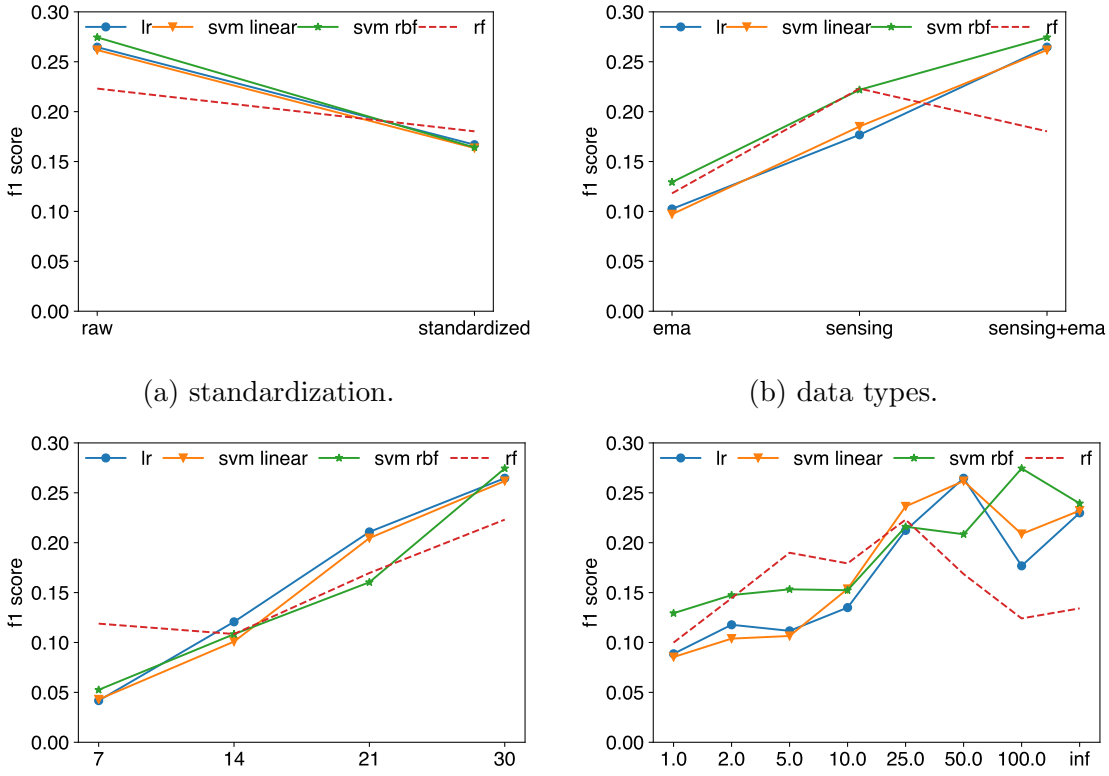
We discuss the result from SVM with linear kernel in detail. The precision of the model is 15.4% and the recall is 51.6%. The classifier predicts 64 cases that are relapses, 10 of which are correct and 54 are incorrectly identified as relapse. 9 relapses are misclassified as non-relapse. Compared with the result with the best F1 score, the model correctly identifies more relapses with the cost of more false positives.

The result shows that in practice, we can bias our models to be more sensitive to relapses with the cost of more false positives. To do so, we could tweak our models by assigning different weights to precision and recall thus obtain a more desirable relapse prediction model.

6.3.4 Prediction performance analysis.

In what follows, we discuss how different model decisions affect the relapse prediction performance. Specifically, we focus on whether or not we standardize each participant's features, what types of data are included in the prediction model (i.e., EMA, sensing, and both EMA and sensing), what prediction window we use (i.e., 7-day, 14-day, 21-day, or 30-day), and whether or not we apply PCA to transform the data and how many PCs we should use if PCA is applied.

Per-participant standardization transforms a participant's passive sensing features and EMA responses according to their first 30 days' data. Specifically, we first compute the mean μ_{30} and standard deviation σ_{30} for each of the features in the first 30 days, then we transform the feature as follows: $v_t = (v - \mu_{30})/\sigma_{30}$, where v is the original feature vector and v_t is the transformed feature vector. Figure 6.2(a) shows the best F1 scores obtained from four classifiers with or without per-participant



(a) standardization.

(b) data types.

(c) prediction window length.

(d) number of PCs. inf: no PCA applied.

Figure 6.2: Predictin F1 score from different models.

standardization. Applying per-participant standardization leads to worse F1 scores. Logistic regression and SVM models show overall similar F1 scores, where the F1 scores decrease from about 0.256 to 0.169 after applying per-participant standardization. The results show that per-participant standardization does not improve the prediction performance. We suspect the absolute behavioral levels (e.g., sleep duration), which are eliminated by standardization, are helpful in predicting relapse.

Data types. We inspect how EMA responses and passive sensing data help predicting relapses. Figure 6.2(b) shows the best F1 scores obtained from four classifiers with three settings: predict using only EMA responses, predict using only passive

sensing data, and predict using both EMA responses. All classifiers perform poorly using only EMA responses, where the F1 scores are around 0.1. However, the F1 scores significantly improve when we use passive sensing data for prediction, where SVM with RBF kernel achieves the best performance with $F1 = 0.222, \text{precision} = 22.2\%, \text{recall} = 23.2\%$. Random forest achieves similar prediction performance. Logistic regression and SVM with linear kernel, however, perform worse than RBF SVM and random forest. We suspect the non-linearity of the RBF kernel and random forest helps reducing under-fitting. *We achieve the best prediction performance by combining both EMA and passive sensing.* SVM with RBF kernel achieves the best performance with $F1 = 0.274, \text{precision} = 26.8\%, \text{recall} = 28.4\%$. Logistic regression and linear SVM achieve slightly worse performance whereas random forest achieve a worse F1 score compared with using only sensing data. In summary, we can predict relapses more accurately using passive sensing data compared with only using self-report EMA. Combining passive sensing data and EMA self-reports further improves the prediction performance.

Prediction window length. Figure 6.2(c) shows the best F1 scores obtained from four classifiers with four different prediction window settings: 7-day, 14-day, 21-day, and 30-day. The F1 scores increase for all classifiers as we increase the window length from 7 days to 30 days. We suspect behavioral patterns over a longer period are more indicative of future relapses. Therefore, we find better prediction results with longer prediction windows. However, increasing the prediction window length reduces the number of examples available for training and testing prediction models. Take the RBF SVM as an example, the classifier achieves $F1 = 0.053$ with 7-day window, which is 0.034 higher than the baseline shown in Table 6.1, whereas it achieves $F1 = 0.274$

with 30-day window, which is 0.193 higher than the baseline. We suspect summarizing behavioral features in shorter windows leads to more noise in the feature data because of the short term behavior changes whereas longer windows smooth the behavioral data so that the features captures participants' behaviors more accurately.

PCA. Figure 6.2(d) shows the best F1 scores obtained from four classifiers with different PCA settings. As we include more principle components (PCs) in the predictions, the F1 scores increase for all classifiers. Two linear classifiers, linear SVM and logistic regression, achieve best F1 scores when using 50 PCs, which is higher than using the raw feature data without PCA transformation. RBF SVM achieves best F1 score when using 100 PCs, which again is higher than using the raw feature data. Random forest achieves the best F1 score when using 25 PCs, however, the prediction performance of random forest is worse than other three classifiers. The results show transforming the features using PCA reduces the feature dimensionality and generates more useful features by combining different features together. We discuss particular PCs in later sections.

6.3.5 Useful Features.

In what follows, we present features selected by L1 regularization in logistic regression training. We do not transform the features using PCA so that we can interpret how features are related to relapse. We choose to present logistic regression coefficients instead of other classifiers because it is easier to interpret parameters in logistic regression - positive coefficients indicate positive correlations whereas negative coefficients indicate negative correlations. We first present features selected by the model using both sensing and EMA data, then we present features selected by the model

using only sensing data. The selected features and their regression coefficients are presented in Table 6.4.

Table 6.4: L1 regularization selected features in logistic regression.

sensing+EMA precision=22.6%, recall=36.8%		sensing precision=12.5%, recall=42.1%	
feature	coeff	feature	coeff
conversation duration morning	2.631	conversation duration morning	0.659
on foot duration evening	2.553	number of voice frames night	-0.319
number of voice frames morning	2.540	number of calls made	-0.251
visit education places	-2.139	on foot duration evening	0.193
number of visited places evening	-1.952	visit parks and outdoors places	0.084
EMA item median response time	1.876	duration of calls made in the morning	-0.008
EMA_seeing_things	-1.650		
audio amplitude afternoon	-1.620		
visit travel places	-1.597		
visit residence places	-1.506		

Sensing and EMA. The logistic regression model achieves 22.6% of precision and 36.8% of recall using both sensing and EMA data. The prediction window is 30 days. The l1 regularization selects 81 out of 144 features in training. We present the top 10 features with the largest absolute coefficients. We find that participants who have more conversations in the morning, walk more in the evening but visit fewer places in the evening, visit less educational, travel, and residential places, report lower score in seeing things, but spend more time responding to EMAs are more likely to relapse. Please note, the L1 regularization also selects 5 EMA items (i.e., depressed, calm, voices, think, harm) and positive scores to predict relapses. Specifically, participants who report higher scores in depressed, voices, and harm items, and lower scores in calm, think, and positive score are more likely to relapse.

Sensing only. The logistic regression model achieves 12.5% of precision and 42.1% of recall using both sensing and EMA data. The prediction window is 30 days. The l1 regularization selects 6 out of 130 features in training. Specifically, participants who have more conversations in the morning, spend more time walking in the evening, visit more parks and outdoor places, makes fewer phone calls are more likely to relapse.

6.3.6 Behavioral Principle Components

In what follows, we present the top 5 PCs with the largest absolute logistic regression coefficients. The top 5 PCs, their regression coefficients, and characteristics.

PC 1 describes a behavioral pattern in which participants spend less time responding to EMAs, report lower scores in all EMA items, and makes more phone calls. *Participants whose behaviors are more similar to PC 1 are less likely to relapse.*

PC 19 describes a behavioral pattern in which participants visit more places, spend less time at nightlife, arts and entertainment, parks and outdoor, and gym places, spend more time at residence, medical and education places, receive more phone calls but do not make many calls, have more conversation during the evening, and spend more time responding to EMA questions. *Participants whose behaviors are more similar to PC 19 are more likely to relapse.*

PC 7 describes a behavioral pattern in which participants make and receive more calls, have less conversation in the morning but visit more places in the evening. *Participants whose behaviors are more similar to PC 19 are less likely to relapse.*

PC 36 describe a behavioral pattern in which participants visit less places related to medical, gym, library but visit more places relate to arts and entertainment, home,

Table 6.5: Characteristics of principle components with largest absolute regression coefficients. A positive coefficient indicate the principle component is positively correlated with relapses (i.e., larger PC weight indicates higher probability of relapse).

PC	coefficient	features
1	-437.7	low EMA item scores, respond EMAs fast, more phone calls especially in the evening and night, ride bikes
19	395.8	visit more places; visit less places relate to nightlife, arts and entertainment, parks and outdoor, and gym; visit less places relate to residence, medical and education; respond EMAs slow; receive more phone calls but make fewer phone calls in the afternoon; have more conversations in the evening
7	-348.0	make and receive more calls, have less conversation in the morning, visit more places in the evening, ambient light is bright at night.
36	307.0	visit less places relate to medical, gym, and library; visit more places relate to arts and entertainment, home, and residence; report higher EMA score in items including hearing voices, harm, and feel less hopeful
8	305.7	visit more places relate to medical; more SMS use at nigh; more conversation at night; fewer phone calls in the morning; bright at night; more phone use at night; wake up late.

residence. They report high scores in hearing voices, harm, and feel less hopeful. *Participants whose behaviors are more similar to PC 19 are more likely to relapse.*

PC 8 describe a behavioral pattern in which participants visit more places related to medical, have more SMS use, phone use, and conversations at night, less calls in the

morning, bright at night, and wake up late. *Participants whose behaviors are more similar to PC 19 are more likely to relapse.*

6.4 Discussion and Conclusion

The CrossCheck system discussed in this thesis shows promise in using mobile phones and passive sensing to predict schizophrenia relapses. In what follows, we discuss our results, limitations, and future work.

6.4.1 Relapse classifier design considerations.

Our results show that per-participant standardization fails to improve the prediction performance. The standardized feature values indicate how many standard deviations the true value is from the feature mean value, which measures the within-individual behavior differences. Applying per-participant standardization for relapse prediction assumes similar deviations from a participant's average behaviors across all participants account for relapse. However, our results show that applying per-participant standardization leads to poorer prediction performance, *which may indicate that the between-individual differences are more predictive of relapse than within-individual differences.*

We find the 30-day time window is the best time window to predict relapse. Summarizing behavioral features in shorter windows leads to more noise in the feature data because of the short term behavior changes whereas longer windows smooth the behavioral data. Also, the pre-relapse behavioral changes might be gradual and the behavioral changes may start many days before the relapse. A shorter time window

may lead to lower resolution of changes in behaviors (i.e., behaviors in consecutive time windows are similar despite one is relapse the the other one is not) thus more challenging for the classifiers to identify relapse signals.

We show that using passive sensing data greatly improves the prediction performance compared with using only self-report EMA. Combining both passive sensing and EMA further improves the prediction performance. *Our results indicate that passive sensing data has the potential to unobtrusively monitor and predict relapse in the future.*

Semantic location features, phone calls, and conversational features are good predictors of relapse. Transforming the behavioral feature space using PCA provides interpretable behavioral patterns. Our findings show that behavioral features that are more interpretable are more likely to be indicative of relapses. Future work should focus on designing features that are interpretable and capture people's higher level behaviors (e.g., go to work, socializing with friends, exercising) by combining different sensor streams. Exploring behavioral patterns (e.g., *behavioral principle components*) would further gives more insight to relapse. It is important that the results are interpretable by clinicians.

Misclassifying relapse as non-relapse may have severe consequences compared with misclassifying non-relapse as relapse. Misclassifying a relapse as non-relapse may lead to non-action (e.g., fail to deliver intervention in time) and miss the best opportunity to treat the patient, whereas misclassifying a non-relapse as relapse may lead to unnecessary clinical visits thus increased cost. We show that our models achieve 53.7% of recall with the cost of lower precision, which is still better than the baseline defined in Section 6.3.1. *For practical use, we need to carefully evaluate the precision*

recall tradeoff and select the best prediction model that maximize recall (i.e., identify as many patients at risk as possible) while minimizing the unnecessary cost due to misclassifying non-relapses as relapses.

6.4.2 Limitations, future work and conclusion remark

The CrossCheck relapse prediction system presented in this thesis shows promise in using mobile phones and passive sensing to predict schizophrenia relapses. The system and models show reasonable 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 schizophrenia relapses.

We also recognize limitations of our work. Our dataset only had 27 relapses from 20 participants. Therefore, the current dataset is small and imbalanced. Our small dataset poses many challenges to build a relapse predictor. First, over-fitting the training data becomes much harder to avoid. Because the data was collected from a small number of participants (i.e., 61 participants) over short period (i.e., 1 year), there could be high correlations between examples. Therefore, the model could also over-fit to the test data as well. Second, outliers become much more dangerous because they have more weight (higher outliers to normal data points ratio) to skew the model. Any noise in general becomes a real issue because we might have enough data to filter out the noise in training. The imbalanced nature of our relapse dataset makes building the prediction model even more challenging because prediction models have even less data to learn the schizophrenia relapse behavioral characteristics. To address such limitations, we applied re-sampling and missing data imputations. However, any data augmentation would introduce bias to the dataset.

We carefully designed a 2-level 3-fold crossvalidation to account for the potential bias introduced by the re-sampling and imputation, however, we need to validate our method on more data to reaffirm our findings.

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 solutions 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 for reasons of model performance.

Future work may improve predicting relapse in schizophrenia patients in the following areas. First, future studies could aim to collect data from more patients over a much longer time (e.g., 5 years). Therefore, we would have more data to build a more robust relapse predictor. Second, future studies could incorporate new sensing technologies in assessing relapse risk. For example, researchers could include wearable in the study to measure patients' physiological signals, which might be more indicative of impending relapses. Finally, researchers could look into other relapse modeling techniques. For example, researchers could treat relapse prediction as an anomaly detection problem. That is, a relapse manifest itself as a significant change in behavioral norms. Anomaly detection treat the data as time series. Each participant has their own behavioral data time series. Anomaly detection usually learns

the time series patterns and make prediction on the same time series. Therefore, such modeling technique require a large longitudinal dataset. Researchers also need to consider how to incorporate data from different individuals to build the prediction model.

We presented and evaluated different prediction model design considerations and found that linear models (e.g., logistic regression and linear SVM) using PCA-transformed passive sensing and self-report EMA features best predict relapses with 30-day time window. We discussed features and behavioral patterns that are predictive with relapses. Although our prediction performance might not be good enough to use in clinical practices, our results show promises in using passive sensing to help clinicians better identifying patients at risks of relapses.

Chapter 7

Conclusion

We have witnessed a growing number of studies using smartphones to study mental health. Smartphones provide a feasible and unobtrusive method to continuously collect behavioral data from people. Mental health and psychology researchers have begun to use smartphones to assess depression, bipolar disorders, anxiety, schizophrenia, post traumatic stress disorders, and personality. In the first part of this thesis, we built the StudentLife sensing system and used the system to collect data from college students. We showed that we could use the smartphone data collected by the StudentLife sensing system to predict academic performance and mental health. We inferred a number of behaviors that are closely related to students' life on campus. These behaviors are closely related to students' academic performance and mental health. The results from the StudentLife study encouraged us to apply the same mental health sensing technology in a more challenging community: people with serious mental illnesses. In the second part of this thesis, we presented the CrossCheck study, a RCT aims to track schizophrenia patients' symptoms and predict impending relapses. We built the CrossCheck sensing system based on the StudentLife sensing

system and deployed the system to people with schizophrenia. During the study, we used the smartphone data to identify and reach out to patients at-risk.

7.1 Insights

This thesis addressed a number of technical challenges related to mental health sensing. We built power efficient smartphone sensing systems, modeled human behaviors from the sensing data, built behavioral features that capture students' life and depression symptoms, built schizophrenia symptom prediction system, and evaluated models to predict schizophrenia relapses. We focused on finding relationship between passive sensing data and mental health, and used smartphone sensing data to predict people's mental wellbeing. We explored new research questions that were not addressed before, specifically: 1) How can we infer college students' on campus activities (e.g., studying, partying, been distracted) from smartphone data? 2) Can we learn college students' term behavioral trends from their smartphone data? 3) How to use smartphone sensing to assess college students' mental health, and academic performance? 4) What depression symptom features can we derive from smartphone sensor streams? 5) How to use smartphones to assess schizophrenia patients' symptoms, and how can we use the data for reaching out patients at risk, and 6) Can we predict impending schizophrenia relapses?

The contribution of this thesis are summarized as follows:

First, we presented the StudentLife Android sensing system and a longitudinal study using the system to assess mental health, academic performance and behavioral trends of a student body. We collected a large number of smartphone data, a number of validated mental health measurements, and GPAs from 48 Dartmouth

students over a 10-week term in 2013. We discussed behavioral features that are specifically designed to capture students' life on campus. The features incorporate multiple sensor streams from smartphones and our knowledge about the campus. We observed trends in the sensing data, termed the *Dartmouth term lifecycle*. The trends show how students' stress, positive affect, conversation levels, sleep, and daily activity patterns change as the term progresses and the workload increases. We identified correlations between automatic sensing data and a broad set of well-known mental well-being measures. We proposed for the first time a model that can predict a student's cumulative GPA using automatic behavioral sensing data from smartphones. The results showed great potential in assessing people's mental health using smartphones. We presented a follow-up study, in which we upgraded the StudentLife sensing system to support iPhones and Microsoft Band 2. We collected smartphone and wearable data from 83 undergraduate student across two 9-week terms during the winter and spring term in 2016. We proposed a set of passive sensor based symptom features derived from phones and wearables that we hypothesized proxy 5 out of the 9 major depressive disorder symptoms defined in DSM-5. We identified a number of correlations between the symptom features and PHQ-8, and we showed that we could predict PHQ-4 and PHQ-8 using the proposed symptom features.

Second, we presented CrossCheck, a year long randomized control trial (RCT)[50] conducted in collaboration with a large psychiatric hospital in New York City, NY, which aimed to track symptoms in people with schizophrenia and predict impending relapses. We recruited 61 participants in the smartphone arm from 2014 to 2016, in which participants carried study phones with our sensing app built on the StudentLife core sensing system. The data collection phase concluded in June 2017.

7.1 Insights

We identified 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. We presented and evaluated models that predict participants' aggregated ecological momentary assessment (EMA) scores that measure several dynamic dimensions of mental health and functioning in people with schizophrenia. We found 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. We built the CrossCheck symptom prediction system, which was 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. The CrossCheck symptom prediction system predicted participants' BPRS scores each week and sent the predictions to our research staff. Our research staff used the predictions to determine whether or not a participant is at risk and reach out when a participant is determined as at-risk. We investigated the feasibility of using passive sensing data and/or self-report EMAs to predict relapses. We discussed the challenges to build a relapse prediction system and solutions to the challenges. We evaluated different methods to predict relapses, including different prediction window setups, feature space transformations, training data resampling, missing data imputation, and four different binary classifiers (i.e., linear SVM, RBF SVM, logistic regression, and random forest). We presented a new 2-level 3-fold cross-validation method, which combines training data resampling and hyper-

parameter selection. The 2-level 3-fold cross-validation method is a robust method to evaluate relapse predictions. We showed that we can predict whether or not a participant is going to relapse with precision = 26.8%, recall = 28.4%. We also showed that we could tweak the model to maximize recall with a slightly reduced precision with precision = 15.4%, recall = 51.6%. Our results show potentials in using mobile phones and passive sensing to predict schizophrenia relapses. However, there are still a lot of work to be done to make the schizophrenia relapse prediction system a viable tool for clinicians. For example, a major challenge of developing an accurate relapse predictor is that we do not have a large dataset that contains enough relapse cases. Future studies could aim to collect data from more patients over a much longer time (e.g., 5 years). If a large dataset is available, researchers could apply other relapse modeling techniques. For example, researchers could treat relapse prediction as an anomaly detection problem. That is, a relapse manifest itself as a significant change in behavioral norms. Anomaly detection treat the data as time series. Each participant has their own behavioral data time series. Anomaly detection usually learns the time series patterns and make prediction on the same time series. Therefore, such modeling technique require a large longitudinal dataset. Researchers also need to consider how to incorporate data from different individuals to build the prediction model.

Finally, we released the StudentLife dataset to the research community and we will release our StudentLife core sensing system to help future studies in assessing mental health using smartophones.

7.2 Future Work

There are many ways our work can be improved and extended. In what follows, we discuss what areas future studies could work on.

Large scale studies with more diverse populations. The scale of our studies is small. The studies presented in this thesis recruited fewer than 100 participants. Furthermore, our participants are either college students from the same university or schizophrenia patients living in the same urban area. Smaller number of participants (i.e., small sample size) decreases statistical power, which may lead to more false negative findings (i.e., Type II errors, failing to reject a false null hypothesis). Findings from studies with homogeneous participants may not apply to a more general population. We need to conduct larger scale studies with more diverse participants (e.g., people live in different regions, different occupation) to confirm our findings.

Consolidating the sensing technology. The research community and the industry have made a lot of progress in advancing sensing technology. However, more work can be done to further consolidate the sensing technology. Microphone is a power sensor on the phone. We use microphone to infer whether or not a user is around conversations. However, we cannot determine whether the user is involved in the conversation or not. Future work could improve conversation detection to detect whether the user is speaking (i.e., speaker identification), how many people the user talked to, and the user's emotion (i.e., emotion detection from speech). Future work could also work on detecting a user's acoustic surroundings (e.g., speech, music, factory). Future researchers could also incorporate new technology (e.g., wearables) for mental health sensing. Wearables provide better ways to measure physical activities, sleep, and physiological signals. However, wearables from different vendors may gen-

erate measurements with different accuracy. Researchers in the future could work on building systems that make consistent measurements across different wearables.

Behavioral modeling. Making sense of the inferred behaviors is the key to mental health sensing. Clinicians need to be able to interpret the meaning of features. In this thesis, we presented many daily features derived from the smartphone data inferred behaviors. Future researchers could focus on modeling people's behaviors from different angels. For example, we could compute behavioral lifestyles (e.g., people who are early to bed, early to rise and have a regular day schedule) using multiple sensor streams. Such behavioral patterns/features are more interpretable and may be stronger indicators of mental health.

Mental health prediction models. A lot of work need to be done to advance machine learning models that predict mental health from smartphone and wearable data. In this thesis, we focus on traditional prediction models (e.g., logistic regression, random forest, SVM, gradient boosting). We would also explore applying deep learning models for mental health prediction. Deep learning models have revolutionized many fields (e.g., computer vision, speech recognition). Its capability of learning good feature representations automatically from the data and make accurate predictions is the perfect fit for mental health prediction from smartphone sensing, where we have high dimensional data and feature engineering is challenging. However, we need to address challenges in applying deep learning models for mental health prediction. First, deep learning models need large data sets are needed to make sure the models are trained properly (i.e., achieve good prediction performance). Therefore, we need to conduct large scale studies with a large number of participants. The studies should last long enough to guarantee enough data needed for training the models. Furthermore, it

is difficult to determine how much data is needed to properly train a deep learning model. Typically, the volume of data required is determined by the complexity of the model (e.g., how many layers in the model, the number of input features, the number of classes). Second, training a deep learning model usually require high-performance hardware and the deep learning models might not be able to run on the phone. Finally, Deep learning models are essentially a blackbox. It is difficult to interpret the predictions made by the models (e.g., why the model determines a person is going to relapse).

New mental health classification paradigm. There are different approaches that can be taken to address the mental health classification and detection problem. The National Institute of Mental Health (NIMH) has launched the Research Domain Criteria (RDoC) [110, 120] project to create a framework for studying mental disorders. The RDoC framework centers around dimensional psychological constructs that are relevant to human behavior and mental disorders [110, 120]. The psychological constructs include negative valence systems, positive valence systems, cognitive systems, systems for social processes, arousal/regulatory systems. RDoC proposes to measure the systems using molecular, genetic, neurocircuit and behavioral assessments [110, 120]. We imagine that future mobile sensing approaches for mental health assessment could focus on developing new sensing modalities and physiological and behavioral features to predict the RDoC constructs.

Technology acceptance by clinicians. We need to collaborate more with clinicians to best understand how our system could inform treatment. We need to advance sensing and mental health prediction technology to provide robust and accurate mental health predictions. Clinicians should be able to interpret the mental health

outcome predictions and understand what behavioral changes lead to such predictions.

Mental health interventions. The ultimate goal of mental health sensing using smartphones is to keep people healthy and prevent mental illness relapse. Future researcher could investigate behavioral changes that are precursor to worse mental health states. Researchers could develop realtime symptom monitoring systems and provide interventions when adverse behaviors are observed. But researcher should develop validated intervention constructs first. For example, what behavioral changes and inferred symptom changes warrant interventions, and what interventions should be applied. Our work on weekly BPRS prediction and reaching out to participant-at-risk is a step toward this goal.

7.3 Final Comment

The contributions made by this thesis push the boundaries of how researchers would use smartphones to continuously and unobtrusively monitor people's mental health. The chapters of this thesis collectively provide a picture of the potential for modeling human behaviors from smartphone and wearable sensing data. Through the contributions of the StudentLife study and CrossCheck study, we have laid a solid foundation for further exploration of applying smartphone sensing in mental health and other disciplines. Since the StudentLife study described in Chapter 2 published in 2014, there are a growing number of studies applied the similar methodology described in this thesis to investigate using smartphones to assess mental illnesses, personality traits, mood, academic performance, and work performance. The CampusLife project aims to collect data from campus communities. The project collects data from mobile and

wearable devices and social media. The goal of the CampusLife project is to understand wellness for young adults, as well as how to perform such experimentation. We hope this thesis not only provides building blocks for future research but also acts as a useful guide to identify new directions in mental health sensing using smartphones.

We still have a long way to go to make mental health sensing using smartphone phones a reality. Our research will turn people's smartphones and wearable devices into mental health monitors and keep people mentally health. I strongly believe that smartphones sensing technology will eventually evolve into facilitate better mental health care. In this thesis, we have presented some early steps toward this goal.

Chapter 8

Refereed Publications as a Ph.D. Candidate

My refereed publications as a Ph.D. candidate are listed below. Work in preparation and technical reports are omitted.

Conference/Workshop Publications

Chuang-Wen You, Nicholas D. Lane, Fanglin Chen, Rui Wang, Zhenyu Chen, Thomas J. Bao, Martha Montes-de-Oca, Yuting Cheng, Mu Lin, Lorenzo Torresani, and Andrew T. Campbell. 2013. CarSafe app: alerting drowsy and distracted drivers using dual cameras on smartphones. In Proceeding of the 11th annual international conference on Mobile systems, applications, and services (MobiSys '13). ACM, New York, NY, USA, 13-26.

Zhenyu Chen, Mu Lin, Fanglin Chen, Nicholas D. Lane, Giuseppe Cardone, Rui Wang, Tianxing Li, Yiqiang Chen, Tanzeem Choudhury, and Andrew T. Campbell. 2013. Unobtrusive sleep monitoring using smartphones. In Proceedings of the 7th

International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth '13). ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), ICST, Brussels, Belgium, Belgium, 145-152.

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.

Fanglin Chen, Rui Wang, Xia Zhou, and Andrew T. Campbell. 2014. My smartphone knows i am hungry. In Proceedings of the 2014 workshop on physical analytics (WPA '14). ACM, New York, NY, USA, 9-14.

Rui Wang, Gabriella Harari, Peilin Hao, Xia Zhou, and Andrew T. Campbell. 2015. SmartGPA: how smartphones can assess and predict academic performance of college students. In Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15). ACM, New York, NY, USA, 295-306.

Rui Wang, Andrew T. Campbell, and Xia Zhou. 2015. Using opportunistic face logging from smartphone to infer mental health: challenges and future directions. In Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers (UbiComp/ISWC'15 Adjunct). ACM, New York, NY, USA, 683-692.

Sophia Haim, Rui Wang, Sarah E. Lord, Lorie Loeb, Xia Zhou, and Andrew T. Campbell. 2015. The mobile photographic stress meter (MPSM): a new way to measure stress using images. In Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers (UbiComp/ISWC'15 Adjunct). ACM, New York, NY, USA, 733-742.

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.

Gabriella M. Harari, Weichen Wang, Sandrine R. Müller, Rui Wang, and Andrew T. Campbell. 2017. Participants' compliance and experiences with self-tracking using a smartphone sensing app. In Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers (UbiComp '17). ACM, New York, NY, USA, 57-60.

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. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 1, 3, Article 110 (September 2017), 24 pages.

Rui Wang, Weichen Wang, Alex daSilva, Jeremy F. Huckins, William M. Kelley, Todd F. Heatherton, and Andrew T. Campbell. 2018. Tracking Depression Dynamics in College Students Using Mobile Phone and Wearable Sensing. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 2, 1, Article 43 (March 2018), 26 pages.

Journal Publications

Ben-Zeev D, Scherer EA, Wang R, Xie H, Campbell AT. Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health. Psychiatric rehabilitation journal. 2015 Sep;38(3):218.

Harari, G. M., Gosling, S. D., Wang, R., and Campbell, A. T. (2015) Capturing Situational Information with Smartphones and Mobile Sensing Methods. Eur. J. Pers., 29: 509–511.

Ben-Zeev, D., Scherer, E.A., Wang, R., Xie, H. and Campbell, A.T., 2015. Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health. Psychiatric rehabilitation journal, 38(3), p.218.

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., 2015. Mobile behavioral sensing for outpatients and inpatients with schizophrenia. Psychiatric services, 67(5), pp.558-561.

Harari, G.M., Lane, N.D., Wang, R., Crosier, B.S., Campbell, A.T. and Gosling, S.D., 2016. Using smartphones to collect behavioral data in psychological science:

opportunities, practical considerations, and challenges. Perspectives on Psychological Science, 11(6), pp.838-854.

Harari, G.M., Gosling, S.D., Wang, R., Chen, F., Chen, Z. and Campbell, A.T., 2017. Patterns of behavior change in students over an academic term: A preliminary study of activity and sociability behaviors using smartphone sensing methods. Computers in Human Behavior, 67, pp.129-138.

Ben-Zeev, D., Scherer, E.A., Brian, R.M., Mistler, L.A., Campbell, A.T. and Wang, R., 2017. Use of multimodal technology to identify digital correlates of violence among inpatients with serious mental illness: a pilot study. Psychiatric services, 68(10), pp.1088-1092.

Harari, G.M., Müller, S.R., Mishra, V., Wang, R., Campbell, A.T., Rentfrow, P.J. and Gosling, S.D., 2017. An Evaluation of Students' Interest in and Compliance With Self-Tracking Methods: Recommendations for Incentives Based on Three Smartphone Sensing Studies. Social Psychological and Personality Science, 8(5), pp.479-492.

Ben-Zeev, D., Brian, R., Wang, R., Wang, W., Campbell, A.T., Aung, M.S., Merrill, M., Tseng, V.W., Choudhury, T., Hauser, M. and Kane, J.M., 2017. CrossCheck: Integrating self-report, behavioral sensing, and smartphone use to identify digital indicators of psychotic relapse. Psychiatric rehabilitation journal, 40(3), p.266.

Book Chapter

Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., Zhou, X., Ben-Zeev, D. and Campbell, A.T., 2017. StudentLife: Using smartphones to assess mental

health and academic performance of college students. In Mobile Health (pp. 7-33). Springer, Cham.

Bibliography

[1] *CS65 Smartphone Programming*, http://www.cs.dartmouth.edu/~campbell/cs65/cs65.html.

[2] *Dartmouth College Weekly Schedule Diagram*, http://oracle-www.dartmouth.edu/dart/groucho/timetabl.diagram.

[3] *Depression*, http://www.nimh.nih.gov/health/topics/depression/index.shtml.

[4] *funf-open-sensing-framework*, https://code.google.com/p/funf-open-sensing-framework/.

[5] *PACO*, https://code.google.com/p/paco/.

[6] *StudentLife Dataset 2014*, http://studentlife.cs.dartmouth.edu/.

[7] *SurveyMonkey*, https://www.surveymonkey.com/.

[8] Saeed Abdullah, Mark Matthews, Ellen Frank, Gavin Doherty, Geri Gay, and Tanzeem Choudhury, *Automatic detection of social rhythms in bipolar disorder*, Journal of the American Medical Informatics Association **23** (2016), no. 3, 538–543.

[9] Saeed Abdullah, Mark Matthews, Elizabeth L Murnane, Geri Gay, and Tanzeem Choudhury, *Towards circadian computing: early to bed and early to rise makes some of us unhealthy and sleep deprived*, Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing, ACM, 2014, pp. 673–684.

[10] Saeed Abdullah, Elizabeth L Murnane, Mark Matthews, and Tanzeem Choudhury, *Circadian computing: sensing, modeling, and maintaining biological rhythms*, Mobile health, Springer, 2017, pp. 35–58.

[11] Nadav Aharony, Wei Pan, Cory Ip, Inas Khayal, and Alex Pentland, *Social fMRI: Investigating and shaping social mechanisms in the real world*, Pervasive and Mobile Computing **7** (2011), no. 6, 643–659.

[12] Carolyn M Aldwin, *Stress, coping, and development: An integrative perspective*, Guilford Press, 2007.

[13] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Julien Epps, Matthew Hyett, Gordon Parker, and Michael Breakspear, *Multimodal depression detection: Fusion analysis of paralinguistic, head pose and eye gaze behaviors*, IEEE Transactions on Affective Computing (2016).

[14] American College Health Association, *American college health association-national college health assessment ii: Reference group executive summary fall 2016*, Hanover, MD: American College Health Association (2016).

[15] Apple, *Core motion*, 2017, https://developer.apple.com/reference/coremotion.

[16] Haya Ascher-Svanum, Baojin Zhu, Douglas E Faries, David Salkever, Eric P Slade, Xiaomei Peng, and Robert R Conley, *The cost of relapse and the predictors of relapse in the treatment of schizophrenia*, BMC psychiatry **10** (2010), no. 1, 2.

[17] Daniel Ashbrook and Thad Starner, *Using gps to learn significant locations and predict movement across multiple users*, Personal and Ubiquitous computing **7** (2003), no. 5, 275–286.

[18] American Psychiatric Association et al., *Diagnostic and statistical manual of mental disorders (dsm-5®)*, American Psychiatric Pub, 2013.

[19] Min Aung, Faisal Alquaddoomi, Cheng-Kang Hsieh, Mashfiqui Rabbi, Longqi Yang, JP Pollak, Deborah Estrin, and Tanzeem Choudhury, *Leveraging multimodal sensing for mobile health: a case review in chronic pain*, IEEE Journal of Selected Topics in Signal Processing **10** (2016), no. 5, 1–13.

[20] Sunlee Bang, Minho Kim, Sa-Kwang Song, and Soo-Jun Park, *Toward real time detection of the basic living activity in home using a wearable sensor and smart home sensors*, Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE, IEEE, 2008, pp. 5200–5203.

[21] Nikola Banovic, Tofi Buzali, Fanny Chevalier, Jennifer Mankoff, and Anind K Dey, *Modeling and understanding human routine behavior*, Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, ACM, 2016, pp. 248–260.

[22] Ling Bao and Stephen S Intille, *Activity recognition from user-annotated acceleration data*, International Conference on Pervasive Computing, Springer, 2004, pp. 1–17.

[23] Ian Barnett, John Torous, Patrick Staples, Luis Sandoval, Matcheri Keshavan, and Jukka-Pekka Onnela, *Relapse prediction in schizophrenia through digital phenotyping: a pilot study*, Neuropsychopharmacology (2018), 1.

[24] P Bech, N-A Rasmussen, L Raabæk Olsen, V Noerholm, and W Abildgaard, *The sensitivity and specificity of the major depression inventory, using the present state examination as the index of diagnostic validity*, Journal of affective disorders **66** (2001), no. 2, 159–164.

[25] Aaron T Beck, David Guth, Robert A Steer, and Roberta Ball, *Screening for major depression disorders in medical inpatients with the beck depression inventory for primary care*, Behaviour research and therapy **35** (1997), no. 8, 785–791.

[26] Aaron T Beck, Robert A Steer, Gregory K Brown, et al., *Beck depression inventory*, (1996).

[27] Dror Ben-Zeev, *Mobile technologies in the study, assessment, and treatment of schizophrenia*, Schizophrenia bulletin (2012), sbr179.

[28] Dror Ben-Zeev, Christopher J Brenner, Mark Begale, Jennifer Duffecy, David C Mohr, and Kim T Mueser, *Feasibility, acceptability, and preliminary efficacy of a smartphone intervention for schizophrenia*, Schizophrenia bulletin (2014), sbu033.

[29] Dror Ben-Zeev, Kristin E Davis, Susan Kaiser, Izabela Krzsos, and Robert E Drake, *Mobile technologies among people with serious mental illness: opportunities for future services*, Administration and Policy in Mental Health and Mental Health Services Research **40** (2013), no. 4, 340–343.

[30] Dror Ben-Zeev, Susan M Kaiser, Christopher J Brenner, Mark Begale, Jennifer Duffecy, and David C Mohr, *Development and usability testing of focus: A smartphone system for self-management of schizophrenia.*, Psychiatric rehabilitation journal **36** (2013), no. 4, 289.

[31] Dror Ben-Zeev, Gregory J McHugo, Haiyi Xie, Katy Dobbins, and Michael A Young, *Comparing retrospective reports to real-time/real-place mobile assessments in individuals with schizophrenia and a nonclinical comparison group*, Schizophrenia bulletin **38** (2012), no. 3, 396–404.

[32] Dror Ben-Zeev, Stephen M Schueller, Mark Begale, Jennifer Duffecy, John M Kane, and David C Mohr, *Strategies for mhealth research: Lessons from 3 mobile intervention studies*, Administration and Policy in Mental Health and Mental Health Services Research (2014), 1–11.

[33] 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, *Mobile behavioral sensing for outpatients and inpatients with schizophrenia*, Psychiatric services **67** (2015), no. 5, 558–561.

[34] Dror Ben-Zeev, Michael A Young, and Patrick W Corrigan, *DSM-V and the stigma of mental illness*, Journal of Mental Health **19** (2010), no. 4, 318–327.

[35] Yoav Benjamini and Yosef Hochberg, *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.

[36] Yoav Benjamini and Daniel Yekutieli, *The control of the false discovery rate in multiple testing under dependency*, Annals of statistics (2001), 1165–1188.

[37] Randall J Bergman, David R Bassett Jr, and Diane A Klein, *Validity of 2 devices for measuring steps taken by older adults in assisted-living facilities.*, Journal of physical activity & health **5** (2008).

[38] Christoph Bergmeir and José M. Benítez, *On the use of cross-validation for time series predictor evaluation*, Information Sciences **191** (2012), 192–213.

[39] Vinay Bettadapura, Edison Thomaz, Aman Parnami, Gregory D Abowd, and Irfan Essa, *Leveraging context to support automated food recognition in restaurants*, Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on, IEEE, 2015, pp. 580–587.

[40] Max Birchwood, Jo Smith, Fiona Macmillan, Bridget Hogg, Rekha Prasad, Cathy Harvey, and Sandy Bering, *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** (1989), no. 03, 649–656.

[41] Max Birchwood, Elizabeth Spencer, and Dermot McGovern, *Schizophrenia: early warning signs*, Advances in Psychiatric Treatment **6** (2000), no. 2, 93–101.

[42] *Brief psychiatric rating scale (bprs) expanded version (4.0).*, 2017.

[43] Dena M Bravata, Crystal Smith-Spangler, Vandana Sundaram, Allison L Gienger, Nancy Lin, Robyn Lewis, Christopher D Stave, Ingram Olkin, and John R Sirard, *Using pedometers to increase physical activity and improve health: a systematic review*, Jama **298** (2007), no. 19, 2296–2304.

[44] Leo Breiman, *Random forests*, Machine learning **45** (2001), no. 1, 5–32.

[45] PRABIR BURMAN, EDMOND CHOW, and DEBORAH NOLAN, *A cross-validatory method for dependent data*, Biometrika **81** (1994), no. 2, 351–358.

[46] P Burton, L Gurrin, and P Sly, *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** (1998), no. 11, 1261–91.

[47] Alison L Calear and Helen Christensen, *Systematic review of school-based prevention and early intervention programs for depression*, Journal of adolescence **33** (2010), no. 3, 429–438.

[48] A Colin Cameron and Frank AG Windmeijer, *R-squared measures for count data regression models with applications to health-care utilization*, Journal of Business & Economic Statistics **14** (1996), no. 2, 209–220.

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

[50] Thomas C Chalmers, Harry Smith, Bradley Blackburn, Bernard Silverman, Biruta Schroeder, Dinah Reitman, and Alexander Ambroz, *A method for assessing the quality of a randomized control trial*, Controlled clinical trials **2** (1981), no. 1, 31–49.

[51] Yin-Wen Chang, Cho-Jui Hsieh, Kai-Wei Chang, Michael Ringgaard, and Chih-Jen Lin, *Training and testing low-degree polynomial data mappings via linear svm*, Journal of Machine Learning Research **11** (2010), no. Apr, 1471–1490.

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

[53] Zhenyu Chen, Mu Lin, Fanglin Chen, Nicholas D Lane, Giuseppe Cardone, Rui Wang, Tianxing Li, Yiqiang Chen, Tanzeem Choudhury, and Andrew T Campbell, *Unobtrusive sleep monitoring using smartphones*, Proceedings of the 7th International Conference on Pervasive Computing Technologies for Healthcare, 2013.

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

[55] Tanzeem Choudhury, Sunny Consolvo, Beverly Harrison, Jeffrey Hightower, Anthony LaMarca, Louis LeGrand, Ali Rahimi, Adam Rea, G Bordello, Bruce

Hemingway, et al., *The mobile sensing platform: An embedded activity recognition system*, Pervasive Computing, IEEE **7** (2008), no. 2, 32–41.

[56] Philip I Chow, Karl Fua, Yu Huang, Wesley Bonelli, Haoyi Xiong, Laura E Barnes, and Bethany A Teachman, *Using mobile sensing to test clinical models of depression, social anxiety, state affect, and social isolation among college students*, Journal of medical Internet research **19** (2017), no. 3.

[57] Jack Cohen, *Statistical power analysis for the behavioral sciencies*, Routledge, 1988.

[58] Sheldon Cohen, Tom Kamarck, and Robin Mermelstein, *A global measure of perceived stress*, Journal of health and social behavior (1983), 385–396.

[59] Sunny Consolvo, David W McDonald, Tammy Toscos, Mike Y Chen, Jon Froehlich, Beverly Harrison, Predrag Klasnja, Anthony LaMarca, Louis LeGrand, Ryan Libby, et al., *Activity sensing in the wild: a field trial of ubifit garden*, Proceedings of the SIGCHI conference on human factors in computing systems, ACM, 2008, pp. 1797–1806.

[60] Patrick Corrigan and Alicia Matthews, *Stigma and disclosure: Implications for coming out of the closet*, Journal of mental health **12** (2003), no. 3, 235–248.

[61] Corinna Cortes and Vladimir Vapnik, *Support-vector networks*, Machine learning **20** (1995), no. 3, 273–297.

[62] Roddy Cowie and Ellen Douglas-Cowie, *Automatic statistical analysis of the signal and prosodic signs of emotion in speech*, Spoken Language, 1996. IC-

SLP 96. Proceedings., Fourth International Conference on, vol. 3, IEEE, 1996, pp. 1989–1992.

[63] John G Csernansky, Ramy Mahmoud, and Ronald Brenner, *A comparison of risperidone and haloperidol for the prevention of relapse in patients with schizophrenia*, New England Journal of Medicine **346** (2002), no. 1, 16–22.

[64] Andrea Dal Pozzolo, Olivier Caelen, Reid A Johnson, and Gianluca Bontempi, *Calibrating probability with undersampling for unbalanced classification*, Computational Intelligence, 2015 IEEE Symposium Series on, IEEE, 2015, pp. 159–166.

[65] Dartmouth College Office of Institutional Research, *Dartmouth student health survey*, 2016, http://www.dartmouth.edu/ oir/2016-dartmouth-health-survey-final-web-version.pdf.

[66] Kadir Demirci, Mehmet Akgönül, and Abdullah Akpinar, *Relationship of smartphone use severity with sleep quality, depression, and anxiety in university students*, Journal of behavioral addictions **4** (2015), no. 2, 85–92.

[67] Ed Diener, Derrick Wirtz, William Tov, Chu Kim-Prieto, Dong-won Choi, Shigehiro Oishi, and Robert Biswas-Diener, *New well-being measures: Short scales to assess flourishing and positive and negative feelings*, Social Indicators Research **97** (2010), no. 2, 143–156.

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

[69] Lisa Dixon, Gretchen Haas, Peter J Weiden, John Sweeney, and Allen J Frances, *Drug abuse in schizophrenic patients: clinical correlates and reasons for use*, Am J Psychiatry **148** (1991), no. 2, 224–230.

[70] Olive Jean Dunn, *Multiple comparisons among means*, Journal of the American Statistical Association **56** (1961), no. 293, 52–64.

[71] Nathan Eagle and Alex Pentland, *Reality mining: sensing complex social systems*, Personal and ubiquitous computing **10** (2006), no. 4, 255–268.

[72] Enrique Echeburúa, Montserrat Gómez, and Montserrat Freixa, *Prediction of relapse after cognitive-behavioral treatment of gambling disorder in individuals with chronic schizophrenia: A survival analysis*, Behavior Therapy **48** (2017), no. 1, 69–75.

[73] Daniel Eisenberg, Ezra Golberstein, and Sarah E Gollust, *Help-seeking and access to mental health care in a university student population*, Medical care **45** (2007), no. 7, 594–601.

[74] Ahmed El-Rabbany, *Introduction to gps: the global positioning system*, Artech house, 2002.

[75] Jon D Elhai, Robert D Dvorak, Jason C Levine, and Brian J Hall, *Problematic smartphone use: A conceptual overview and systematic review of relations with anxiety and depression psychopathology*, Journal of affective disorders **207** (2017), 251–259.

[76] Jon D Elhai, Jason C Levine, Robert D Dvorak, and Brian J Hall, *Non-social features of smartphone use are most related to depression, anxiety and problematic smartphone use*, Computers in Human Behavior **69** (2017), 75–82.

[77] Jane Elith, John R Leathwick, and Trevor Hastie, *A working guide to boosted regression trees*, Journal of Animal Ecology **77** (2008), no. 4, 802–813.

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

[79] L.V. Fausett and W. Elwasif, *Predicting performance from test scores using backpropagation and counterpropagation*, Neural Networks, 1994. IEEE World Congress on Computational Intelligence., 1994 IEEE International Conference on, vol. 5, Jun 1994, pp. 3398–3402 vol.5.

[80] *Foursquare place api*, `https://developer.foursquare.com/places-api`, 2018.

[81] Kenneth R Fox, *The influence of physical activity on mental well-being*, Public health nutrition **2** (1999), no. 3a, 411–418.

[82] Daniel Joseph France, Richard G Shiavi, Stephen Silverman, Marilyn Silverman, and D Mitchell Wilkes, *Acoustical properties of speech as indicators of depression and suicidal risk*, Biomedical Engineering, IEEE Transactions on **47** (2000), no. 7, 829–837.

[83] David A Freedman, *Statistical models: theory and practice*, cambridge university press, 2009.

[84] Jerome H. Friedman, *Greedy function approximation: A gradient boosting machine*, The Annals of Statistics **29** (2001), no. 5, 1189–1232.

[85] Susan R Furr, John S Westefeld, Gaye N McConnell, and J Marshall Jenkins, *Suicide and depression among college students: A decade later.*, Professional Psychology: Research and Practice **32** (2001), no. 1, 97.

[86] Steven J Garlow, Jill Rosenberg, J David Moore, Ann P Haas, Bethany Koestner, Herbert Hendin, and Charles B Nemeroff, *Depression, desperation, and suicidal ideation in college students: results from the american foundation for suicide prevention college screening project at emory university*, Depression and anxiety **25** (2008), no. 6, 482–488.

[87] Ginger.io, *Ginger.io*, 2017, https://ginger.io/.

[88] John F Gleeson, David Rawlings, Henry J Jackson, and Patrick D McGorry, *Early warning signs of relapse following a first episode of psychosis*, Schizophrenia research **80** (2005), no. 1, 107–111.

[89] *Google activity recognition api.*, `https://developers.google.com/android/reference/com/google/android/gms/location/ActivityRecognitionApi`, 2017.

[90] Eric Granholm, Catherine Loh, and Joel Swendsen, *Feasibility and validity of computerized ecological momentary assessment in schizophrenia*, Schizophrenia bulletin **34** (2008), no. 3, 507–514.

[91] Isabelle Guyon and André Elisseeff, *An introduction to variable and feature selection*, Journal of machine learning research **3** (2003), no. Mar, 1157–1182.

[92] Max Hamilton, *A rating scale for depression*, Journal of neurology, neurosurgery, and psychiatry **23** (1960), no. 1, 56.

[93] Nils Y. Hammerla and Thomas Plötz, *Let's (not) stick together: Pairwise similarity biases cross-validation in activity recognition*, Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (New York, NY, USA), UbiComp '15, ACM, 2015, pp. 1041–1051.

[94] James A Hanley and Barbara J McNeil, *The meaning and use of the area under a receiver operating characteristic (roc) curve.*, Radiology **143** (1982), no. 1, 29–36.

[95] Gabriella M Harari, Nicholas D Lane, Rui Wang, Benjamin S Crosier, Andrew T Campbell, and Samuel D Gosling, *Using smartphones to collect behavioral data in psychological science: opportunities, practical considerations, and challenges*, vol. 11, Sage Publications Sage CA: Los Angeles, CA, 2016, pp. 838–854.

[96] C Haring, R Banzer, A Gruenerbl, S Oehler, G Bahle, P Lukowicz, and O Mayora, *Utilizing smartphones as an effective way to support patients with bipolar disorder: Results of the monarca study*, European Psychiatry **30** (2015), 558.

[97] Treniece Lewis Harris and Sherry Davis Molock, *Cultural orientation, family cohesion, and family support in suicide ideation and depression among african american college students*, Suicide and Life-Threatening Behavior **30** (2000), no. 4, 341–353.

[98] Steven A Harvey, Elliot Nelson, John W Haller, and Terrence S Early, *Lateralized attentional abnormality in schizophrenia is correlated with severity of symptoms*, Biological Psychiatry **33** (1993), no. 2, 93–99.

[99] Graeme Hawthorne, *Measuring social isolation in older adults: development and initial validation of the friendship scale*, Social Indicators Research **77** (2006), no. 3, 521–548.

[100] HealthRhythms, *Healthrhythms*, 2018, https://www.healthrhythms.com/.

[101] James Hedlund, *The brief psychiatric rating scale (bprs): A comprehensive review*, 1980.

[102] Yoshinosuke Henmi, *Prodromal symptoms of relapse in schizophrenic outpatients: retrospective and prospective study*, Psychiatry and clinical Neurosciences **47** (1993), no. 4, 753–775.

[103] Marvin I Herz and Charles Melville, *Relapse in schizophrenia.*, The American Journal of Psychiatry (1980).

[104] Tin Kam Ho, *Random decision forests*, Document analysis and recognition, 1995., proceedings of the third international conference on, vol. 1, IEEE, 1995, pp. 278–282.

[105] Sepp Hochreiter and Jürgen Schmidhuber, *Long short-term memory*, Neural computation **9** (1997), no. 8, 1735–1780.

[106] Jin-Hyuk Hong, Julian Ramos, and Anind K Dey, *Toward personalized activity recognition systems with a semipopulation approach*, IEEE Transactions on Human-Machine Systems **46** (2016), no. 1, 101–112.

[107] Jin-Hyuk Hong, Julian Ramos, Choonsung Shin, and Anind K Dey, *An activity recognition system for ambient assisted living environments*, International Competition on Evaluating AAL Systems through Competitive Benchmarking, Springer, 2012, pp. 148–158.

[108] Peter J Huber et al., *Robust estimation of a location parameter*, The Annals of Mathematical Statistics **35** (1964), no. 1, 73–101.

[109] Joel W Hughes and Catherine M Stoney, *Depressed mood is related to high-frequency heart rate variability during stressors*, Psychosomatic medicine **62** (2000), no. 6, 796–803.

[110] Thomas Insel, Bruce Cuthbert, Marjorie Garvey, Robert Heinssen, Daniel S Pine, Kevin Quinn, Charles Sanislow, and Philip Wang, *Research domain criteria (rdoc): toward a new classification framework for research on mental disorders*, 2010.

[111] Oliver P John and Sanjay Srivastava, *The big five trait taxonomy: History, measurement, and theoretical perspectives*, Handbook of personality: Theory and research **2** (1999), 102–138.

[112] Richard Kadison and Theresa Foy DiGeronimo, *College of the overwhelmed: The campus mental health crisis and what to do about it.*, Jossey-Bass, 2004.

[113] Andrew H Kemp and Daniel S Quintana, *The relationship between mental and physical health: insights from the study of heart rate variability*, International Journal of Psychophysiology **89** (2013), no. 3, 288–296.

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

[115] David Kimhy, Inez Myin-Germeys, Jasper Palmier-Claus, and Joel Swendsen, *Mobile assessment guide for research in schizophrenia and severe mental disorders*, Schizophrenia bulletin (2012), sbr186.

[116] Laurence J Kirmayer, James M Robbins, Michael Dworkind, and Mark J Yaffe, *Somatization and the recognition of depression and anxiety in primary care.*, The American journal of psychiatry (1993).

[117] Ron Kohavi et al., *A study of cross-validation and bootstrap for accuracy estimation and model selection*, Ijcai, vol. 14, Stanford, CA, 1995, pp. 1137–1145.

[118] Alex Kopelowicz, Joseph Ventura, Robert Paul Liberman, and Jim Mintz, *Consistency of brief psychiatric rating scale factor structure across a broad spectrum of schizophrenia patients*, Psychopathology **41** (2007), no. 2, 77–84.

[119] S.B. Kotsiantis and P.E. Pintelas, *Predicting students marks in hellenic open university*, Advanced Learning Technologies, 2005. ICALT 2005. Fifth IEEE International Conference on, July 2005, pp. 664–668.

[120] Michael J Kozak and Bruce N Cuthbert, *The nimh research domain criteria initiative: background, issues, and pragmatics*, Psychophysiology **53** (2016), no. 3, 286–297.

[121] Kurt Kroenke and Robert L Spitzer, *The phq-9: a new depression diagnostic and severity measure*, Psychiatric Annals **32** (2002), no. 9, 509–515.

[122] Kurt Kroenke, Robert L Spitzer, and Janet BW Williams, *The phq-9*, Journal of general internal medicine **16** (2001), no. 9, 606–613.

[123] Kurt Kroenke, Robert L Spitzer, Janet BW Williams, and Bernd Löwe, *An ultra-brief screening scale for anxiety and depression: the phq–4*, Psychosomatics **50** (2009), no. 6, 613–621.

[124] Kurt Kroenke, Tara W Strine, Robert L Spitzer, Janet BW Williams, Joyce T Berry, and Ali H Mokdad, *The phq-8 as a measure of current depression in the general population*, Journal of affective disorders **114** (2009), no. 1, 163–173.

[125] Min Kwon, Dai-Jin Kim, Hyun Cho, and Soo Yang, *The smartphone addiction scale: development and validation of a short version for adolescents*, PloS one **8** (2013), no. 12, e83558.

[126] Nicholas D Lane, Emiliano Miluzzo, Hong Lu, Daniel Peebles, Tanzeem Choudhury, and Andrew T Campbell, *A survey of mobile phone sensing*, Communications Magazine, IEEE **48** (2010), no. 9, 140–150.

[127] Nicholas D Lane, Mashfiqui Mohammod, Mu Lin, Xiaochao Yang, Hong Lu, Shahid Ali, Afsaneh Doryab, Ethan Berke, Tanzeem Choudhury, and Andrew Campbell, *Bewell: A smartphone application to monitor, model and promote wellbeing*, 5th International ICST Conference on Pervasive Computing Technologies for Healthcare, 2011.

[128] Kung-Yee Liang and Scott L Zeger, *Longitudinal data analysis using generalized linear models*, Biometrika **73** (1986), no. 1, 13–22.

[129] Georgia Tech Campus Life, *Campus life — optimizing the student environment*, 2017, http://www.quantifiedcampus.gatech.edu/.

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

[131] Hong Lu, Jun Yang, Zhigang Liu, Nicholas D Lane, Tanzeem Choudhury, and Andrew T Campbell, *The jigsaw continuous sensing engine for mobile phone applications*, Proceedings of the 8th ACM conference on embedded networked sensor systems, 2010.

[132] Laurens Van Der Maaten and Geoffrey Hinton, *Visualizing Data using t-SNE*, Journal of Machine Learning Research **9** (2008), 2579–2605.

[133] Anmol Madan, Manuel Cebrian, David Lazer, and Alex Pentland, *Social sensing for epidemiological behavior change*, Proceedings of the 12th ACM international conference on Ubiquitous computing, ACM, 2010, pp. 291–300.

[134] Marek Malik, *Heart rate variability*, Annals of Noninvasive Electrocardiology **1** (1996), no. 2, 151–181.

[135] Christopher D Manning, Christopher D Manning, and Hinrich Schütze, *Foundations of statistical natural language processing*, MIT press, 1999.

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

[137] Daniel Martinez, *Predicting student outcomes using discriminant function analysis.*, (2001).

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

[139] Charles E McCulloch and John M Neuhaus, *Generalized linear mixed models*, Wiley Online Library, 2001.

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

[141] Microsoft, *Microsoft band*, 2016, https://www.microsoft.com/microsoft-band/en-us.

[142] Emiliano Miluzzo, Nicholas D Lane, Kristóf Fodor, Ronald Peterson, Hong Lu, Mirco Musolesi, Shane B Eisenman, Xiao Zheng, and Andrew T Campbell, *Sensing meets mobile social networks: the design, implementation and evalu-*

ation of the CenceMe application, Proceedings of the 6th ACM conference on Embedded network sensor systems, 2008.

[143] Mindstrong, *Mindstrong health*, 2018, https://mindstronghealth.com/.

[144] Megan A Moreno, Lauren A Jelenchick, Katie G Egan, Elizabeth Cox, Henry Young, Kerry E Gannon, and Tara Becker, *Feeling bad on facebook: Depression disclosures by college students on a social networking site*, Depression and anxiety **28** (2011), no. 6, 447–455.

[145] Richard Morriss, Indira Vinjamuri, Mohammad Amir Faizal, Catherine A Bolton, and James P McCarthy, *Training to recognize the early signs of recurrence in schizophrenia*, Schizophrenia bulletin **39** (2013), no. 2, 255–256.

[146] Abderrahamane Mouhab, *Predictive inference: An introduction*, Journal of the American Statistical Association **90** (1995), no. 432, 1489–1491.

[147] Christopher JL Murray, Jerry Abraham, Mohammed K Ali, Miriam Alvarado, Charles Atkinson, Larry M Baddour, David H Bartels, Emelia J Benjamin, Kavi Bhalla, Gretchen Birbeck, et al., *The state of us health, 1990-2010: burden of diseases, injuries, and risk factors*, JAMA **310** (2013), no. 6, 591–606.

[148] Christopher JL Murray, Theo Vos, Rafael Lozano, Mohsen Naghavi, Abraham D Flaxman, Catherine Michaud, Majid Ezzati, Kenji Shibuya, Joshua A Salomon, Safa Abdalla, et al., *Disability-adjusted life years (dalys) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the global burden of disease study 2010*, The Lancet **380** (2013), no. 9859, 2197–2223.

[149] Venet Osmani, *Smartphones in mental health: detecting depressive and manic episodes*, IEEE Pervasive Computing **14** (2015), no. 3, 10–13.

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

[151] John E Overall and Donald R Gorham, *The brief psychiatric rating scale*, Psychological reports **10** (1962), no. 3, 799–812.

[152] 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, *Scikit-learn: Machine learning in Python*, Journal of Machine Learning Research **12** (2011), 2825–2830.

[153] Skyler Place, Danielle Blanch-Hartigan, Channah Rubin, Cristina Gorrostieta, Caroline Mead, John Kane, Brian P Marx, Joshua Feast, Thilo Deckersbach, et al., *Behavioral indicators on a mobile sensing platform predict clinically validated psychiatric symptoms of mood and anxiety disorders*, Journal of Medical Internet Research **19** (2017), no. 3.

[154] John P Pollak, Phil Adams, and Geri Gay, *PAM: a photographic affect meter for frequent, in situ measurement of affect*, Proceedings of the SIGCHI conference on Human factors in computing systems, 2011.

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

[156] Kiran K Rachuri, Mirco Musolesi, Cecilia Mascolo, Peter J Rentfrow, Chris Longworth, and Andrius Aucinas, *Emotionsense: a mobile phones based adaptive platform for experimental social psychology research*, Proceedings of the 12th ACM international conference on Ubiquitous computing, 2010, pp. 281–290.

[157] Anshul Rai, Krishna Kant Chintalapudi, Venkata N Padmanabhan, and Rijurekha Sen, *Zee: Zero-effort crowdsourcing for indoor localization*, Proceedings of the 18th annual international conference on Mobile computing and networking, ACM, 2012, pp. 293–304.

[158] Alvin C Rencher, *Methods of multivariate analysis*, vol. 492, John Wiley & Sons, 2003.

[159] Till Roenneberg, *Chronobiology: the human sleep project*, Nature **498** (2013), no. 7455, 427–428.

[160] Cristobal Romero, Pedro G Espejo, Amelia Zafra, Jose Raul Romero, and Sebastian Ventura, *Web usage mining for predicting final marks of students that use moodle courses*, Computer Applications in Engineering Education **21** (2013), no. 1, 135–146.

[161] Daniel W Russell, *UCLA loneliness scale (version 3): Reliability, validity, and factor structure*, Journal of personality assessment **66** (1996), no. 1, 20–40.

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

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

[164] Sohrab Saeb, Mi Zhang, Christopher J Karr, Stephen M Schueller, Marya E Corden, Konrad P Kording, and David C Mohr, *Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study*, Journal of medical Internet research **17** (2015), no. 7.

[165] Takaya Saito and Marc Rehmsmeier, *The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets*, PloS one **10** (2015), no. 3, e0118432.

[166] SAMHSA, *Key substance use and mental health indicators in the united states: Results from the 2015 national survey on drug use and health*, 2015, https://www.samhsa.gov/data/sites/default/files/NSDUH-FFR1-2015/NSDUH-FFR1-2015/NSDUH-FFR1-2015.htm.

[167] Henry Scheffe, *The analysis of variance*, vol. 72, John Wiley & Sons, 1999.

[168] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller, *Kernel principal component analysis*, International Conference on Artificial Neural Networks, Springer, 1997, pp. 583–588.

[169] Suzanne C Segerstrom and Lise Solberg Nes, *Heart rate variability reflects self-regulatory strength, effort, and fatigue*, Psychological science **18** (2007), no. 3, 275–281.

[170] John Shawe-Taylor and Nello Cristianini, *Kernel methods for pattern analysis*, Cambridge university press, 2004.

[171] Saul Shiffman, Arthur A Stone, and Michael R Hufford, *Ecological momentary assessment*, Annu. Rev. Clin. Psychol. **4** (2008), 1–32.

[172] Choonsung Shin and Anind K Dey, *Automatically detecting problematic use of smartphones*, Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing, ACM, 2013, pp. 335–344.

[173] Tom AB Snijders, *On cross-validation for predictor evaluation in time series*, pp. 56–69, Springer, 1988.

[174] Robert L Spitzer, Kurt Kroenke, Janet BW Williams, Patient Health Questionnaire Primary Care Study Group, et al., *Validation and utility of a self-report version of prime-md: the phq primary care study*, Jama **282** (1999), no. 18, 1737–1744.

[175] Robert L Spitzer, Kurt Kroenke, Janet BW Williams, and Bernd Löwe, *A brief measure for assessing generalized anxiety disorder: the gad-7*, Archives of internal medicine **166** (2006), no. 10, 1092–1097.

[176] 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, *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** (2015), no. 12, 1485–1496.

[177] Allan Stisen, Henrik Blunck, Sourav Bhattacharya, Thor Siiger Prentow, Mikkel Baun Kjærgaard, Anind Dey, Tobias Sonne, and Mads Møller Jensen, *Smart devices are different: Assessing and mitigating mobile sensing heterogeneities for activity recognition*, Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems, ACM, 2015, pp. 127–140.

[178] Yoshihiko Suhara, Yinzhan Xu, and Alex'Sandy' Pentland, *Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks*, Proceedings of the 26th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, 2017, pp. 715–724.

[179] Cheryl D Swofford, John W Kasckow, Geri Scheller-Gilkey, and Lawrence B Inderbitzin, *Substance use: a powerful predictor of relapse in schizophrenia*, Schizophrenia research **20** (1996), no. 1, 145–151.

[180] Ashay Tamhane, Shajith Ikbal, Bikram Sengupta, Mayuri Duggirala, and James Appleton, *Predicting student risks through longitudinal analysis*, Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (New York, NY, USA), KDD '14, ACM, 2014, pp. 1544–1552.

[181] N Tarrier, Christine Barrowclough, and JS Bamrah, *Prodromal signs of relapse in schizophrenia*, Social Psychiatry and Psychiatric Epidemiology **26** (1991), no. 4, 157–161.

[182] Shelley E Taylor, William T Welch, Heejung S Kim, and David K Sherman, *Cultural differences in the impact of social support on psychological and biological stress responses*, Psychological Science **18** (2007), no. 9, 831–837.

[183] Edison Thomaz, Irfan Essa, and Gregory D Abowd, *A practical approach for recognizing eating moments with wrist-mounted inertial sensing*, Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ACM, 2015, pp. 1029–1040.

[184] Robert Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society. Series B (Methodological) (1996), 267–288.

[185] Mickey T Trockel, Michael D Barnes, and Dennis L Egget, *Health-related variables and academic performance among first-year college students: Implications for sleep and other behaviors*, Journal of American college health **49** (2000), no. 3, 125–131.

[186] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman, *Missing value estimation methods for dna microarrays*, Bioinformatics **17** (2001), no. 6, 520–525.

[187] Fani Tsapeli and Mirco Musolesi, *Investigating causality in human behavior from smartphone sensor data: a quasi-experimental approach*, EPJ Data Science **4** (2015), no. 1, 24.

[188] Catrine Tudor-Locke, Susan B Sisson, Tracy Collova, Sarah M Lee, and Pamela D Swan, *Pedometer-determined step count guidelines for classifying walking intensity in a young ostensibly healthy population*, Canadian Journal of Applied Physiology **30** (2005), no. 6, 666–676.

[189] Verily, *Tackling mental health at verily*, 2017, https://blog.verily.com/2017/05/tackling-mental-health-at-verily.html.

[190] Theo Vos, Ryan M Barber, Brad Bell, Amelia Bertozzi-Villa, Stan Biryukov, Ian Bolliger, Fiona Charlson, Adrian Davis, Louisa Degenhardt, Daniel Dicker, 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** (2015), no. 9995, 743–800.

[191] Theo Vos, Abraham D Flaxman, Mohsen Naghavi, Rafael Lozano, Catherine Michaud, Majid Ezzati, Kenji Shibuya, Joshua A Salomon, Safa Abdalla, Victor Aboyans, et al., *Years lived with disability (ylds) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the global burden of disease study 2010*, The Lancet **380** (2013), no. 9859, 2163–2196.

[192] Fabian Wahle, Tobias Kowatsch, Elgar Fleisch, Michael Rufer, and Steffi Weidt, *Mobile sensing and support for people with depression: a pilot trial in the wild*, JMIR mHealth and uHealth **4** (2016), no. 3.

[193] Strother H Walker and David B Duncan, *Estimation of the probability of an event as a function of several independent variables*, Biometrika **54** (1967), no. 1-2, 167–179.

[194] He Wang, Souvik Sen, Ahmed Elgohary, Moustafa Farid, Moustafa Youssef, and Romit Roy Choudhury, *No need to war-drive: Unsupervised indoor localization*, Proceedings of the 10th international conference on Mobile systems, applications, and services, ACM, 2012, pp. 197–210.

[195] 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, *Crosscheck: Toward passive sensing and detection of mental health changes in people with schizophrenia*, Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (New York, NY, USA), UbiComp '16, ACM, 2016, pp. 886–897.

[196] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell, *Studentlife: Assessing mental health, academic performance and behavioral trends of college students using smartphones*, Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (New York, NY, USA), UbiComp '14, ACM, 2014, pp. 3–14.

[197] Rui Wang, Gabriella Harari, Peilin Hao, Xia Zhou, and Andrew T Campbell, *Smartgpa: how smartphones can assess and predict academic performance of college students*, Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing, ACM, 2015, pp. 295–306.

[198] Rui Wang, Gabriella Harari, Peilin Hao, Xia Zhou, and Andrew T. Campbell, *Smartgpa: How smartphones can assess and predict academic performance of*

college students, Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (New York, NY, USA), UbiComp '15, ACM, 2015, pp. 295–306.

[199] Rui Wang, Weichen Wang, Min SH Aung, Dror Ben-Zeev, Rachel Brian, Andrew T Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, Emily A Scherer, et al., *Predicting symptom trajectories of schizophrenia using mobile sensing*, Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **1** (2017), no. 3, 110.

[200] Rui Wang, Weichen Wang, Alex daSilva, Jeremy F Huckins, William M Kelley, Todd F Heatherton, and Andrew T Campbell, *Tracking depression dynamics in college students using mobile phone and wearable sensing*, Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **2** (2018), no. 1, 43.

[201] Jun-ichiro Watanabe, Saki Matsuda, and Kazuo Yano, *Using wearable sensor badges to improve scholastic performance*, Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication, ACM, 2013, pp. 139–142.

[202] Jun-Ichiro Watanabe, Kazuo Yano, and Saki Matsuda, *Relationship between physical behaviors of students and their scholastic performance*, Ubiquitous Intelligence and Computing, 2013 IEEE 10th International Conference on and 10th International Conference on Autonomic and Trusted Computing (UIC/ATC), IEEE, 2013, pp. 170–177.

[203] David Watson, Lee A Clark, and Auke Tellegen, *Development and validation of brief measures of positive and negative affect: the panas scales.*, Journal of personality and social psychology **54** (1988), no. 6, 1063.

[204] Gary S Wilkinson and GJ Robertson, *Wide range achievement test (wrat4)*, Psychological Assessment Resources, Lutz (2006).

[205] Rui Xu, Donald Wunsch, et al., *Survey of clustering algorithms*, Neural Networks, IEEE Transactions on **16** (2005), no. 3, 645–678.

[206] Rachel Yehuda, *Post-traumatic stress disorder*, New England journal of medicine **346** (2002), no. 2, 108–114.

[207] Yosef Hochberg Yoav Benjamini, *Controlling the false discovery rate: A practical and powerful approach to multiple testing*, Journal of the Royal Statistical Society. Series B (Methodological) **57** (1995), no. 1, 289–300.

[208] Amelia Zafra, Cristóbal Romero, and Sebastián Ventura, *Multiple instance learning for classifying students in learning management systems*, Expert Systems with Applications **38** (2011), no. 12, 15020 – 15031.

[209] S L Zeger and K Y Liang, *An overview of methods for the analysis of longitudinal data.*, Statistics in medicine **11** (1992), no. 14-15, 1825–39.

[210] Hui Zou and Trevor Hastie, *Regularization and variable selection via the elastic net*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **67** (2005), no. 2, 301–320.