Protocol

Passive Mobile Self-Tracking of Mental Health by Veterans with Serious Mental Illness

NCT 05023252
April 29, 2020

This project studies passive mobile sensing with Veterans in treatment for serious mental illness (SMI). Data are used for self-tracking behaviors and symptoms. While passive mobile sensing has been feasible, acceptable and safe in patients with serious mental illness, these are studied for the first time in VA. Analytics are developed that use passive data to predict behaviors and symptoms. The project has these objectives:

1. Conduct user-centered design of passive mobile self-tracking to support Veterans' management of their mental health.
2. Study the feasibility, acceptability and safety of passive self-tracking of mental health that includes feedback of mental health status to the Veteran.
3. Use mobile sensor and phone utilization data to develop individualized estimates of sociability, activities, and sleep as measured by weekly interviews.
4. Study the predictive value of using data on sociability, activities, and sleep to identify exacerbations of psychiatric symptoms.

The informatics platform will capture an array of passive sensing data off participant's smartphone device to track metrics indicative of behavioral patterns and physiological well-being. These metrics are all passively collected, requiring no active user input, to minimize user burden and ensure maximal user participation in the study. Comparison data will also be collected in the form of weekly assessments conducted by research staff via telephone.

The project will perform statistical analysis on the data to find which short- and long-term features could be correlated with clinical status. These features include both short- and long-term behavioral data, i.e. what an individual has done recently and how their behavior has changed over time. To generate these long-term features, our team will perform time-series analysis on the data and find which parts of participants' daily lives have changed over the course of the study. We will compare these changes to surveys collected at different points throughout the duration of the research study. After generating the relevant features, we use machine learning techniques to study a classifier which predicts whether an individual is in a risk group or not.

<u>Methods</u>

This study that will enroll patients with mental health diagnoses at risk of symptoms who own a smartphone. Participants will download a smartphone application to their phones which will capture an array of data for a period of nine months. Initially, we conduct focus groups and in-lab usability testing with patients and interviews with clinicians to ensure that development of this functionality is consistent with this VA population. We will conduct 3 focus groups with patients (N=18) and interviews with clinicians (N=10) and make resulting modifications to the app. After these, we conduct in-lab usability testing with 8 patients and make resulting modifications to the app.

Then participants will be recruited and enrolled into the mobile sensing trial where they will download a smartphone application to their phones which will collect an array of data passively for a period of nine months. Participants will receive research assessments from a research assistant at baseline and final follow-up. They will also be asked to complete brief surveys with a research assistant

by phone once a week. A subsample will also complete a semi-structured qualitative interview by phone at 3- months and in person at 9- months. For the duration of each participant's involvement, mobile sensor data are collected. Data are transmitted daily to the server.

For all phases of the study, patients are recruited based on being in treatment for a serious mental illness, being at risk of symptoms, and using a mobile phone. Participants will receive assessments from a research assistant at baseline and follow-up. This baseline and follow-up assessments will be conducted by a research assistant. Patient interviews assess recent medication adherence, housing status, substance abuse, psychiatric symptoms, cognitive deficits associated with SMI, and functioning. In addition, each week, research staff will contact participants by phone, and administer brief assessments of sociability, activity, sleep, and mood and psychotic symptoms. At the conclusion of the study period, the participant will come to the VA to meet with research staff for a final quantitative follow-up interview which includes the same measures from the baseline assessment. At 3 and 9 months, a semi-structured interview is conducted with a subsample by phone to elucidate the feasibility, acceptability and safety of mobile sensing. At the end of the study period, the app will be deleted from the phone, and the participant will be notified that it is no longer functional.

Recruitment and screening

A list of potentially eligible participants at the GLA VA will be identified from VA data. We will obtain demographic information, mental health diagnoses, variables determining high risk status (per inclusion criteria), name, last 4 of social security number, address, and telephone number. These data are available through VA Informatics and Computing Infrastructure (VINCI) accessing the Corporate Data Warehouse (CDW), or medical center data systems. We will use an opt-out letter to invite individuals to find out more about the study. If eligible participants do not opt out, they will be contacted by the research team to inquire about study interest. The research team will field calls in response to the letter (both those who want to opt out and those who self-initiate a call for more study information). We will also post flyers at clinics of the Greater Los Angeles VA inviting participation and field calls in response to these.

Patient Participant Eligibility

The study involves distinct patient samples: participants in 3 focus groups (n=6 each), participants in an individual in-lab usability trial (n=8), and users of the app for 9 months (n=100). The inclusion and exclusion criteria are the same for all samples, as are the recruitment methods. An individual can participate in one focus group, an in-lab usability trial, and the mobile sensing trial. Furthermore, individuals can participate in the in-lab usability trials or mobile sensing trial and not have participated in the focus groups.

Inclusion criteria: 1) chart diagnosis of SMI, defined as a diagnosis of schizophrenia, schizoaffective disorder, or bipolar disorder, confirmed by the patient's treating psychiatrist; 2) risk for symptoms based on having had, during the past year, psychiatric hospitalization, psychiatric emergency care, lived at a crisis treatment program, or more than 6 outpatient treatment visits; and, 3) ownership of a phone with a data plan.

Exclusion criteria: 1) under age 18. 2) Patient has a conservator/legally authorized representative who makes their medical decisions.

Staff participant eligibility

Inclusion criteria: clinicians who are involved in treating individuals with SMI at GLA VA.

Exclusion criteria: none

Staff Stakeholder Interviews

A purposive sample of 10 clinicians who are involved in treating individuals with SMI at the GLA VA will be interviewed in person before the intervention is implemented and after. These interviews will be conducted after interviewees complete a verbal informed consent process. The baseline interviews will assess acceptability of mobile sensing from a clinician perspective, including usefulness as a tool to improve clinical assessment and care and their recommendations for program refinements. The follow-up interviews will ask about how to reach the appropriate population, reflections on our findings, intention for referrals, implementation issues, and resources needed for sustainability and incorporation into clinical practice. These 30-minute interviews will be digitally recorded and transcribed.

Patient Participant Eligibility Check

For those interested in participating, additional eligibility criteria will be confirmed by self-report by phone. We will ask the prospective participant if they own a smartphone and help them to determine if it has a data plan. We will confirm their primary psychiatric diagnosis. After confirming eligibility criteria, potential participants will be scheduled to receive further information, confirm the eligibility assessment, and conduct full informed consent. We will tell potential participants to bring their cell phone and a recent cell phone bill to that meeting to confirm phone platform and data plan.

Participant Consent: After explaining the study, if the Veteran is interested, research staff will schedule an appointment to review and sign the informed consent form in-person. Research staff will use informed consent procedures validated in individuals with serious mental illness and limited English language proficiency. These involve an iterative process of querying the participant's understanding with a 10-item true/false test on key consent information and providing feedback until an acceptable level of understanding is achieved (no less than 100% correct after no more than 3 tries). It is our experience that almost all (more than 95%) outpatients with serious mental illness can successfully complete this informed consent process. Patients unable to successfully answer all questions on the consent quiz after 3 attempts will not be eligible to enroll.

Patient Focus Groups

We will conduct 3 focus groups, approximately 2 weeks apart, to elucidate acceptance and usability regarding the app and mobile monitoring. Each focus group will entail participants discussing perceived usefulness, ease of use, attitudes, self-efficacy, norms, facilitating conditions and behavioral intent. The focus groups will be co-led by the qualitative methods expert and research staff. Each focus group will include 6 Veterans who meet inclusion criteria. Focus groups will be digitally audio recorded and professionally transcribed.

Patient In-Lab Usability Trials

After app modifications have been made in response to the focus groups and clinician interviews, in-lab usability trials will be conducted. The trials will be conducted in an office at the VA. They will be co-led by the qualitative methods expert and research staff. Users will participate individually and will be asked to complete specific tasks (e.g., find the dashboard of your sleep). Topics include downloading and opening the app, reviewing dashboards, closing the app, and managing the app on the user's phone. The user will be asked to verbalize their thoughts as they go through the experience of navigating the app. Research staff logs what the participant did and said in as much detail as possible. Errors are also counted (e.g., wrong pathways, failure to find content).

Patient Participant Assessments and App Download

For participants in the mobile sensing trial, after the informed consent process, research staff administer the baseline research assessment. Then the project app is downloaded to the participant's

phone. It is activated by staff using a code. Each time the app is activated on a device, a unique ID is created by the server. This ID connects data from the app with patient identifiers kept on the secure server. After the app is activated, it runs in the background and collects data passively with no further action from the user.

Participants will receive phone calls from study staff on a weekly basis throughout a 9-month study period to complete assessments by phone. Based on our experience with these assessment instruments, as staff get to know individual patients the time required for assessments becomes relatively brief. A semi-structured qualitative interview will also be conducted at 3 months by phone with a subsample of participants until data saturation is achieved. At the conclusion of the 9 month study period, the participants meet with research staff for a final follow-up interview. At this point, the project app will be deleted from the phone, and the participant will be notified that it is no longer functional. The final assessment will also include a qualitative component for a subsample of participants until data saturation is achieved.

Study Measures

Screening

Demographic and Psychiatric: VistA data, extracted by a study analyst, will be used to obtain psychiatric diagnosis and age. VistA data will be used to identify whether the individual has received treatment from GLA VA during the past year and is at high risk for symptoms as defined in the inclusion criteria.

Phone Ownership and Data Plan: We will ask to see the potential participant's cell phone to confirm cell phone platform type, and see a cell phone bill to confirm that they have a data plan.

Baseline and Final Follow-up

Demographics: Demographic data including age, race, ethnicity, marital status, educational attainment, annual family income, and household roster; and psychiatric illness history will be collected from the patient at baseline only.

Phone data plan: Phone data plan information will be collected from each patient's recent phone bill brought to the baseline assessment. We will also collect their phone number.

Psychopathology, Substance Abuse, Cognitive Deficits, Functioning: Psychopathology will be measured using the Brief Psychiatric Rating Scale (BPRS) at both baseline and 9 month follow-up. Psychopathology will be measured using the Brief Psychiatric Rating Scale (BPRS). We will use the BPRS Positive Symptoms Factor (psychosis), Activation Factor (mania), and Affect Factor (depression) (9 items total).

Cognitive deficits of SMI will be measured by brief assessment. We will assess verbal learning using the Hopkins Verbal Learning Test-Revised (HVLT-R), and processing speed using the Digit Symbol Coding Test at baseline only. Occupational, social, symptom-related and overall functioning will be assessed using the MIRECC GAF at baseline only, which has been validated in Veterans with SMI.

Housing Stability: Housing stability for the past 30 days will be assessed using the Residential Time-Line Follow-Back Inventory (RTLFBI). The RTLFBI has been used in prior research to collect retrospective housing history over a specified period of time among people who have experienced homelessness. All residences over 30 days will be recorded, and coded into one of 34 previously defined locations, which will be further categorized as "stable" (e.g., own apartment, room or studio apartment in a supportive housing program, transitional housing, group home, boarding home) versus "unstable" (e.g., living on the streets, in public places, in emergency shelter-type accommodations) settings, using established methodology. The number of days spent in any of the locations categorized as "stable" or "unstable" will be summed and divided by the total number of days of residency reported at the interview,

and reported as a percentage. RTLFBI has been demonstrated sensitive to change. This will be collected at baseline only.

Medication Possession Ratio (MPR): MPR has been shown to be a valid measure of medication adherence in Veterans with SMI. MPR assesses the extent to which dispensed medications provide coverage for the given interval. We will calculate a MPR for the 12 months pre-baseline using pharmacy data to assess medication adherence.

Healthcare Utilization: The Service Use and Resources Form (SURF) will be completed by patients at baseline only. SURF has been successfully used to assess costs and service utilization in numerous multi-site studies with patients with SMI, including CATIE. For the present study, we will use the SURF questionnaire to collect information on inpatient and outpatient services utilization, including type of facility and type of clinician contact, for visits both inside and outside VA settings.

Sociability, Activity, Sleep and Symptoms: We will administer the measures for these 4 domains described below at both baseline and the follow-up.

Weekly research assessments

Sociability: Social withdrawal and functioning will be assessed using the Abbreviated Lubben Social Network Scale (6 items) and the Objective Frequency of Social Contact scale (6 items).

Activity: Physical activity, routines and habits will be assessed using the Short International Physical Activity Questionnaire (4 items) and the Independence Performance and Prosocial domains of the Social Functioning Scale (15 items).

Sleep: Sleep duration and quality will be assessed using components 1, 2, 3 and 6 from the Pittsburgh Sleep Quality Index (5 items) and the Insomnia Problem items from the Insomnia Severity Index (3 items).

Symptoms: Symptom severity will be assessed using the Brief Psychiatric Rating Scale. We will use the BPRS Positive Symptoms Factor (psychosis), Activation Factor (mania), and Affect Factor (depression) (9 items total).

Safety: We will monitor adverse events on a weekly basis using data from weekly assessment phone calls and the medical record. Based on prior research studies of mobile sensing in people with SMI, and our prior mobile informatics research in patients with SMI at VA, we do not expect any safety problems.

Formative evaluation assessments

Acceptability: At 3 and 9 months, a semi-structured interview will be conducted with a subsample of participants until data saturation is achieved. The interview will be audio recorded and professionally transcribed. The interview will assess acceptability of mobile sensing, including usefulness, ease of use, attitudes and behavioral intention towards use. We study what participants like and do not like and why, their perception of the usefulness of mobile sensing as a tool to improving their assessment and care, and their recommendations for program refinements. These interviews will be digitally recorded. Monthly, we will calculate the number of days that the app was recording and transmitting data for each participant.

Safety: We will monitor the rate of serious adverse events throughout participation.

Data Collected from Mobile Application

The app collects passive behavioral data for assessment in three domains that have previously been proven to be measurable using these data: sociability, activities, and sleep. Sensor data collected from the app is encrypted and cached locally on the device, securely transmitted to the platform's remote server when a data connection is available. Analyses will seek to correlate behavioral patterns with emotional well-being. Analyses will also conduct time-series analysis to monitor shifts over time. We

anticipate detecting acute health stress through shifts in behavior patterns. The app will passively collect behavior data in three domains: sociability, activity, and sleep.

Participants will be able to view the tracking of their activities, sociability and sleep over time, similar to how people currently view their step counts in mobile health app dashboards. Each of the behavioral domains in this project will be presented on a 1-100 scale, as numbers graphed over time.

Sociability explores social characteristics of the individual. Dynamic and active social lives are correlates of mental health. Concurrently, sudden changes in sociability, particularly a notable increase in introversion and isolation could be indicators of diminishing mental health. Sociability assesses social characteristics by collecting information on calls and texts made with the phone (the content of these communications is not assessed or recorder) and public interaction by measuring ambient sound in social environments. Specific measures of Sociability we will attempt to gather:

- Communication: regular communication with a diverse set of individuals by phone call, text messaging, or use of messaging apps
- Public Interaction: regular communication with others in person or spending time in a social environment

Activity measures an individual's day-to-day routine by monitoring activity, movement, location, and habits. Adhering to regular routines, engaging in exercise, and productive activity outside the home are indicators of mental health. Changes in routines, lack of exercise, and limited intentional movement outside the home can be signs of mental health issues. Specific measures of Activity we will gather:

- Routine: a regular and structured daily and weekly routine
- Exercise: regular exercise
- Activities: Engagement in activities, particularly outside of the individual's residence

Sleep measures will focus on regular, uninterrupted sleep, in sufficient quantities represents one of key measures of mental health. Sleep measures collect data on ambient light, ambient sound, phone movement, lack of communication, and lack of use of the phone. Specific measures of Sleep we will attempt to gather:

- Quality: Is the individual having uninterrupted sleep
- Duration: what is the length of sleep the individual is having

Table 1. Data for Modeling of Behavioral Domains

Behavioral Domain	Input Data from Phone	Model Output Classifiers
Activities physical activity intensity and duration (walking, running, climbing stairs, riding bicycle) vs. sedentary organized activities: number and duration outside the individual's residence regular structured activities throughout each day (location and duration)	• accelerometer sensor: acceleration force data for three coordinate axes • linear accelerometer Android software: acceleration force data for three coordinate axes (excludes gravity) • gyroscope sensor: rotation data for three coordinate axes • rotational vector Android software: rotation vectors for three coordinate axes and scalar • step counter Android software • significant motion Android software: motion leading to change in location • Activity Recognition Google API: still, on foot, on a bicycle, in a vehicle • Fused Location Google API (uses GPS & wifi): location after 5 meters of movement	• Short International Physical Activity Questionnaire (4 items) • Independence, Performance and Prosocial domains of the Social Functioning Scale (15 items)

Sleep total sleep duration uninterrupted sleep regular daily sleep and wake times	Sleep: • ambient light sensor: indoors, outdoors or dark location • ambient sound sensor: volume • significant motion Android software: motion leading to change in location • Fused Location Google API (uses GPS & wifi): location after 5 meters of movement • phone unlock, screen interactions and on-time duration • log of phone calls placed • log of messages sent • apps opened • app duration of use	• Pittsburgh Sleep Quality Index Components 1, 2, 3, 6 (5 items) • Insomnia Problem items from the Insomnia Severity Index (3 items)
Sociability communication in person or in a public social environment communication with a diverse set of individuals communication with repeated partners	• log of phone calls placed with phone numbers • log of phone calls received with phone numbers • log of messages sent with phone numbers • log of messages received with phone numbers • social media apps opened • social media apps: keystrokes and duration used • messaging or email apps opened • messaging or email apps: keystrokes and duration used • ambient sound sensor: volume • Activity Recognition Google API: still, on foot • Fused Location Google API (by GPS & wifi): location after 5 meters of movement	• Objective Frequency of Social Contact scale (6 items) • Abbreviated Lubben Social Network Scale (6 items)

Data Analyses

Aims 1 & 2. Conduct user-centered design of passive mobile self-tracking to support Veterans' management of their mental health. Study the feasibility, acceptability and safety of passive self-tracking of mental health that includes feedback of mental health status to the Veteran.

Before deployment of the app, user-centered design is conducted to inform further development of the app and intervention. Using ATLAS.ti, transcripts of clinician interviews, patient focus groups, and notes from in-lab patient usability testing will be coded deductively for major subthemes regarding usefulness, ease of use, attitudes and behavioral intention towards use, acceptability, strengths and weaknesses, barriers and facilitators of use, safety, and recommendations about revisions and implementation efforts. Results inform modifications of the app, and refinement of methods for enrollment and maintenance of participation.

During deployment of the app, formative evaluation is conducted to understand and strengthen development, deployment and future implementation. Feasibility is characterized by studying the extent to which potential subjects enroll, maintain involvement and complete study assessments. Acceptability is characterized by counting days during which patients use mobile sensing and dashboards, and by analyzing data from semi-structured interviews. Safety is measured using the rate of serious adverse events. The following may be considered serious adverse events: death, life-threatening experience, inpatient hospitalization, persistent or significant disability, or incapacity. Important medical events that do not result in death or require hospitalization may be considered serious adverse events if they jeopardize the participant or require medical or surgical intervention to prevent one of the outcomes above.

Using multivariate analyses, we study whether quantitative baseline patient characteristics described above are associated with feasibility, acceptability, and safety. Qualitative data from patients 3-month and 9-month interviews are used to evaluate themes related to the feasibility, acceptability and safety of sensing. Qualitative data from post-intervention clinician interviews are used to evaluate themes related to patient engagement, findings, implementation, and resources for sustainability and incorporation into routine practice. Using ATLAS.ti, transcripts of interviews will be coded deductively for major subthemes regarding acceptability, strengths and weaknesses, barriers and facilitators of use, safety, and recommendations about future revisions and implementation efforts. The qualitative findings will be triangulated with findings from the other measure of acceptability and safety to examine whether concordance exists among these measures of the intervention. Should analytic anomalies in the quantitative findings arise, the qualitative data will be explored to illuminate possible explanations for unexpected findings. At 3 and 9 months, 26 individuals should be sufficient to provide data saturation regarding acceptability of the intervention using qualitative data.

Aims 3 & 4. Use mobile sensor and phone utilization data to develop individualized estimates of sociability, activities, and sleep as measured by weekly interview. Study the predictive value of using data on sociability, activities, and sleep to identify exacerbations of psychiatric symptoms.

We study digital phenotypes that are informative as predictors of behaviors and symptoms. These include both short- and long-term behavioral data, i.e. what an individual has done recently and how their behavior and symptoms have changed over time. There will be multiple real time data streams collected that will be continuous and intermittent sequence activities to assess behaviors (Table 1). Even though the data can be represented as a function of time t, they are being collected at vastly different frequencies and durations, and with different noise and error rates. Thus, we use a two-phase analytics pipeline of modeling. In phase 1, we conduct data preprocessing that derives features capturing events from the data collected. In phase 2, we conduct predictive analytics using machine learning that models correlations between sensor and phone utilization data, and behavioral assessments and symptoms.

From the project's mobile app, we obtain highly detailed patient level information over a 9 month follow up period and transmit it to the server database. In the data preprocessing phase, we reduce these mobile data to continuous summaries that retain more information than categorical ones. Sensor and utilization data are captured at varying frequencies for the different domains. For measures that are updated in seconds and minutes, we summarize data using activity profiles (over a day or a week), defined as the amount of time the person spends at an activity level (e.g., above a certain threshold) or engaged in an activity (e.g., walking, exercising, riding in a motor vehicle). Movement measured through the accelerometer will be transformed to an activity profile by defining activities from no movement to faster movement (using thresholds). Light and sound levels will be thresholded from low to high and the amount of time spent at each level captured. In this way, problems with temporal alignment of data are mitigated, transforming data into densely measured functional data, conceptualized as longitudinal functional data. Here the profile constitutes the functional observation varying over levels (summarizing over a day or a week) and days/weeks constitute the longitudinal repetitions. For measurements summarized over days (e.g., total phone calls, total hours of sleep), data are conceptualized as longitudinal data. We also include static patient measures from baseline that could improve the accuracy of models, such as demographics, prior medication adherence, housing status, psychiatric symptoms, cognitive deficits associated with SMI, and social and occupational functioning.

For data obtained from smartphones, an issue that will arise frequently is missing data. When adjusting for missing data in creating activity profiles, we will consider various approaches. One approach is a similarity assumption which states that the proportion of time that a participant spends at a given activity level during the observed times is similar to the proportion in the unobserved times. This is similar to a missing at random assumption in longitudinal data analysis. For example, time spent at different

activity levels (activity profile) could be rescaled to include time over unobserved periods as well. Another approach is to use the planned machine learning approaches to also evaluate missing data patterns.

In Aim 3 we develop individualized estimates of sociability, activities and sleep. We will aggregate daily mobile data on a weekly basis to predict behavioral measures that are also obtained weekly. Summary statistics such as mean, variance, min, or max will be used for aggregation as deemed appropriate. We will then employ various machine learning algorithms to build, train, and select prediction models for each of the patient behavioral assessment domains using features of sensor and phone utilization data (Table 1). Model development will proceed in three stages. First, we will utilize methods of unsupervised learning to cluster sensor and phone utilization data and extract features to reduce the dimensionality. Second, we will explore a number of supervised learning procedures to build candidate prediction models for patient behavioral assessments. Finally, we will use cross-validation to estimate the out-of-sample performance of the prediction models, and to select best model(s) from among the candidates.

To reduce the number of features upon which to build the prediction models, while preserving the maximum amount of information, the first strategy is to use principal component analysis (PCA). PCA reduces the dimensionality of a data set by identifying linear combinations of the features, the principal components, which maximize variability while remaining mutually uncorrelated. Building the models using fewer candidate features reduces the risk of overfitting, and using the principal components in place of the original features prevents problems of multicollinearity. The second dimension reduction strategy is to use FDR (false discovery rate) threshold. FDR-based thresholding gives an optimal way of adapting to unknown sparsity (only a small number of effects are significantly large) and may be viewed as an estimation procedure with an appropriate penalization scheme accounting for the number of selected variables.

Given the reduced number of features, various supervised machine learning algorithms will be used to train prediction models for patient behaviors and symptoms. At the preliminary analysis stage, a set of supervised learning algorithms will be explored to evaluate cross-sectional associations between mobile data features and patient behavioral assessments, with the goal of narrowing down the pool of candidate predictors. First, we will consider random forests. A random forest is a collection of decision trees, each built using a random sample of features, and trained on a bootstrap sample from the original data. Random forests are useful for identifying complex interactions among features due to the use of decision tree learning, while protecting against overfitting by means of bootstrap aggregation. Second, we will consider support vector machines (SVM). SVM is a method of supervised learning which classifies instances into categories by estimating a hyperplane that separates instances belonging to different classes. SVM may be used if the outcomes of interest, behaviors and symptoms, can be meaningfully categorized (e.g., high vs. low). Selected predictors from these approaches will be used in the next step of the model building process that takes into account the temporal correlation and time series nature of mobile data and patient assessments. We will use hidden Markov models to predict patients' current behaviors and symptoms based on current features and past patients' behaviors and symptoms. We view patients' features as observed states, and patients' behaviors and symptoms are hidden states that enjoy the Markov property. We will also use linear or generalized linear mixed effects models to predict patient behavioral assessments using selected features from the preliminary stage, while taking into account within-patient correlation in repeated measurements over time.

Models developed using supervised machine learning methods will be evaluated in terms of out-of-sample predictive performance using leave-one-out cross-validation (LOOCV). For time series modeling, models will be trained on times 1, ..., t-1, and evaluated on time t, for each time t. For cross-sectional modeling, models will be trained on all but patient i, and evaluated on patient i, for each patient i. Performance will be averaged across hold-out times/patients to obtain a LOOCV estimate of

generalization error. Quadratic loss will be used to evaluate prediction error. The selected model will be the one that minimizes LOOCV mean squared error.

In Aim 4, outcome data are collected weekly for each patient on three scales (psychosis, mania, depression). These scales each have a range of 1 to 7. The primary outcome will be the scale indicating the most severe symptoms in a 2-month window. We will evaluate the association between mobile data features identified in Aim 3 with this primary outcome using a mixed effects model to accommodate within patient data correlation due to repeated measures. Specifically, we will consider feature data from fixed time windows preceding the most severe symptoms. The outcome data distribution will be examined prior to model fitting, and an appropriate link function will be used in the mixed effects model.

Sample size and power considerations: In Aim 3, we extract features of mobile sensor and phone utilization data to estimate behaviors in the domains of sociability, activity, and sleep. The selected features should have a high correlation with behaviors to be a good estimate. Of 100 participants enrolled, assuming 30% attrition, we expect to have complete data on 70 participants (and substantial longitudinal data for a larger group). This sample size will provide over 95% power to detect a correlation of at least 0.50 at the significance level 0.05 using a two-sided test. In Aim 4, our inclusion criteria select for greater than average symptom risk, so we anticipate substantial symptoms. Previous research on 21 patients with SMI for a median of 134 days found that sensing data predicted symptom indicator scores with correlations in the range of 0.5 to 0.8. Given possible confounders, and since our sample may be more heterogeneous than prior research, the correlation in our study is anticipated to be lower than what was reported previously. With 100 enrolled, and assuming 30% attrition, we will have over 85% power to detect correlations of at least 0.35 between predictors of interest and symptoms assessments at the significance level of 0.05 using a two-sided test.