A review of physiological and behavioral monitoring with digital sensors for neuropsychiatric illnesses

Erik Reinertsen, *Georgia Institute of Technology*
Gari D. Clifford, *Emory University*

Final published version: http://dx.doi.org/10.1088/1361-6579/aabf64
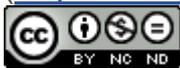
Copyright information:

A review of physiological and behavioral monitoring with digital sensors for neuropsychiatric illnesses

Erik Reinertsen[1] and **Gari D. Clifford**[1,2]

[1]Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA, USA

[2]Department of Biomedical Informatics, Emory University, Atlanta, GA, USA

Abstract

Physiological, behavioral, and psychological changes associated with neuropsychiatric illness are reflected in several related signals, including actigraphy, location, word sentiment, voice tone, social activity, heart rate, and responses to standardized questionnaires. These signals can be passively monitored using sensors in smartphones, wearable accelerometers, Holter monitors, and multimodal sensing approaches that fuse multiple data types. Connection of these devices to the internet has made large scale studies feasible and is enabling a revolution in neuropsychiatric monitoring. Currently, evaluation and diagnosis of neuropsychiatric disorders relies on clinical visits, which are infrequent and out of the context of a patient's home environment. Moreover, the demand for clinical care far exceeds the supply of providers. The growing prevalence of context-aware and physiologically relevant digital sensors in consumer technology could help address these challenges, enable objective indexing of patient severity, and inform rapid adjustment of treatment in real-time. Here we review recent studies utilizing such sensors in the context of neuropsychiatric illnesses including stress and depression, bipolar disorder, schizophrenia, post traumatic stress disorder, Alzheimer's disease, and Parkinson's disease.

1. Introduction

Neuropsychiatric illness comprises 13-16% of the total global burden of disease measured in disability life-adjusted years (DALYs) for all ages, which exceeds the burden of cardiovascular disease or cancer (Vigo et al. 2016). One in four people in the world will be affected by mental or neurological disorders at some point in their lives, yet only a small fraction of the 450 million people affected will receive treatment due to pervasive underdiagnosis, a lack of trained healthcare professionals, stigma, and other reasons (Sayers 2001). These illnesses are more prevalent among older people and will contribute even more to overall global disease as life expectancy improves. The burden of mental and substance use disorders increased by 37% between 1990 and 2010, which for most disorders was driven by population growth and aging (Whiteford et al. 2013). The prevalence of dementia continues to rise, and by 2050 an estimated 13.8 million Americans will have Alzheimer's disease (AD; see Table A1 for definitions of abbreviations and acronyms used in this review) or another dementia. In 2016 in the United States, total payments for healthcare, long-term care, and hospice services for people 65 years or older with dementia were estimated to be $230.1 billion, and caregivers provided 18.2 billion hours of unpaid assistance (Alzheimer's Association 2016). The lack of effective interventions for neuropsychiatric illness is partially

due to limited understanding of underlying mechanisms, but also due to under-distribution of medications and human resources in low- and middle-income countries, in which disease burden measured in DALYs is disproportionately high (Collins et al. 2011).

Autonomic nervous system (ANS) dysfunction occurs in neuropsychiatric illness, resulting in altered heart rate (HR), heart rate variability (HRV), galvanic skin response, skin conductance and temperature, and respiratory rate (Draghici et al. 2016; Karemaker 2017). Due to the prevalence of HR sensors in wearable devices, and a substantial amount of literature exploring HRV measurements as markers of ANS modulation, we review studies that utilize HR and HRV measurements. Note HRV is not one metric; rather, it encompasses several types of metrics such as time domain (Stein et al. 1994; Kleiger et al. 2005; Bauer et al. 2017), frequency domain (Akselrod et al. 1981; Montano et al. 2009), and complexity measures such as entropy (Costa et al. 2002). Changes in these metrics have been reported in patients with stress (Thayer et al. 2012), major depressive disorder (MDD; Kemp et al. 2010), bipolar disorder (BD; Henry et al. 2010), schizophrenia (Chang et al. 2009), post traumatic stress disorder (PTSD; Liddell et al. 2016), Alzheimer's disease (Femminella et al. 2014) and Parkinson's disease (PD; Maetzler et al. 2013).

Neuropsychiatric illness is also associated with alterations in behavior, especially physical movements and social routine. Patients with MDD, BD, or schizophrenia can be significantly more sedentary than age- and gender-matched healthy controls (Vancampfort et al. 2017). Diminished motor function, the presence of tremor, and coordination issues also occur in movement disorders such as Parkinson's disease. On the other hand, locomotor agitation can be a sign of mania or psychosis which may be part of the presentation of schizophrenia or BD. These abnormalities are detectable by smartphones and wearable devices with accelerometers or global positioning system (GPS) sensors. Because modern smartphones and most wearables marketed to consumers for fitness purposes (which are often used in academic studies for healthcare applications, including many referenced here) have accelerometers, we also review studies that analyze locomotor activity. Behavior can also be inferred from social activity data, such as phone calls, text messages, social media use, and web browser history. Importantly, passive monitoring via digital sensors can yield information about a patient's physiology and behavior in the 99% of the time they are not seeing a clinician, during which they take actions and are influenced by their environment in ways that profoundly impact their health (Asch et al. 2012). Together these data could give us a richer understanding of the day-to-day variability of neuropsychiatric illness, enable the monitoring of patient status before (rather than after) symptoms reach a level warranting intervention, and reduce biases and inaccuracy intrinsic in subjective questionnaires (Karow et al. 2008; Copeland et al. 2017).

Monitoring is distinct from diagnosis. The latter is performed by clinicians who take a comprehensive history, perform a physical exam, utilize questionnaires and surveys, order and interpret laboratory tests and imaging, and exclude alternate diagnoses. Clinical diagnoses often form the ground truth for subsequent monitoring efforts. For example, a machine learning algorithm can associate patterns in digital sensor data – such as alterations in heart rate variability or locomotor activity – with questionnaire results indicating severity, or a clinical diagnosis. Passive monitoring could augment diagnosis by providing clinicians

with additional information, capture behavioral and biological variation not accounted for by current diagnostic categories, and enable the discovery of novel illness phenotypes.

Digital sensors in smartphones and wearables generate a vast amount of high-frequency high-dimensional time series data that require new methods of analysis. In contrast, data used by clinicians – self-reported symptoms, lab tests, and vital signs – are subjective, infrequently sampled, and small-scale. Univariate significance testing and regression models are commonly used to perform hypothesis testing on these traditional data, but such methods are poorly suited for the analysis of data from digital sensors. Rather, approaches from signal processing, information theory, and complexity science are needed. Features of interest in digital sensor data include statistical moments, e.g. the mean or the variance of a signal, time-domain characteristics, frequency-domain characteristics such as power spectral density attributes or wavelet coefficients, and complexity measures such as entropy (Johnson et al. 2016). These features are used to train machine learning algorithms that perform regression, continuous parameter prediction, and classification of outputs such as disease phenotype or questionnaire score (Obermeyer et al. 2016). Excellent machine learning algorithms generalize in the sense that they accurately classify inputs from data used to train the algorithm as well as novel input from an external set of data not used for training. Of note, univariate statistical significance does not guarantee predictivity or clinical utility of a biomarker (Lo et al. 2015). Methods focusing on p-values can miss useful "weak features" – those that do not significantly differ by output class when assessed via univariate statistical tests, yet can be used as input to train a multivariate machine learning algorithm that achieves high accuracy.

In this review we summarize recent studies utilizing smartphones, wrist-worn wearables, and physiological patches for passive monitoring of some prevalent and debilitating neuropsychiatric illness: stress, MDD, BD, schizophrenia, PTSD, AD, and PD (Table A3). Sensors used in these studies measure accelerometry, HR, GPS, phone calls, SMS, and more. Examples of aberrations in physiology and behavior detectable by these sensors and associated with the above illnesses are provided in (Table 1). Particular emphasis is placed on studies that classify illness status, or estimate scores from neurological and psychiatric surveys, scales, and questionnaires (summarized in Table A2). We discuss the challenges, limitations, and potential of using these technologies for neuropsychiatric care. Related works and ongoing studies that have yet to yield results but are promising in terms of scope and scale are referenced in these latter sections (Table A4). We do not review smartphone applications ("apps") designed to deliver interventions such as cognitive behavioral therapy, provide general information to patients about their illness, or accompany existing care delivery paradigms (i.e. a mobile version of a patient portal). Furthermore, a thorough technical review of signal processing, information theory, and machine learning used in these studies is beyond the scope of this review. We also do not cover the topic of sleep, a key aspect of neuropsychiatric conditions that has been extensively reviewed elsewhere (Krystal 2012; Behar et al. 2013; Roebuck et al. 2014; Zinkhan et al. 2016).

2. Smartphones

Smartphones are globally ubiquitous, owned by 72% of Americans and 3 billion people worldwide, and are projected to reach a global total of over 5 billion people by 2030 (Poushter 2016). Importantly, studies in the USA, United Kingdom, Canada, and India have found smartphone ownership to not be significantly lower among people with serious mental health conditions compared to the average owner, and ownership by these individuals is projected to increase, mirroring the trend seen in the general population (Torous et al. 2014; Firth et al. 2016). Additionally, people tend to keep their phones with them and check them between 46 to 85 times per day (Andrews et al. 2015; Eadicicco 2016). These data thus reflect social and behavioral manifestations of neuropsychiatric illnesses in the context of daily life rather than in an artificial clinical setting (Insel 2017). For example, GPS location data measured on smartphones can be used to estimate behavioral attributes such as percentage of time a subject spends in certain locations (fig. 1). By evaluating the time of day, day of week, and amount of time spent in each location, the purpose of each location datum can be inferred, e.g. work versus home. Additionally, social interactions in the form of calls and text messages can be monitored and quantified (fig. 2). Geolocation, social network activity, and other attributes reflect behavior and may differ in subjects with neuropsychiatric illness compared to healthy controls. Several investigators have built smartphone apps for collecting sensor and usage data, including Automated Monitoring of Symptom Severity (AMoSS; Palmius et al. 2014), Purple Robot (Schueller et al. 2014), and Beiwe (Torous et al. 2016).

Smartphones can also be used to administer validated questionnaires for evaluating quality of life and mental well-being (Palmius et al. 2017). Although self-reported questionnaires are prone to recall, social desirability, and confirmation biases, they provide a pragmatic best estimate of an individual's mental status and can achieve results comparable to clinician-administered surveys (Spitzer et al. 2012; Ebner-Priemer et al. 2006; Martel 2008; Solhan et al. 2009). The inference of mental health questionnaire results from digital sensor data is a common approach in the literature and could be useful for monitoring the status of subjects who struggle with adherence or have impaired cognition and executive decision-making capacity (Table A2; Mohr et al. 2016; Tsanas et al. 2016; Barrett et al. 2017; Aung et al. 2017). In this section we review recent studies using smartphones to monitor neuropsychiatric illnesses; work that may be related but also involves analysis of heart rate data is reviewed in later sections.

2.1. Stress and depression

MDD is a debilitating disease that is characterized by depressed mood, diminished interests, impaired cognitive function, vegetative symptoms, disturbed sleep, and altered appetite (Otte et al. 2016). The lifetime incidence of MDD in the United States of 12% in men and 20% in women (Belmaker et al. 2008). Affecting up to 300 million people in the world, MDD is the leading cause of disease burden in middle- and high-income countries worldwide. Individuals with MDD have higher medical costs, exacerbated medical conditions, and significantly increased rates of mortality. Compounding this severity, MDD is a heterogeneous disorder with a highly variable course, an inconsistent response to

pharmacological treatment, and no established mechanism. Passively monitoring movement, location, social activity, and voice of patients with MDD could enable continuous assessment of mental well-being and inform context-appropriate clinical responses.

In a study of depression in 48 college undergraduates, Wang et al. used Android smartphones to monitor accelerometry, audio, ambient light, location, and device use over ten weeks (Wang et al. 2014). Depression was simultaneously measured using the PHQ-9 survey, a standardized nine question survey which has been shown to correlate with depression (Kroenke et al. 2001). In addition to making criteria-based diagnoses of depressive disorders, the PHQ-9 has also been shown to be a reliable and valid measure of depression severity (Martin et al. 2006). A PHQ-9 score 10 resulted in a sensitivity of 0.88 and a specificity of 0.88 for major depression in primary care and obstetrics-gynecology populations, and a sensitivity of 0.77 (0.71-0.84) and a specificity of 0.94 (0.90-0.97) in a meta-analysis, although the positive predictive value in an unselected primary care population was only 0.59 (Wittkampf et al. 2007). Students who slept less, held fewer conversations, self-reported higher stress responses, or interacted less during the day were more likely to be depressed ($p < 0.05$). Students started with high positive affect and conversation levels, low stress, and healthy sleep and daily activity patterns. As the term progressed, self-reported stress significantly rose, while positive affect, sleep, conversation and activity decreased. However, this study did not train a classifier, nor was cross-fold validation used. Study results may thus not generalize to out-of-sample data. Furthermore, p-values do not guarantee diagnostic accuracy or clinical utility (Wasserstein et al. 2016).

Burns et al. developed the "Mobilyze!" app to collect GPS, ambient light, recent calls, and other data (Burns et al. 2011). A companion website included feedback graphs illustrating correlations between patients' self-reported states, as well as didactics and tools teaching patients behavioral activation concepts. Mobilyze! was tested for eight weeks in a cohort of seven adult patients with MDD who completed treatment. The Mini-International Neuropsychiatric Interview was used to characterize co-morbid anxiety disorders at baseline, the PHQ-9 was used to evaluate self-reported MDD symptom severity, and the GAD-7 was used to evaluate general anxiety symptom severity. In the Mobilyze! study, record-wise ten-fold cross validation was performed, although we note record-wise cross-validation overestimates predictive accuracy compared to subject-wise (Saeb et al. 2016b). Decision trees were used to estimate location, activity, social environment, and internal states. Generalized estimating equations logistic regression was used to estimate the binary outcome of either presence or absence of MDD diagnosis in the held-out set of sensor values. Categorical states such as location, isolation, and conversational status were estimated with mean accuracies ranging from 60-90%. However, the decision tree models estimated out-of-sample scale-based states such as mood no better than chance, and the results of the binary classification of MDD status were not reported.

Canzian et al. developed the MoodTraces app for the Android operating system to collect GPS data and answers to eight daily questions from the PHQ-8 depression test (Canzian et al. 2015). This study evaluated 28 subjects from the general population rather than a cohort of people diagnosed with depression. Mobility features were extracted from GPS data recorded over a period of 14 days and used to train a support vector machine (SVM) to

predict PHQ score changes from an individual's own past data, i.e. using individualized models. A positive label was defined as a change in PHQ score greater than one standard deviation of that subject's normal PHQ score, and a negative label was defined as a change in PHQ score less than or equal to one standard deviation of a subject's normal PHQ score. Leave-one-out cross validation was performed. Using a horizon window – the number of days between the last day of data collection and the day of the subsequently predicted PHQ score change – of 0 days resulted in a sensitivity of 0.71 and a specificity of 0.87. Interestingly, increasing the horizon window did not dramatically reduce the sensitivity and specificity. This study suggests that personalized models, instead of general ones, should be used to monitor the depressive state of an individual using his/her mobility traces. However, this study did not survey subjects diagnosed with MDD, nor were other digital sensor data from the smartphone utilized.

Ben-Zeev et al. evaluated behavior and mental health in 47 adolescent subjects using Android smartphones and an app developed in-house. GPS, accelerometry, ambient light and sound, and microphone data were recorded (Ben-Zeev et al. 2015). Geospatial activity, sleep duration, and variability in geospatial activity were associated with daily stress levels assessed via the 10-item Perceived Stress Scale ($p < 0.05$ for all). Sensor-derived speech duration, geospatial activity, and sleep duration were associated with changes in depression assessed via the PHQ-9, and sensor-derived kinesthetic activity was associated with loneliness. However, cross-validation was not performed, and features were assessed on the basis of statistical significance rather than classifier predictivity.

Saeb et al. used Android smartphones and an app developed in-house to evaluate 40 adult subjects for depressive symptoms over two weeks (Saeb et al. 2015). 28 of the subjects had sufficient sensor data to analyze. Several features extracted from GPS and phone usage data were related to depressive symptom severity. The lower the location entropy, e.g. more time spent in fewer locations, or the lower the regularity of circadian rhythm, the more likely a subject was to be depressed. Other predictive features included phone usage duration, and phone usage frequency. Elastic net regularization was performed to reduce overfitting, 1000 bootstrapped sets of features and their corresponding PHQ-9 scores were created, and leave-one-out cross validation was performed. A classifier trained on these data to distinguish subjects with PHQ-9 scores greater than or equal to from those with PHQ-9 scores less than 5 achieved an accuracy of 86.5%. Similar findings were reported in a subsequent study; location features from weekend data better predicted depression compared to location features from weekday data – even weeks in advance (Saeb et al. 2016a). These results suggest the relationship between depression and movement is stronger on non-workdays versus workdays when behavior is driven by social expectations. This finding highlights the importance of social context, time scale, and routine in the study of behavioral manifestations of mental illness.

2.2. Bipolar disorder

BD is a mental illness that can present with mania, hypomania, and major depression; manic episodes are characterized by significant changes in activity, energy, mood, behavior, sleep, and cognition (Belmaker et al. 2008). Patients with BD commonly manifest co-morbid

psychiatric disorders, such as anxiety disorders and substance use disorders. Psychotic features such as delusions, hallucinations, and disorganized thinking and behavior can also occur during manic, major depressive, and mixed episodes.

Effective mood forecasting – the prediction of future mood states using current and past data – could support management of BD and provide early warning signs of relapse. Currently, symptoms are monitored using paper diaries or asking patients to recall mood during a clinical visit. Thus, data potentially usable to model mood in BD is sparse and may suffer from recall bias. Moore et al. used time series regression to forecast the next week's depression ratings using self-rated mood data obtained via SMS (Moore et al. 2012). One method used was Gaussian process regression, a Bayesian nonparametric model in which a Gaussian prior distribution over the regression function is assumed. Forecasting by Gaussian process regression requires centering the time series because the prior process is assumed to have a zero mean. The algorithm then finds the optimal value for the hyperparameters θ and the noise variance σ_n^2 by maximizing the marginal likelihood $p(y/x, \theta)$. The predictive equation is used to find the forecast mean, and the original signal bias is added, thus estimating the mood rating. Other types of simple exponential smoothing forecasting were also used for comparison. 153 patients with BD participated in the study; at least 23 responses were obtained from 100 patients whose data was used in the final analysis. Depression was measured via QIDS-SR$_{16}$ and severity of mania was quantified using the ASRM. Questionnaires were administered every week for up to four years. Mood time series data varied widely in length, response interval, and stationarity. Out-of-sample forecasting was performed to estimate expected prediction error. Gaussian process regression did not outperform simpler exponential smoothing approaches. This is not surprising because noisy or undersampled time series will train a smoothing coefficient of zero, and most of the time series from these patients were noisy or lacked serial correlation. The authors concluded effective depression forecasts using this method cannot be made over the period of a week.

Faurholt-Jepsen et al. conducted the "MONitoring, treAtment and pRediCtion of bipolAr disorder episodes" (MONARCA) study, in which software for Android smartphones was used to monitor subjective and objective manifestations of BD alongside with treatment adherence in a bidirectional feedback loop between patients and providers (Maria et al. 2013). Subjective data included mood/irritability, sleep, and alcohol use. Objective data included speech, social, and physical activity. Data were recorded from 17 patients with BD for 3 consecutive months (Maria et al. 2014). Patients were rated every two weeks using the HDRS-17 and YMRS. Depressive symptoms correlated with less movement and fewer outgoing calls.

The MONARCA study was continued by Faurholt-Jepsen et al. in a larger cohort of 61 patients with BD. A linear mixed-effects regression model was used to estimate relationships between independent and dependent variables while accounting for within individual variation and between individual variations over time (Faurholt-Jepsen et al. 2015). The regression analysis was also adjusted for age and sex as possible confounders. Since the goal of the study was to estimate model coefficients rather than accurately classify subjects and estimate model generalizability, regression rather than a cross-validation or bootstrap

approach was used. The duration of incoming and outgoing calls/day correlated with depressive symptoms. Additionally, the number and duration of incoming calls/day correlated with manic symptoms. Self-reported mood and activity data correlated negatively with HDRS-17 scores and positively with YMRS scores. Self-reported sleep quantity negatively correlated with both HDRS-17 and YMRS scores, whereas self-reported stress positively correlated with both. In other words, less sleep or more stress correlated with more depressive and manic symptomatology.

Recently, MONARCA was updated to collect and extract voice features from phone calls using the open-source "Media Interpretation by Large feature-space Extraction" (openSMILE) toolkit (Eyben et al. 2010; Maria et al. 2016). Class imbalance was addressed via random oversampling of the minority class, and a random forest algorithm trained on features derived from voice, objective, and self-reported data achieved an AUC of 0.78 in classifying a depressive state versus a euthymic state, and an AUC of 0.89 in classifying a manic or mixed state versus a euthymic state – although the number of folds used for k-fold cross-validation were not specified.

The group that created MONARCA recently began enrolling patients in the first-ever randomized controlled trial (RCT) to to investigate whether using a smartphone-based monitoring and treatment system, including an integrated clinical feedback loop, reduces the rate and duration of re-admissions more than standard treatment in unipolar disorder and bipolar disorder (Maria et al. 2017).

Grunerbl et al. passively recorded data from ten patients with BD over ten months using Android phones and a monitoring application developed in-house (Grünerbl et al. 2015). Bipolar symptoms were determined via HAMD or YMRS scale tests conducted every three weeks. Social activity features included number of phone calls, length, and unique numbers contacted. Speech and voice features included average speaking length, turn duration, utterances, and frequency-domain features. Four of the ten patients refused to use the study phone to make phone calls, so speech and voice features were determined for the remaining six patients. A naive Bayes algorithm classified records into one of seven states, including depressive, normal, and manic with different degrees. 66% of the data was randomly allocated to serve as training data, and the remaining 33% was allocated to the test set. This cross-validation was repeated 500 times, and resulted in an average 69% classification accuracy using a fusion of features extracted from accelerometer and GPS data. Next, the classifier was revised to weigh class estimates by data quantity; however, we note a quantity-based weighing introduces its own form of bias and does not guarantee improvement of classifier accuracy. On days with data from multiple modalities (phone, sound, acceleration, and location), class estimates were calculated for each modality and weighed by the amount of data as to favor modalities with more data and penalize modalities with fewer data. The class with the highest estimated probability was selected. Weighing estimates by data quantity dramatically improved classification accuracy to 76%, with both sensitivity and positive predictive value of 97%. However, the authors did not report the timing between feature collection and symptom assessment via questionnaire. Additionally, the variance across different folds was not reported.

Beiwinkel et al. developed "Social Information Monitoring for Patients with Bipolar Affective Disorder" (SIMBA), a smartphone app to track daily mood, physical activity, and social communication (Beiwinkel et al. 2016). SIMBA was tested with a cohort of 13 patients diagnosed with BD. Random-coefficient multilevel models were computed to analyze the relationship between smartphone data and externally rated manic and depressive symptoms. Lower self-reported mood in the monitoring period prior to a clinical visit predicted higher overall levels of clinical depressive symptoms ($p < 0.05$). A decline in social communication and physical activity predicted an increase in clinical depressive symptoms ($p < 0.05$). Lower physical activity but higher social communication predicted higher overall levels of clinical manic symptoms ($p < 0.05$). Lastly, a decrease in physical activity predicted an increase in clinical manic symptoms ($p < 0.05$). This study evaluated prediction rather than classification, as the outcome of interest was temporally later compared to the time at which smartphone data was assessed. However, no cross-validation or external validation cohort data set was utilized to assess generalizability of the model.

BD is characterized by disturbances in rhythmicity of sleep and social routine. Abdullah et al. used passive smartphone sensor data gathered via a customized Android app called "MoodRhythm" to measure rhythmicity in seven BD patients over four weeks (Abdullah et al. 2016). Measured data included accelerometry, ambient light, microphone audio, calls and SMS, and phone usage such as screen unlocks and recharging. Patients were administered the Social Rhythm Metric (SRM) questionnaire, although the frequency was unclear. The most predictive features were derived from these data via recursive feature elimination, and included: number of location clusters, distance traveled, frequency of conversation, and duration of non-sedentary activity. A support vector regression model was trained on data over a rolling window of seven days, and ten-fold cross-validation was performed. The average square root of average of squared errors between predicted and actual SRM scores was 1.40; the SRM ranges from zero to seven. To classify subjects as either "normal social rhythm" or "unstable", a cutoff SRM score of 3.5 was selected due to the population SRM score being ≈3.5. A person with an SRM score < 3.5 was considered unstable, whereas a person with an SRM score 3.5 was considered stable. An SVM using the same features as before achieved a PPV of 0.85 and a sensitivity of 0.86.

Palmius et al. designed the Automated Monitoring of Symptom Severity app, or "AMoSS" ((Palmius et al. 2014)). This app collected location, activity, battery usage, daily self-reported mood (through a six-axis seven-point Likert scale mood survey, 'Mood Zoom') and social networking behavior via de-identified lists of recipients and senders of text messages and phone calls, including length/duration and time of day (fig. 3). Collecting data up to several samples per second, AMoSS is a comprehensive mHealth monitoring system for mental health. Physiological monitoring devices recorded HR and blood pressure from a total of 100 participants, including patients with BD and matched controls. Early results demonstrated correlations between diagnoses of BD and borderline personality disorder with mood reports (Tsanas et al. 2016).

Subsequent work by the AMoSS team on a subset of these data demonstrated that clinically significant depression could be detected using features extracted from GPS location data (Palmius et al. 2017). Anonymized geographic locations were collected from 22 subjects

with BD and 14 controls over three months, using location data from Samsung Galaxy S III or S4 smartphones running the Android 4.1 operating system. This version of Android featured a geospatial resolution of approximately 100 meters. Depressive symptomatology was self-reported by subjects via the QIDS-SR16 survey. Features were extracted to assess the level and regularity of geographic movements of the subjects, including normalized entropy, location variance, and number of distinct location clusters. For subjects with BD, a linear regression model trained on these features estimated questionnaire scores with a mean absolute error rate of 3.73 points. A quadratic discriminant analysis algorithm was trained to classify depression, and achieved an F_1 score of 0.86 ± 0.02, classification accuracy of 0.85 ± 0.02; sensitivity of 0.84 ± 0.01, and specificity of 0.87 ± 0.05 (median $\pm$ IQR). Results were robust to leave-one-out, 10-fold, 5-fold, and 3-fold cross-validation.

2.3. Schizophrenia

Schizophrenia is a complex and heterogeneous illness characterized by several criteria, including at least two of the following symptoms for one month or longer: delusions, hallucinations, disorganized speech, grossly disorganized or catatonic behavior, or negative symptoms such as diminished emotional expression; furthermore, there must be impairment in work, interpersonal relations, or self-care, as well as continuous signs of the disorder for at least 6 months (Kahn et al. 2015). The lifetime global prevalence of schizophrenia is about 1%. Outcomes vary widely, ranging from total recovery to totally debilitating illness requiring chronic care. Life expectancy for people with schizophrenia is reduced by 20 years compared to people without the illness. Pharmacological treatments for schizophrenia can relieve psychotic symptoms but usually fail to meaningfully improve social, cognitive and professional functioning. Psychosocial interventions can be useful but are resource-intensive and inconsistently delivered. Finally, schizophrenia tends to be diagnosed years after symptoms begin. Compared to depression and BD, relatively little work on mobile health has focused on schizophrenia.

Wang et al. collected passive smartphone sensor data from 21 outpatients diagnosed with schizophrenia and recently discharged from hospital over a period ranging from 2 - 8.5 months (Wang et al. 2016). Samsung Galaxy S5 phones running the Android operating system were equipped with the "CrossCheck" app developed in-house that monitors type and duration of physical activity, sleep duration, number and durations of phone conversations, number of SMS, geolocation, phone and app usage, and ambient light and noise. Every three days, Ecological Momentary Assessment (EMA) questions were administered and sensor data were aggregated. Generalized estimating equations were used to map associations between features and EMA responses. Higher scores in attributes related to positive perception of mental well-being – including calm, hopeful, sleeping well, social, and ability to think clearly – were associated with waking up earlier, having fewer conversations, fewer phone calls, and fewer SMS. Higher scores in questions related to negative perception were associated with staying stationary more in the morning but less in the evening, visiting fewer new places, having fewer conversations but making more phone calls and SMS, and using the phone less. Gradient boosted regression trees were used to predict EMA scores from these features. Models trained on an individual's data could estimate EMA scores for that individual with a correlation between prediction and outcome

of $r = 0.77$ and $p < 0.001$. However, outcomes predicted via leave-one-out cross validation did not correlate with actual outcomes, suggesting high variance of feature phenotypes between individuals.

Staples et al. recently reported a three-month observational study of both self-reported and objective measures of sleep in schizophrenia (Staples et al. 2017). Using the Beiwe app (available for both iPhones and Android phones), 13 subjects diagnosed with chizophrenia were given tri-weekly EMAs. Passive data were continuously collected, including accelerometry, GPS, screen use, and anonymized call and SMS activity. Sleep quality was assessed in a clinical setting using the PSQI, which was compared to both EMAs and sleep estimates based on passively collected accelerometer data. A cross-validated linear regression model with mean phone-based EMA scores as the outcome and mean paper-based PSQI scores as the predictor classified 85% (11/13) of subjects as exhibiting high or low sleep quality. Accelerometry moderately correlated with subject self-assessments of sleep duration ($r = 0.69$, 95% CI $0.23 - 0.90$). Active and passive phone data predicted concurrent PSQI scores with a mean average error of 0.75, and future PSQI scores with a mean average error of 1.9, with scores ranging from 0-14.

Among individuals who are diagnosed, hospitalized, and treated for schizophrenia, up to 40% of those who are discharged will relapse within one year. Barnett et al. evaluated a smartphone platform for monitoring seventeen patients with schizophrenia undergoing active treatment in order to identify warning signs of relapse, defined as psychiatric hospitalization or an increase in the level of psychiatric care, such as increase in the frequency of clinic visits or referral to a partial or outpatient hospital program (Barnett et al. 2018). Patients were monitored for three months using the Beiwe app, and mobility patterns and social behavior were gathered and analyzed. Features were extracted from the data, including daily distance traveled, time spent at home, number of significant locations visited, total duration of calls, number of missed calls, and number of text messages sent. The app also administered surveys twice per week to assess anxiety, depression, sleep quality, psychosis, the warning symptoms scale, and medication adherence. The rate of behavioral anomalies detected in the 2 weeks prior to relapse was 71% higher than the rate of anomalies during other time periods. Although anomalies were calculated using each patient's own data to account for differences in baseline features, the number of anomalies greatly varied between subjects. Additionally, many subjects did not relapse, as the cohort enrolled only seventeen patients and for only three months. The features captured in patients that did relapse may not have reflected the "potential trajectories and mechanisms that can lead to relapse". The anomaly detection approach demonstrated in this paper could be useful for measuring other outcomes that were not reported but could be clinically useful, such as changes in positive or negative symptoms of schizophrenia.

2.4. Post traumatic stress disorder

PTSD is a psychopathological response to a traumatic event such as violence, a natural disaster, or combat. Symptoms include nightmares of the trauma, hypervigilance, difficulty sleeping, poor concentration, and avoidance of places, activities, or persons that remind the affected individual of the causal incident (Shalev et al. 2017). The lifetime prevalence of

PTSD ranges between 6-30%, is much higher in veterans exposed to combat, and varies by gender and country of origin.

Smartphone apps for PTSD have focused on education about the disorder, delivery of cognitive behavioral therapy, self-assessment of symptoms via questionnaires, and access to crisis support and other relevant resources (Kuhn et al. 2014). Few papers describe the use of digital sensors to passively monitor clinical symptoms of PTSD. However, many smartphone- and wearables-based sensing approaches have focused on anxiety and depression which are common co-morbidities.

Place et al. conducted a 12-week trial with 73 patients who reported at least one symptom of PTSD or depression (Place et al. 2017). Clinical symptoms were assessed by licensed social workers who administered the depression and PTSD modules of the Structured Clinical Interview for Mental Disorders (SCID). An Android app was developed to gather accelerometry, SMS and call, location, device use, and audio data. Extracted features included sum of outgoing calls, count of unique numbers texted, absolute distance traveled, dynamic variation of the voice, speaking rate, and voice quality. Feature reduction was performed to reduce over-fitting and interfeature correlation, and a logistic regression was trained using 10-fold cross validation. Fatigue, interest in activities, and social connectedness were predicted using data from the prior week with AUCs of 0.56, 0.75, and 0.83 respectively. Depressed mood was predicted from audio data with an AUC of 0.74. Finally, subjects reported comfort with sharing personal data with clinicians and medical researchers. However, it was unclear if the predictive model outperformed sample-and-hold estimations of mood from the previous week. This can be viewed using a Bayesian framework, in which the mood state from the previous week informs the prior, and data from the smartphone app is used to update the model and estimate the posterior. Evaluating subjects at several time points affords an opportunity to quantify the additional contribution of passive sensor data to predictive models that use questionnaires or surveys as ground truth.

University of North Carolina, Harvard University, and Verily Life Sciences LLC (South San Francisco, CA) are leading the AURORA study, a 19-institution five-year effort to perform the most comprehensive observational study of mental disorders that occur in the wake of trauma to date (National Institute of Mental Health 2016). Investigators will screen 5,000 people arriving in emergency rooms after trauma. After an initial evaluation and a baseline collection of biological data from blood samples, subjects will be monitored for the next several months through the use of mobile technology, such as wrist wearables and smart phones, to track factors like activity, sleep, and mood. Other assessments will include additional blood samples, functional brain imaging, and psychological tests. Participant involvement will continue over a year, generating a wide variety of detailed information on, for example, health history (including that of earlier trauma), genetics, stress responses (physical and psychological), behavior, and cognition. This collaboration presents a unique opportunity to learn more about the factors that mediate the development of mental illness after trauma, and potentially contribute to new diagnostic and therapeutic approaches. The Aurora study represents a new trend in public-private partnerships, involving multiple

research institutions and and technology companies such as Verily and Mindstrong Health (Palo Alto, CA).

2.5. Dementia and Alzheimer's disease

AD is a progressive, chronic neurodegenerative disorder that primarily affects older adults and is the most common cause of dementia. Selective memory impairment is the most common initial symptom (Wolk et al. 2017). Executive dysfunction and visuospatial impairment also present early, while deficits in language and behavioral symptoms typically manifest later in the disease course. The average life expectancy after a diagnosis of AD ranges from three to eight years; patients generally succumb to complications related to advanced debilitation such as dehydration, malnutrition, and infection. AD has an estimated prevalence of 10–30% in the population > 65 years of age with an incidence of 1–3% (Masters et al. 2015). According to The Alzheimer's Association over 5 million Americans have AD and it is the sixth leading cause of death (Alzheimer's Association 2016). In 2016 more than 15 million caregivers provided 18 billion hours of unpaid care at a value of approximately $221B. Given the devastating cost and prevalence of AD, relatively low cost smartphone and wearable technologies could potentially aid cognitive assessment and improvement, monitor daily activities, and prolong independence and/or improve lives of caregivers.

Tung et al. attempted to distinguish AD from controls using GPS features from smartphones (Tung et al. 2014). A cohort of 19 older adults with mild-to-moderate AD and 33 controls were monitored via wearing GPS-enabled mobile phones for three days. The make and model of the phones were not described. Measures of geographical territory (area, perimeter, mean distance from home, and time away from home) were calculated from GPS data and group differences were tested using two-sample t-tests. Area, perimeter, and mean distance from home were significantly smaller in the AD group compared to controls. Furthermore, area and perimeter were significantly associated with steps per day, Disability Assessment for Dementia questionnaire scores, gait velocity, symptoms of apathy, and depression.

Aguilera et al. demonstrated self-reported daily mood ratings assessed via SMS are proxies for PHQ-9 scores (Aguilera et al. 2015). A cohort of 33 people with a diagnosis of depression who were undergoing group cognitive behavioral therapy received a daily SMS asking them to report mood on an ordinal scale of 1-9. They were also asked questions about thoughts and activities. The subjects further received a PHQ-9 to complete each week that they attended the therapy group. Daily and one-week average SMS mood scores were significantly associated with PHQ-9 scores. The authors noted "SMS-based mobile mood ratings, when assessed daily, may provide a more accurate indicator of longitudinal symptom levels than the PHQ-9, as the PHQ-9 may be subject to a recency bias". While short SMS-based surveys can provide a more nimble and higher-resolution picture of how a patient feels in the moment, they are not intended to replace more thorough survey instruments such as the PHQ-9, which focuses on the diagnostic criteria of the DSM-IV for MDD (Kroenke et al. 2001).

Batista et al. developed an Android app to monitor people with mild cognitive impairment (MCI) (Batista et al. 2015). Specifically, the app raises alarms under certain conditions, such

as an AD patient leaving a defined geographic zone (e.g. home), not moving after a certain amount of time, moving at too high a speed (suggesting they are utilizing transportation), or the phone battery level reaching too low a level. 16 subjects with varying stages of AD participated in a pilot study in which user perception of the app was assessed. The authors selected the most inexpensive Android smartphones available via Amazon to make the platform affordable for more patients and caregivers.

2.6. Parkinson's disease

PD is a neurodegenerative movement disorder that affects 1-3% of the population   65 years old (Poewe et al. 2017). PD is caused by a marked reduction of dopaminergic neurons in the substantia nigra, and subsequent disruption of dopaminergic neurotransmission in the basal ganglia. The diagnosis of PD is clinical as it classically presents with asymmetric resting tremor, rigidity, and bradykinesia (Nutt et al. 2005). The economic burden of PD is estimated to be $23B in the USA and projected to increase to $50B by the year 2040 (Dodel et al. 1998). Current dopamine replacement therapy with Carbidopa and Levodopa treats symptoms of the disease and is not curative; PD continues to progress resulting in significant disability, worsening quality of life, and eventual need for advanced care and nursing home placement (Huse et al. 2005).

Neurologists can adjust medication selection and dosage to manage symptoms. However, clinical visits merely provide a snapshot of a patient's condition, which fluctuates within and across days. Furthermore, recall of symptoms can be inaccurate or incomplete. During the exam, some motor aspects of the disease such as nighttime akinesia can be absent. Continuous and passive measurements of physiology and behavior could provide complementary information to clinicians, and potentially reduce the recall bias associated with retrospective surveys and diaries.

Several groups have used the accelerometers and processing power of smartphones to assess walking and hand movement, especially during maneuvers. Some studies evaluated univariate correlations between clinical scales versus features such as tremor, amplitude of leg movements, and frequency of finger tapping. More recently, machine learning approaches have been utilized to classify activities, estimate PD severity, distinguish PD from controls, or quantify movement of PD patients. We direct the interested reader to an article by Kubota, Chen, and Little that provides a review of relevant machine-learning algorithms applied to large-scale wearable sensor data in Parkinson's disease (Kubota et al. 2016).

Woods et al. developed a smartphone app that uses discrete wavelet transforms and support vector machines to discriminate between Parkinson's and essential postural tremors (ET) with over 96% accuracy (Woods et al. 2014). 14 subjects with PD and 18 subjects with ET were evaluated via the motor portion of the UPDRS. Subjects performed several motor tasks while holding a smartphone to record tremor: holding phone with eyes open and closed, attending to active tremor hand, attending to a laser target, and counting backwards by 3. DB8 wavelet decomposition was performed to produce five frequency bins; the energy in the bands 3.5–7.0 Hz and 7.0–14.0 Hz were of particular interest as they are the dominant tremor frequencies of PD, ET and physiological tremor. Analysis of variance between PD

and ET for the six tasks showed a significant main effect of task $F(3.4, 104.8) = 6.93$, $p <$ 0.05 and a significant interaction of Group × Task $F(3.494, 104.831) = 4.709$, $p < 0.05$. The mean tremor amplitude at 3.5–7.0 Hz was significantly lower in PD than in ET for all tasks ($p < 0.05$). Resultant wavelet data were used to train an SVM to classify the type of tremor. Using five-fold cross validation, the classifier achieved an accuracy of 96.4%.

Ellis et al. developed the "SmartMOVE" app to quantify gait variability (Ellis et al. 2015). The accuracy of using a smartphone gyroscope to calculate successive step times and step lengths) was validated against two heel contact-based measurement devices: heel-mounted foot switch sensors to capture step times, and an instrumented pressure sensor mat to capture step length). 12 subjects with PD and 12 age-matched controls walked along a 26-m path during self-paced and metronome-cued conditions, with all three devices recording simultaneously. Four attributes of gait were calculated. Mixed factorial analysis of variance revealed several instances in which between-group differences (e.g., increased gait variability in PD patients relative to controls) yielded medium-to-large effect sizes. Cueing-mediated changes (e.g., decreased gait variability when PD patients walked with auditory cues) yielded small-to-medium effect sizes, whereas device-related measurement error yielded small-to-negligible effect sizes. Despite a small sample size, between-group effect sizes were greater than within-group or device-related effect sizes. However, factors that contribute to variance in outcomes are less intuitive and interpretable compared to factors that contribute to direction and magnitude of outcomes.

Kassavetis et al. recorded accelerometry using a smartphone on 14 subjects with PD who performed various tasks, e.g. holding the phone in their palm with outstretched arms, performing pronation-supination, and tapping the screen (Kassavetis et al. 2016). Metrics such as amplitude and frequency were calculated for each maneuver. Clinical severity of motor symptoms was assessed with the MDS-UPDRS. Five subscores – rest tremor, postural tremor, pronation-supination, leg agility, and finger tapping – significantly correlated with eight parameters of the data collected with the smartphone.

Albert et al. recorded accelerometry via smartphone from eight subjects with PD and 18 controls, who performed a number of different activities for at least one minute (Albert et al. 2017). These activities included walking, standing, sitting, holding, or not wearing the phone. Features extracted from these accelerometry data included statistical moments, root mean square, extremes, Fourier components, and cross product means between different accelerometry axes. Automatic feature selection was performed, although the particular method was not specified. Ten-fold record-wise cross validation was performed, and an SVM classified activities from these features with an accuracy of 96.1% for controls and 92.2% for PD patients. Regularized logistic regression achieved slightly inferior results. Next, the SVM was trained on data from healthy subjects and but applied to test set data from PD patients. This lowered classification accuracy to 60.3%. Subject-wise cross-validation on PD patients resulted in an accuracy of 75.1%, whereas the same approach but for controls resulted in an accuracy of 86.0%.

Pan et al. developed an Android app called "PD Dr" to monitor accelerometry from 40 subjects with PD (Pan et al. 2015). Subjects performed three motor performance tasks: hand

resting tremor, walking, and turning. The phone was attached to the back of the hand, ankle, and pivot leg for evaluating hand tremor, walking, and turning respectively. Hand tremor features included power of motion data between 4–6 Hz, total power of motion data from 0–20 Hz, average acceleration of motion, etc. Gait features included average gait cycle time, stride length, and acceleration, and turning features included the number of steps used to turn 360°. SVMs were trained to classify hand resting tremor from no tremor, achieving a sensitivity of 0.77 and accuracy of 0.82. Gait difficulty was distinguished from normal walking with a sensitivity of 0.89 and an accuracy of 0.81. Three Lasso-regularized logistic regression models were trained to estimate disease stage (Hoehn & Yahr score from 1-5), hand resting tremor UPDRS score, and gait difficulty UPDRS score. The correlation coefficients for PD stage, hand resting tremor, and gait difficult were $r = 0.81$, $r = 0.74$, and $r = 0.79$ respectively. However, the complexity of the measurement protocol limits the translatability of this study into real-world home and clinical environments. Asking a participant to wear a smartphone in various specific locations to assess particular activities may reduce adherence and introduce variance, whereas passive monitoring of motor function without disrupting typical behaviors or requiring specific body locations for wear may be easier for patients.

Capecci et al. identified freezing of gait (FOG) in 20 subjects with PD using smartphone-based accelerometry (Capecci et al. 2016). Subjects were asked to perform the Timed Up and Go (TUG) test while being video-recorded. Clinicians assessed the videos to identify FOG events. Power in the "freeze" band (3–8 Hz) and the "locomotor" band (0.5–3 Hz) of accelerometry data were calculated. The sum of these powers was the freeze index "FI", and the ratio of freeze to locomotor power was coined "EI" (acronym not definde in paper). The Moore-Bächlin Algorithm defined a freezing of gait event when FI and EI both exceeded one standard deviation above the mean. This approach achieved a sensitivity of 70.1% and a specificity of 84.1%. A second approach that also utilized information about step cadence achieved a sensitivity of 87.6% and a specificity of 95.0%. Of note, in this work the smartphone provided both sensing and computational function, whereas much previous work only used the smartphone for sensing. However, the performance of this approach has yet to be studied on a larger cohort with several PD types and stages.

Kostikis et al. evaluated hand tremor in 25 subjects with PD and 20 age-matched controls using an iPhone app developed in-house (Kostikis et al. 2014). A physician evaluated each participant to determine their UPDRS score. Next, subjects wore a custom-built glove case to attach an iPhone to their right hand. The app assessed four metrics of hand tremor via accelerometry, including the magnitude of acceleration and rotation rate vector. A random forest was trained via bagging to classify PD from healthy subjects, and out-of-bag results were reported. The classifier achieved a maximum AUC of 0.94, a sensitivity of 0.82, and a specificity of 0.90.

Lee et al. performed a similar but larger-scale study on 103 subjects with PD (Lee et al. 2016). Using a smartphone monitoring approach, the investigators sought to (i) validate their app against MDS-UPDRS motor assessment and the two-target tapping test; (ii) generate a prediction model for UPDRS scores; (iii) assess repeatability of the app, and (iv) examine compliance and user-satisfaction. Subjects used the app at home over three days. Features

significantly correlated with MDS-UPDRS-III ($r = 0.28 - -0.61$, $p < 0.05$), and a prediction model based on these parameters accounted for 52.3% of variation in UPDRS ($R^2 = 0.523$, $F(4, 93) = 25.48$, $p < 0.05$). 48 subjects underwent repeat assessment under identical conditions, and repeatability of features and predicted UPDRS scores was moderate with intraclass correlation coefficient of $0.584 - -0.763$ ($p < 0.05$). A follow-up survey identified that subjects were comfortable with the app.

Speech degrades as PD progresses, with voice amplitude decreasing and breathiness increasing. Clinical speech scientists have used speech signal processing algorithms to characterize dysphonias such as those that occur in PD. Smartphones can record voice, either via standalone app or passively during phone calls or merely ambient background monitoring, which suggests the possibility for remote evaluation of PD. Tsanas et al. used statistical machine-learning techniques to map features from speech signal processing algorithms – such as spectral energy or amplitude – to UPDRS scores (Tsanas et al. 2011). Specifically, 42 subjects with PD performed self-administered speech tests that did not require their physical presence in the clinic, and clinicians evaluated PD symptomatology using UPDRS. 6,000 recordings were generated, and robust feature selection algorithms (LASSO and elastic net regression) were used to select the optimal subset of the speech features. These features were used to train a random forest learner, which estimated both motor and total UPDRS scores within about two points ($p < 0.001$ via the Wilcoxon rank sum test). Interestingly, linear best fit models between dysphonia features and UPDRS scores achieved low correlations in a univariate sense, but fusing multiple weak features using a machine learning approach enabled accurate UPDRS score estimation. This same group also utilized a similar approach but for dichotomizing PD subjects from healthy controls rather than estimating UPDRS scores (Tsanas et al. 2012). 132 dysophnia features were calculated from 263 samples recorded from 43 subjects. Feature selection was performed and features were used to train random forests and SVMs, which achieved almost 99% overall classification accuracy. Estimating UPDRS scores using passive and remote sensing could enable tracking patient status outside of the clinic, whereas dichotomizing (or estimating the probability of dysphonia features being close to those from a subject with PD) could be utilized for screening high-risk individuals for further evaluation by a neurologist.

3. Wearable accelerometers

Locomotor activity is altered in neuropsychiatric illnesses, due to impaired motor function, weakness, volitional and behavioral changes, or abnormal sleep patterns and circadian rhythms (Teicher 1995). Non-invasive body-worn accelerometers can measure these changes, and were first explored for assessing circadian rhythms (Witting et al. 1990; Sadeh et al. 2002). However, continual monitoring of locomotor activity was not feasible until recently due to poor battery life, the inability to wirelessly transmit data, and low patient adherence with research-grade instrumentation. Only recently have these technological constraints been overcome. Today, personal activity monitoring devices such as fitness bracelets or patches – also known as "wearables" – are affordable and widely available to public consumers. This is partly due to the global saturation of the smartphone market, which consequently reduced the cost of manufacturing and distributing similar component parts. Today, wearables house sensors that detect heart rate, activity, ambient light, and

sleep. These devices have been used in the studies revealing disturbances in 24-hour routine and circadian rhythm associated with neuropsychiatric illnesses such as BD and schizophrenia (fig. 4). While only 2-4% of individuals in the United States have a wearable device, the market is estimated to increase to 115 million units in 2018 and generate %50 billion of revenue (Gandhi et al. 2014; Statista 2017). Here we review recent studies using wearable accelerometers to monitor neuropsychiatric illnesses.

3.1. Stress and depression

Winkler et al. used actigraphy to demonstrate bright light therapy (BLT) normalizes disturbances in the circadian rest-activity cycle associated with seasonal affective disorder (SAD) (Winkler et al. 2005). 17 SAD patients and 17 age-matched controls were treated with BLT for four weeks and monitored via wrist actigraphy. SAD patients had 33% lower total and 43% lower daylight activity in the first week compared with control subjects. Furthermore, SAD patients demonstrated altered relative amplitude and phase of the sleep-wake cycle, as well as lower sleep efficiency. BLT treatment restored these alterations, and increased both total and daylight activity of SAD patients. Interdaily stability (IS) of activity – a measure of regularity of the circadian rhythm between days, which reflects the strength of coupling of the rhythm to external cues such as light – increased by 9% in both patients and controls.

Nakamura et al. measured activity in 14 patients with MDD and 11 age-matched controls using wrist-worn accelerometers (Nakamura et al. 2007). Data from resting and active periods were fit to cumulative power law distributions with the form $P(x \quad a) \sim a^{-\gamma}$. A period of activity was considered resting or active if the counts were cumulatively below or above a threshold value, respectively. The average scaling exponent $\bar{\gamma} = 0.74 \pm 0.12$ among depressed patients and $\bar{\gamma} = 0.92 \pm 0.08$ among controls. This difference was associated with a significantly longer mean resting period duration in depressed patients (15.64 ± 6.19 minutes) than in controls (7.72 ± 1.44 minutes).

Vallance et al. studied the associations between accelerometer-derived physical activity and sedentary time with depression in 2,862 adults from the 2005–2006 US National Health and Nutrition Examination Survey (NHANES) (Vallance et al. 2011). Depression occurred in 6.8% of the sample and was assessed via the PHQ-9 questionnaire. Depressed and non-depressed subjects significantly differed across several demographic, medical, and behavioral characteristics. Subjects wore accelerometers (ActiGraph, Ft. Walton Beach, FL) for seven days on their right hip, attached by a belt. Subjects in the highest quartile of moderate-to-vigorous intensity physical activity had 2.7-fold lower odds of depression compared to subjects in the lowest quartile. Sedentary time was associated with significantly increased odds of depression in overweight/obese adults, but not normal weight adults. However, only simple measures of activity were computed. More sophisticated metrics reflecting structure of routine, complexity, and disturbances in circadian rhythms may better distinguish depressed from healthy people.

Sano et al. searched for physiological or behavioral markers for stress by collecting 5 days of data from 18 subjects (Sano et al. 2013). A wrist sensor monitored accelerometry and skin conductance, and a custom app tracked call, SMS, location, and screen on/off time. The app

also administered surveys to assess stress, mood, sleep, tiredness, general health, alcohol or caffeinated beverage intake, and electronics usage. Using features screen on, mobility, and call or activity levels, an SVM classified subjects into high or low perceived stress groups with 75% accuracy. Higher reported stress level was related to activity level, SMS and screen on/off patterns.

Sano et al. repeated this approach in a subsequent study, but also included academic performance among the outcomes (Sano et al. 2015). Subjective and objective data was gathered using mobile phones, surveys, and wearable sensors worn by 66 subjects for 30 days. Sequential forward feature selection was performed to find the best combinations of features derived from data including survey scores (from surveys that were not the outcome being estimated), accelerometry, skin temperature, calls and SMS, and internet and email use. An SVM achieved a classification accuracy of 90% in distinguishing individuals in the upper 20th percentile from the lower 20th percentile in three different scores: Pittsburg Sleep Quality Index, perceived stress scale, and mental health composite. Features from wearable devices resulted in higher classification accuracy than when using features from smartphones, with the exception of classifying high versus low GPA.

O'Brien et al. monitored activity in 29 older adults with MDD for seven days. Subjects underwent neuropsychological assessment and quality of life (QoL) (36-item Short-Form Health Survey) and activities of daily living (ADL) scales (Instrumental Activities of Daily Living Scale) (O'Brien et al. 2016). A wrist-mounted actigraph used in this study was developed as part of the OpenMovement initiative, a collection of open-source hardware sensors and software tools for research use (https://openlab.ncl.ac.uk/things/open-movement/). Physical activity, jerk, and entropy were all significantly lower in depressed subjects, and reductions in locomotor activity were associated with reduced ADL, lower quality of life, lower associated learning, and a higher depression rating scale score.

Burton et al. performed the first systematic review of activity monitoring in patients with depression and highlighted several limitations of the studies (Burton et al. 2013). 19 eligible studies were identified, and case control studies showed less daytime activity in patients with depression compared to controls (standardized mean difference −0.76, 95% confidence intervals −1.05 to −0.47). However, most studies in the literature contained a source of potential confounding by comparing inpatient subjects with depression to controls in the community. Large differences between groups could have been due to living environment and routine rather than depression. Outpatients have milder forms of depression and less co-morbidity than those admitted to hospital, so results from studies that do not account for treatment setting may not generalize to broader patient populations. Furthermore, not all studies mentioned the make and model of the wrist-worn actigraph used, and the duration of sampling varied considerably from study to study, ranging from 12 hours to > 2 weeks.

3.2. Bipolar disorder

Palmius et al. reported preliminary results from the AMoSS study involving 16 subjects with BD, nine subjects with borderline personality disorder, and 25 controls (Palmius et al. 2014). Subjects were provided with a Samsung Galaxy SIII phone running the Android operating system, a Fitbit One wireless activity and sleep tracker, and a GENEActiv Original wrist-

worn accelerometer. Subjects also took their own temperature and blood pressure at specific times in the study. Furthermore, an ECG was worn for a limited period of time by each participant and selected subjects wore a pulse oximeter overnight. The AMoSS app recorded actigraphy, ambient light, phone call time and duration, SMS time and length, self-reported mood surveys, and blood pressure and temperature recorded from Bluetooth-connected devices. Clinically validated psychiatric questionnaires – including the quick inventory of depressive symptomatology (QIDS), Altman self-rating mania scale, generalized anxiety disorder (GAD-7), and EuroQol EQ-5D – were administered weekly. Although data collection was still ongoing, gross differences in the regularity of actigraphy between borderline personality disorder and controls were evident.

Bullock et al. measured locomotor activity over seven days in 36 subjects with high trait vulnerability and 36 subjects with low trait vulnerability for BD (Bullock et al. 2014). Patients wore wrist actigraphs (Mini-Mitter Actiwatch-L; Respironics, Inc., Bend, OR). Vulnerability to BD was determined via a self-reported General Behavior Inventory (GBI), taken by 358 potential subjects. The top and bottom 10% of the GBI distribution formed the high- and low-vulnerability groups. Relative amplitude (RA) is defined as:

$$RA = \frac{M10 - L5}{M10 + L5}$$

where M10 is the mean activity level during the most active ten hours, and L5 is the mean activity level during the least active five hours. RA was lower in the high-vulnerability group than in the low-vulnerability group, while IV and IS did not significantly differ.

Krane-Gartiser et al. evaluated actigraphy recordings from 18 hospitalized patients with mania, 12 hospitalized patient with bipolar depression, and 28 controls (Krane-gartiser et al. 2014). Actiwatch Spectrum actigraphs were used. From each participant, the first period of 64 minutes containing two consecutive minutes of zero activity counts were selected for subsequent feature extraction. Features included the standard deviation (SD), Root Mean Square of the Successive Differences (RMSSD), autocorrelation (with lag of one), sample entropy, "symbolic dynamics" (a simplified version of sample entropy; Guzzetti et al. 2005), and "Fourier analysis" (ratio between variance in high and low frequencies of the spectrum). Mean activity of both manic and depressed patients was lower than in controls. SD/min in % of mean was higher in depressed patients than in both manic patients and controls. Although no difference was found in sample entropy of activity by patient group, symbolic dynamics were lower in depressed patients compared to manic patients and controls. Finally, the ratio between variance of high frequency (HF) power and variance of low frequency (LF) power was higher, and the autocorrelation was lower, in manic patients compared to other groups.

Hyperactivity is seen in both pediatric BD and attention-deficit hyperactivity disorder (ADHD). Faedda et al. accurately distinguished youth with BD (N=48) from those with ADHD (n=65) and typically developing controls (n=42) using features derived from five minutes of belt actigraphy data (Faedda et al. 2016). Features were selected on the basis of significance rather than predictivity, and included diurnal skew, L5, RA, and bipolar vulnerability index (VI). VI is the integrated area of shape coefficients of the gamma

function fit to the distribution of Morlet wavelet coefficients at scales from 0.2–2 hr. Bagging was performed whereby 75% of the data was used as the training set and model performance was assessed using the remaining 25% of the data as the test set. This was repeated 500 times. Although the cohort included three classes of subjects, a binary classification task was performed to distinguish BD from non-subjects with BD. Several classifiers were used, including random forest, artificial neural networks, SVM, multinomial regression, and partial least squares, achieving area under the ROC curves ranging from 0.75 to 0.78.

3.3. Schizophrenia

Martin et al. found older schizophrenia patients have more disrupted sleep and circadian rhythms (Martin et al. 2005). 28 older schizophrenia patients (mean age=58.3 years) and 28 age- and gender-matched controls were monitored for three days using Actillume wrist actigraphs (Ambulatory Monitoring, Inc., Ardsley, New York). Minute-by-minute activity and light exposure were recorded. Patients spent longer in bed, had more disrupted nighttime sleep, slept more during the day, and had less robust circadian rhythms of activity and light exposure compared to controls.

Apiquian et al. evaluated rest-activity characteristics in 20 unmedicated and non-hospitalized schizophrenia patients and 20 controls for five days using a wrist-worn actigraph (Actiwatch-16) (Apiquian et al. 2017). Compared to controls, untreated patients showed significantly lower levels of motor activity and more sleep time.

Walther et al. investigated the relationship between objective measures of motor activity and PANSS scores (Walther et al. 2009b). 55 schizophrenia patients were monitored for 24 hours via wrist actigraphy. Low activity levels were correlated with high PANSS negative syndrome subscale scores. Interestingly, actigraphic parameters did not correlate with motor-specific questions of the PANSS, challenging the validity of the questionnaire.

This same research group subsequently used 24-hour actigraphy to differentiate schizophrenia subtypes in a cohort of 60 hospitalized patients (35 paranoid, 12 catatonic, 13 disorganized) (Walther et al. 2009a). Activity level and movement index (proportion of 2-second periods with nonzero activity) was highest in paranoid schizophrenics, whereas the mean duration of uninterrupted mobility was highest in catatonic schizophrenics.

Berle et al. used actigraphy to evaluate patterns of motor activity in 23 schizophrenia patients, 23 depressed patients, and 32 control subjects who did not have a history of mood or psychotic systems (Berle et al. 2017). Total motor activity was lower in patients diagnosed with schizophrenia or depression than in controls. However, IS was 18% higher in schizophrenia patients compared to controls, whereas IS did not differ between depressed patients and controls. IV was 18% lower in schizophrenia patients and 8% lower in depressed patients compared to controls.

Hauge et al. revisited this same cohort of patients, but analyzed activity data using Fourier analysis and entropy measurements (Hauge et al. 2011). For each patient, these features were derived from the first 300-minute segment of activity data that contained   4

consecutive minutes of zero activity. RMSSD/SD was significantly lower in schizophrenia patients compared to either depressed patients or controls. Sample entropy of activity was significantly lower in depressed patents compared to either schizophrenia patients or controls. Finally, the ratio between variance of HF power and variance of LF power was significantly higher in depressed patients compared to controls.

Wichniak et al. recorded seven days of actigraphy using Actiwatch AW4 devices (Cambridge Neurotechnology Inc., UK) in 73 patients with schizophrenia and 36 age-and sex-matched controls (Wichniak et al. 2011). Mental status was measured via the PANSS and CDSS questionnaires. Schizophrenia patients had lower mean 24-hour activity and mean 10-hour daytime activity levels, and spent more time in bed. Lower activity was associated with higher PANSS and CDSS scores.

Sano et al. recorded seven days of actigraphy (Actigraph Mini-Motionlogger; Ambulatory Monitors Inc., Ardsley, NY, USA) in 19 schizophrenia patients and 11 controls (Sano et al. 2012). Resting periods obeyed a power-law cumulative distribution whereas active periods obeyed a stretched exponential distribution. Distribution parameters differed among schizophrenia patients and controls. For resting periods, the average scaling exponent values (mean $\pm$ standard deviation) were $\bar{\gamma} = 0.86 \pm 0.03$ for schizophrenia patients and $\bar{\gamma} = 0.99 \pm 0.03$ for controls. For active periods, the average stretching parameters were $\bar{\beta} = 0.57 \pm 0.02$ for schizophrenia patients and $\bar{\beta} = 0.64 \pm 0.02$ for controls.

Evaluating the distribution of rest-activity periods was also previously described by Nakamura et al. 2007 and Sano et al. 2012. Fasmer et al. 2015 used this approach in a cohort of 24 patients with schizophrenia, 23 with depression, and 29 controls. 12 days of actigraphy data were recorded per patient using Actiwatches (Cambridge Neurotechnology Ltd., England, UK). For active periods, average scaling exponent values (mean $\pm$ standard deviation) were $\bar{\gamma} = 0.77 \pm 0.13$ for schizophrenia patients, $\bar{\gamma} = 0.88 \pm 0.13$ for depressed patients, and $\bar{\gamma} = 0.82 \pm 0.01$ for controls. For inactive periods, average scaling exponent values (mean standard deviation) were $\bar{\gamma} = 0.81 \pm 0.17$ for schizophrenia patients, $\bar{\gamma} = 0.93 \pm 0.18$ for depressed patients, and $\bar{\gamma} = 0.71 \pm 0.11$ for controls. Length of active and inactive periods and scaling exponents for both active and inactive periods correlated with IS, whereas only length of active periods and scaling exponents for inactive periods correlated with IV. The authors concluded the distribution of active and inactive periods differed in depressed compared to schizophrenic patients.

Shin et al. assessed correlations between locomotor activity and symptom severity of 61 subjects with schizophrenia (Shin et al. 2016). Subjects wore a Fitbit Flex device for a week to assess their activity, and completed the PANSS questionnaire to assess schizophrenia symptoms. Subjects with a high total PANSS score or high positive subscale scores had significantly lower levels of physical activity than the other groups.

3.4. Dementia and Alzheimer's disease

Kuhlmei et al. evaluated the relationship between actigraphy features and measures of apathy in two groups of patients with mild cognitive impairment (MCI) and dementia (Kuhlmei et al. 2013). The cohort consisted of 32 patients with dementia, 21 patients with

MCI, and 23 elderly controls. Apathy and depression were evaluated via the Apathy Evaluation Scale (AES) and the Beck Depression Inventory (BDI), respectively. Apathy (indicated by a higher AES score) was associated with reduced daytime activity regardless of diagnosis ($r = -0.50$, $p < 0.01$). This effect was greater in patients with dementia than in patients with MCI.

3.5. Parkinson's disease

Several studies have evaluated locomotor activity in patients with (Parkinson's disease) PD using non-phone based accelerometers, and have focused on the identification of specific motor behaviors such as gait alterations, freezing of gait, and balance deficits (Maetzler et al. 2013). Here we review papers specifically describing the use of wrist- or finger-worn sensors for monitoring PD.

Patel et al. used wearable accelerometry to estimate the severity of motor dysfunction in PD patients (Patel et al. 2009). 12 patients with mild to moderate PD were recruited for the study. Accelerometers were placed on two locations per extremity (upper and lower). Subjects performed several motor tasks including finger to nose movements, finger tapping, and sitting. Features extracted from accelerometer data included range of amplitude, root mean square value, cross-correlation metrics, frequency-based metrics, and entropy. Video of study subjects was annotated by clinical experts. These annotations served as multi-class labels for an SVM, which was trained on features derived from an optimal window length of at least 5 seconds of accelerometer data. The classifier achieved 5-7% estimation error.

Stamatakis et al. estimated Unified Parkinson's Disease Rating Scale (UPDRS) scores using features derived from a finger-tapping test recorded via finger-worn accelerometer from 36 PD patients and ten age-matched controls (Stamatakis et al. 2013). Features included mean movement frequency, number of halts and hesitations, acceleration, and angle. Leave-one-out cross-validation and greedy backwards feature selection was performed, and a logistic regression model was trained to estimate UPDRS scores (ranging from 0 to 4 where 0 indicates no problems with the motor task and 4 indicates inability to perform the task due to slowing or interruptions) with AUCs ranging from 0.92 to 0.97.

Roy et al. measured unscripted activity in 11 PD patients using tri-axial accelerometry and surface electromyography (sEMG) (Roy et al. 2011). Sensor and video data were recorded for four hours; the video data was annotated by individuals trained to identify PD motor signs. Accelerometry features included lowpass and highpass energy, lag of first peak of autocorrelation, and ratio of height of first peak to peak at origin in autocorrelation. sEMG features included the root mean square energy, as well as the lag and peak ratio (as described for accelerometry). These features were used to train a dynamic neural network to detect tremor and dyskinesia, as well as severity. The classifier was tested on a validation set of four controls and eight patients not used to train the model. The neural network achieved sensitivities of 95.2% - 97.2% for detecting tremor and 91.9% - 95.0% for detecting dyskinesia. These results demonstrate accurate classification of PD motor activities, particularly using training data from different patients than test data. However, it is unclear if similar performance is achievable using commercial wrist-worn fitness trackers, or without sEMG data.

Griffiths et al. recorded activity continuously for ten days in 34 PD subjects with idiopathic Levodopa-responsive PD, and from 10 age-matched controls (Griffiths et al. 2012). Parkinson's Kinetigraphs (PKG; Global Kinetics Corporation) were used to record activity. Unified Parkinson Disease Rating Scale (UPDRS) scores were obtained for each subject. A bradykinesia score (BKS) was produced by establishing the maximum acceleration in each 2-minute epoch of acceleration recordings and calculating the mean spectral power (0.2 and 4 Hz) surrounding this peak. A dyskinesia score (DKS) was calculated similarly except using the mean acceleration rather than the maximum. BKS and DKS values significantly differed in PD patients versus controls. Activity-based estimates of clinical dyskinesia were quantitatively comparable to those from three neurologists. Finally, improvement in scores in response to changes in medication could be assessed in individual patients.

Niwa et al. investigated rest activity and autonomic function by evaluating actigraphy and ECG data in a cohort of 27 PD patients and 30 age-matched controls (Niwa et al. 2011). Actigraphy was recorded for seven days using Mini-Motionlogger Actigraphs (Ambulatory Monitoring, Ardsley, NY, USA). Ambulatory ECG was also recorded for 24 hours. Nine PD patients classified as Hoehn-Yahr (HY) stage 1 or 2 and were considered early stage, and 18 PD patients classified as HY stage 3 or 4 and were considered late stage. Disease duration, medication status, score on the Unified Parkinson's Disease Rating Scale (UPDRS), and score on the Mini Mental State Examination (MMSE) were determined. Sleep episodes out of bed and wake episodes in bed were higher in the PD patients than in the control subjects. Several rest activity features including activity of duration, mean activity, and ratio of out-of-bed versus in-bed activity significantly differed by PD status. HRV analysis revealed a decline in total power and HF with increasing stage of PD. In summary, these results show alterations in circadian rhythm, locomotor activity, and autonomic nervous system function associated with PD.

Because PD is a movement disorder, the relationship between accelerometer data and intensity of physical activity may differ in healthy people versus those with PD. Nero et al. 2015 defined accelerometer cut points for different walking speeds in 30 older adults with mild-to-moderate PD (mean age=73 years). Subjects wore a waist accelerometer and walked at self-defined brisk, normal, and slow speeds. Walking speed was also measured. Through receiver operating characteristic analysis, cut points were generated for different levels of walking speed in counts per 15 seconds. Sensitivity and specificity were 61-100%.

Kheirkhahan et al. estimated a clinically meaningful mobility phenotype of walking speed using waist-worn accelerometers in a cohort of 1,135 older adults with impaired mobility (Kheirkhahan et al. 2016). Although this study was not specific to PD, it is relevant given how mobility is impaired in PD and is a major predictor of overall health in older adults. Regularized linear regression was performed to estimate walk speed from features; the best model achieved a sensitivity of 0.70, a specificity of 0.80 and a classification rate of 76%. The most predictive features were average counts per min during active periods, average length of bout of activity, average activity count where half of the total activity is accumulated, standard deviation of activity counts, and number of steps per day. Interestingly, these features were not universally correlated with walking speed in univariate

analyses, which suggests that the combination of features may result in better classification performance compared to using single features.

4. Holter monitoring

Much literature has established a bidirectional relationship between changes in HRV and neuropsychiatric illness. People with severe mental illness have worse cardiovascular outcomes than healthy controls, and people with serious cardiovascular illness are more likely to develop certain neuropsychiatric illnesses (Newcomer et al. 2007; Sowden et al. 2009). The interplay between mental and cardiovascular health is believed to be mediated by alterations in the ANS, endocrine effectors such as cortisol and catecholamines, activation of pro-inflammatory cytokines, and lifestyle and environmental exposures such as diet, exercise, and social support (Grippo et al. 2009). In particular, HR and HRV measures can be measured noninvasively, and reflect the state of the ANS. Alterations in HR and HRV may thus provide an objective and passively measurable marker of clinical status in neuropsychiatric illnesses ranging from MDD to BD to schizophrenia (Henry et al. 2010; Cohen et al. 2000).

Cardiac monitoring from body-worn instrumentation is known as Holter monitoring and was originally performed via large stationary ECG devices. Recently, body-worn patches adhering to the skin have been developed to measure HR via ECG, actigraphy, and even metabolites in sweat (Rodgers et al. 2015). Adhesive patches have the potential to improve adherence with study protocols and device use because they are unobtrusive and always attached to the patient. Most studies using physiological patches have focused on heart disease, although a few groups have used this technology to evaluate patients with depression, stress, and schizophrenia. Photoplethysmography (PPG) is another approach for assessing HR via optical measurements of changes in blood volume, and has become a popular sensing technique in wearable devices such as fitness bracelets (Allen 2017). Here we summarize several studies that exclusively focus on the analysis of heart rate data, measured via both traditional ECG as well as patch-based sensing modalities. Devices utilizing PPG are reported in the next section on multi-modal sensing.

4.1. Stress and depression

Previous studies of the neural correlates of vagal tone – as measured by HF components of the power spectral density of HR – involved mental stress tasks that affect both cognitive and emotional states. Lane et al. 2009 studied the contributions of emotion to vagal tone by correlating HF power with measures of regional cerebral blood flow (rCBF) derived from positron emission tomography (PET) in 12 healthy women during different emotional states. Happiness, sadness, disgust and three neutral conditions were each induced by film clips and recall of personal experiences. ECG was recorded during each 60-second PET scan, and HF HRV was calculated. HF HRV correlated with rCBF in regions of the brain involved in emotion: medial prefrontal cortex, caudate nucleus, periacqueductal gray, and left mid-insula. These data support the hypothesis that emotion – an important component of the psyche that is affected in neuropsychiatric illness – mediates vagal tone which manifests as altered HRV.

Short-term psychological stress has been shown to modulate sympathovagal activity, which is measurable via Holter monitoring. Delaney et al. 2000 assigned 30 healthy subjects into two age- and sex-matched groups. A psychological stress test was administered in a competitive setting and included a financial incentive to produce psychological strain. Psychological stress decreased the standard deviation of interbeat intervals, increased heart rate, reduced HF power, and increased LF power.

Lee et al. developed a three-electrode ECG patch with built-in R-peak detection, and demonstrated comparable noise quantity and signal-to-noise ratios between their device and a commercial ECG system (Lee et al. 2016). HRV measures calculated from data acquired via the ECG patch was also used to evaluate stress responses in 17 subjects who were administered a color–word interference test and a mental arithmetic test. Both the patch and commercial ECG showed elevated LF/HF ratios in subjects undergoing stressful conditions.

Weenk et al. evaluated stress in twenty surgeons and residents using a smart patch. Subjects filled out the State Trait Anxiety Inventory (STAI) before and immediately after each surgical procedure (Weenk et al. 2017). The patch (Vital Connect, Campbell, CA, USA) was worn for 48 hours, and measured single-lead ECG, respiratory rate, skin temperature, body posture, activity, and steps. Calculated HRV measures included the standard deviation of the average normal-to-normal intervals (SDNN), root mean square of the successive differences (RMSSD), very low frequency power (VLF), LF, and HF power. Stress (%) was estimated using a simple, empirical linear model: HR+a SDNN. Performing surgery decreased SDNN and RMSSD by 40% each, increased the ∗LF/HF ratio by 64%, and increased the stress percentage by 300%. Estimated stress was higher in less experienced surgeons.

Nahshoni et al. compared HRV measures from ten subjects with MDD to those in ten mentally healthy heart transplant recipients, as well as ten healthy control subjects (Nahshoni et al. 2004). ECG was used to record 2,000 interbeat intervals from each subject after 10 minutes of lying supine. No significant differences were noted between MDD and transplant subjects, but both of those groups had lower interbeat intervals, standard deviation of RR intervals, and pointwise correlation dimension (a nonlinear measure of information content of a dynamical system) compared to healthy controls.

Roh et al. developed a soft flexible patch to obtain HRV measures to evaluate patients with depression (Roh et al. 2014). The patch features a rechargeable battery and an integrated reduced instruction set computer which can perform ECG acquisition, signal filtering, R peak extraction, and feature extraction including time- and frequency domain and nonlinear features such as entropy. ECG recordings from 41 volunteers were annotated by experienced researchers; the integrated patch software detected R peaks with a sensitivity of 99.3%, PPV of 100.0%, and error of 0.71%. The patch was also compared to a conventional Holter system by recording ECG data from 12 adult volunteers performing various activities; signal-to-noise levels were comparable, and actually higher in measurements from the patch during walking speeds > 1 km/h. Finally, HRV analysis was performed in 17 adult volunteers using ECG data measured via the patch and a Holter monitor for comparison. Subjects were recorded in three states: 1) rest, 2) while performing the Stroop test – a color-word interference task meant to stimulate cognitive stress – and 3) while performing a

mental arithmetic challenge. HRV parameters differed with mental task, but results from the patch and the Holter monitor were equivalent.

4.2. Bipolar disorder

Changes in HRV that are measurable via Holter monitoring have also been observed in bipolar disorder (BD). Henry et al. recorded cardiac activity via ECG and assessed HRV in 23 acutely hospitalized subjects with manic BD, 14 subjects with schizophrenia, and 23 healthy age- and gender-matched controls (Henry et al. 2010). Time domain, frequency domain, and nonlinear HRV measures were calculated. Psychiatric symptoms were assessed by administration of the BPRS and the YMRS. Compared to controls, subjects with BD had significantly higher mean resting HR, lower root mean square of the successive differences (RMSSD) of RR intervals, lower HF power, and lower entropy. Reduction in parasympathetic tone significantly correlated with higher YMRS scores and the unusual thought content subscale on the BPRS, whereas decreased entropy was associated with increased aggression and diminished personal hygiene on the YMRS scale. In summary, autonomic dysfunction is associated with more severe psychiatric symptoms and may depend on the phase of the illness.

4.3. Schizophrenia

Cardiovascular mortality is elevated in patients with schizophrenia, which may be due to increased prevalence of obesity, smoking, and diabetes, adverse pro-arrhythmic effects of antipsychotic medication, and altered autonomic function. Bär et al. 2017 calculated complexity measures of HRV using short-term ECG recordings from 20 unmedicated subjects with schizophrenia and 20 age- and gender-matched healthy controls. Features included joint symbolic dynamics, compression entropy, fractal dimension and approximate entropy. Complexity of HR time series was significantly reduced in acute schizophrenia. However, when using HR as a covariate, only fractal dimension remained significantly altered.

4.4. Post traumatic stress disorder

Cohen et al. evaluated frequency-domain HRV measures via power spectral density analysis using ECG recordings from 14 subjects with post traumatic stress disorder (PTSD), 11 subjects with panic disorder, and 25 matched controls (Cohen et al. 2000). ECG recordings were made while subjects were resting while recalling the trauma implicated in the development of their PTSD, or the circumstances of a severe panic attack, as appropriate, and again while resting. Controls were asked to recall a stressful life event during recall. Both PTSD and panic disorder groups had elevated HR and low frequency LF power at baseline, suggesting increased sympathetic activity. However, PTSD patients did not respond to recall stress with increases in HR and LF.

Reinertsen et al. used a machine learning approach to dichotomize subjects with PTSD from healthy controls using features such as LF power, statistical moments, and acceleration and deceleration capacity (Reinertsen et al. 2017a). 24-hour single-channel ECG recordings were obtained from 23 subjects with current PTSD, and 25 control subjects with no history of PTSD. RR intervals derived from these data were cleaned and used to calculate HR and

HRV features – including statistical moments, power spectral density components, entropy, and acceleration / deceleration capacity – which were used to train a logistic regression classifier. Performance was assessed via repeated random sub-sampling validation. To reduce noise and activity-related effects, features were calculated from five non-overlapping ten-minute quiescent segments of RR intervals defined by lowest HR, as well as random ten-minute segments as a control method. Feature selection was performed and a median AUC of 0.86 was achieved out-of-sample test set data. This was significantly higher than the AUC using 24 h of data (0.72) or random segments (0.67), demonstrating the utility of a novel HR segmentation approach for improving the classification of PTSD from HR and HRV measures. Further work should prospectively evaluate if classifier output changes significantly with worsening symptomatology or effective treatment of PTSD.

4.5. Alzheimer's disease

Zulli et al. evaluated HRV and other cardiovascular measures via Holter ECG monitor in a cohort of 33 subjects with AD, 39 subjects with mild cognitive impairment (MCI), and 29 cognitively healthy controls (Zulli et al. 2005). QT interval dispersion values were significantly higher in patients with AD than in patients with MCI or controls. Furthermore, HRV time and domain parameters were lower in patients with AD than in patients with MCI and controls, and these differences varied with levels of cognitive impairment. This autonomic cardiac dysfunction may be related to a cholinergic deficit. However, Allan et al. 2005 did not find differences in HRV measures in AD patients. ECG recordings were obtained from 14 AD patients, 20 vascular dementia patients, and 80 healthy controls resting in the supine position for five minutes. Power spectral analysis was performed to calculate VLF, LF, HF, and total power, but no measures differed in AD or vascular dementia compared to controls. Further research is needed to elucidate the interplay between cholinergic and cognitive defects in AD and changes in ANS function, and to evaluate if noninvasive sensing of these alterations can be clinically useful.

4.6. Parkinson's disease

Because dysautonomia is a known characteristic of PD, alterations in complexity and frequency-domain HRV measures may be detectable using Holter monitoring approaches and could reflect clinical status of patients (Poewe et al. 2017). Kallio et al. obtained ECG recordings from 50 patients with PD and 55 healthy controls during normal breathing, paced breathing, Valsalva maneuver, upright tilting, and isometric handgrip (Kallio et al. 2000). Sub-measures of RMSSD, and systolic blood pressure after tilt test, significantly differed in PD patients versus controls. Patients with hypokinesia or rigidity as the initial symptom of PD had a more pronounced HRV deficit than those with tremor onset. This difference might be due to more advanced neuronal damage in hypokinesia or rigidity onset PD, and/or preferential involvement of regions of the central or peripheral ANS that mediate autonomic function.

Previous studies of autonomic responses in PD measured HR over relatively short time periods, which provides a limited view of the autonomic cardiac control mechanisms and do not represent tonic autonomic regulation. Furthermore, PD patients exhibit alterations in various circadian autonomic patterns, such as body temperature and heart rate variation,

which would require longer monitoring times to assess (Brown et al. 2012). Haapaniemi et al. sought to evaluate tonic autonomic regulation by recording ECG via Holter monitor over 24 hours in 54 untreated patients with PD and 47 age-matched healthy controls (Haapaniemi et al. 2001). Power spectral features, instantaneous beat to beat variability, long term continuous variability derived from Poincaré plots, and the slope of the power law fit of the RR intervals were analyzed. All spectral components and the slope of the power law curve were lower in the patients with PD than in controls. UPDRS and motor scores negatively correlated with VLF and LF power and the slope of the power law. Patients with mild hypokinesia had higher HF values than patients with more severe hypokinesia, whereas tremor and rigidity were not associated with HRV measures.

A later study by Oh et al. also evaluated 24-hour Holter ECG and BP recordings of 139 patients with PD and 55 age-matched controls (Oh et al. 2014). There were significant differences in the distribution of non-dipping, the percent of nocturnal BP decrease, nighttime BP level, the standard deviation of heart rate, and nocturnal decrease of heart rate between patients with PD and controls. However, these abnormal diurnal HRV and BP measures were not associated with motor symptom severity, age, gender, or disease duration.

5. Multimodal sensing

Here we review studies that utilize heart rate in addition to other sensor types, including accelerometry, ambient light, and GPS. A patient with milder severity of an illness such as schizophrenia may not demonstrate significant alterations in accelerometry or heart rate-derived features in a univariate sense, but multiple weak features can be aggregated together to train a classifier that accurately infers symptomatology or clinical status. However, commercially available devices with physiologically and behaviorally relevant sensing technologies of high accuracy have only recently reached the market. Furthermore, awareness of the utility of conglomerating several weaker signals is more prevalent amongst machine learning practitioners than statisticians and clinical investigators. Relatively few studies employ multi-sensor fusion approaches, and even fewer focus on neuropsychiatric illness.

Kamdar et al. explored the prediction of emotional state from accelerometry, ambient light, and heart rate data measured via Samsung Gear S smartwatches (Kamdar et al. 2016). Data was collected from 13 healthy subjects in a pilot test. A web app was also developed for users to self-report moods via a Likert scale rating of happiness, energy, and relaxation. The app also capture user keystrokes and mouse patterns. Each subject wore the Gear S watch for at least 6 hours and entered at least three insights over a single day. Several machine learning algorithms were trained using these features: random forest, gradient boosted regressor trees, regularized logistic regression, SVM, and k nearest neighbors. A random forest model explained 51% of the variance of emotional state from device-captured data. However, top features were derived primarily from user interactions rather than passively monitored physiology. Furthermore, no classifier accuracy metrics – such as AUC of classification of mood status – were reported, and the authors also reported high levels of variance in HR measured with the watch compared to a direct pulse measurement, although the latter method was not specified.

AlHanai et al. used a combination of auditory, text, and physiological signals to predict the mood (happy or sad) of 31 narrations from ten subjects as they told either happy or sad stories (AlHanai et al. 2017). Subjects wore wrist-mounted Samsung Simband devices which recorded PPG, ECG, accelerometry, skin impedance, galvanic skin response, and skin temperature. Audio was recorded using Apple iPhones. 386 audio and 222 physiological features were calculated from the data. A subset of 4 audio, 1 text, and 5 physiologic features were identified using sequential forward feature selection: subject movement, cardiovascular activity, energy in speech, probability of voicing, and linguistic sentiment (i.e. negative or positive). A deep neural network was trained using these features to classify if the story was happy or sad. To ensure the real-time utility of the model, classification was performed over 5 second intervals. Model performance was assessed via leave-one-subject-out cross-validation, and the classifier achieved a mean AUC of 0.92.

Osipov et al. measured HR and accelerometry in 16 subjects with schizophrenia and 19 controls using an adhesive monitoring patch (Protues Digital Health, Redwood City, CA) (Osipov et al. 2015). Features calculated on both types of data included basic summary statistics – mean, median, mode, and variance – as well rest-activity characteristics (Van Someren et al. 2009), multiscale sample entropy (Costa et al. 2002), and multiscale transfer entropy (Schreiber 2000). An SVM learned to dichotomize subjects as either having a diagnosis of schizophrenia or being a control. Two-fold cross-validation with repeated random sub-sampling was performed 1000 times. Using HR features resulted in an AUC of 0.85, whereas using activity features resulted in AUC of 0.90. Using both HR and activity features resulted in an AUC of 0.99.

Reinertsen et al. measured HR and locomotor activity in 12 medicated subjects with schizophrenia and 12 healthy controls, and classified contiguous days of data as belonging to a schizophrenia patient or a healthy control (Reinertsen et al. 2017b). Subjects were monitored for 3–4 weeks using a disposable adhesive patch sensor worn on the chest and manufactured by Proteus Biomedical (Redwood City, CA). Features derived from time series data included classical statistical characteristics, rest-activity metrics, transfer entropy, and multiscale fuzzy entropy. The analysis window length, or number of days of data considered per record, was varied from two to eight days. An SVM was trained with these features to classify records as belonging to either a schizophrenia or control subject. Model performance was assessed via subject-wise leave-one-out-crossfold-validation. An analysis window length of eight days resulted in a high AUC of 0.96. Reducing the analysis window length to two days only lowered the AUC to 0.91. The type of most predictive features varied with analysis window length. Classifier output may have represented illness severity or level of ANS dysfunction, although verifying this in future work will require gathering information about symptoms on a daily basis.

Cella et al. monitored 30 subjects with schizophrenia and 25 controls using wrist-worn Empatica E4 devices which measured skin conductance, PPG (from which RR intervals were derived), and accelerometry (Cella et al. 2017). Symptom severity in subjects with schizophrenia was assessed via the PANSS questionnaire. Subjects were monitored for six days, and recordings < 60 minutes were excluded. At least two 8-hour recordings were obtained for each subject, with an average of 3-4 8-hour recordings obtained per subject.

Skin conductance did not vary by patient group, but subjects with schizophrenia had significantly lower SDNN and RMSSD values, as well as lower locomotor activity and fewer hours of structured activity, compared to controls. Chlorpromazine levels were not found to significantly affect any physiological measures.

6. Challenges and Limitations

Extracting clinical insights from sensor data is complicated by many challenges such as noise, insufficient sampling frequency, a lack of standardization and calibration of sensors, signal from phenomena unrelated to illness, the need to integrate with the electronic medical record, and the limited role of regulatory agencies governing the deployment of these technologies.

Noise in ECG data is due to poor contact between the electrode and the skin, patient movement, muscle activity, or power line interference (Clifford 2002). Accelerometer recordings can be noisy due to thermal energy, mechanical vibrations, and the location and manner in which the device is worn (Cemer 2011). Estimation of signal quality indices and data fusion approaches can be used to detect poor quality ECG data, and these methods may also be applicable to other types of digital sensor data (Clifford et al. 2011; Clifford et al. 2012a).

Digital sensors allow for measuring physiology at or above the Nyquist frequency, avoiding the common issue of aliasing, which occurs when data is recorded at a rate less than twice the highest frequency in the signal (Clifford et al. 2012b). A sampling rate of 3-6 Hz for heart rate and 10 Hz for movement is usually sufficient for satisfying the Nyquist criterion (Winter et al. 1972; Clifford 2002). However, manufacturers of consumer devices may prioritize battery life over sampling frequency, and the latter attribute is often not reported in product documentation.

Lack of standardization and calibration hinders comparisons across studies, and more importantly may limit generalizability of approaches to populations that use different technologies. Hundreds of different smartphones and wearable devices house different combinations of sensors, CPU, GPU, and operating systems.

Physiological and behavioral signals can be generated from causes unrelated to neuropsychiatric illness, e.g. heart rate and locomotor activity patterns appear abnormal due to a temporary change in a patient's work schedule rather than a change in depressive symptomatology. Technological advances in battery life, sensor design, GPS, and integration with social media to capture context related to mood could address some of these issues.

Finally, integration between data captured via smartphones and wearables with the electronic medical record (EMR) may be important for providing researchers with a more comprehensive picture of patient health, but this not yet a ubiquitous capability. Some EMR systems can import data from devices used at home – such as blood pressure cuffs and glucose monitors – but few can seamlessly sync with smartphones and wearable devices, and attention has been focused on non-neuropsychiatric chronic conditions (Peeples et al. 2013; Kumar et al. 2016; Validic 2017). These data would not be clinically useful in its raw

form; rather, algorithms and visualizations are needed to transform information into insight that aids clinicians as they make decisions about patient management. To ensure healthcare providers can make use of passively monitored sensor data, smartphone and wearables-based sensing approaches should be technically and culturally inter-operable with the electronic medical record.

Large-scale aggregation of patient data could potentially benefit individuals and society, but raise issues around privacy and protection. Concerns over the ethical and societal impact of this field have grown as research funding agencies strongly encourage researchers to share collected data, and as the role of ubiquitous internet-connected mobile technology in our lives continues to expand. Patients face unique and dire ramifications of having their clinical information exposed due to the stigma of neuropsychiatric illness. For example, although methods used to identify patients with depression could be used for positive applications such as monitoring the status of patients in a rural area or enabling cohort discovery for therapeutic trials, similar methods could also be used less moral purposes, such as discrimination to reduce the risk of insurance payouts or future employer costs. Anonymizing health data before analysis has been proposed as one method to better ensure privacy and compliance with regulations (Emam et al. 2015).

Elhai and Frueh 2015 summarizes practical solutions for clinicians and researchers to secure electronic patient communication and records (Elhai et al. 2015). Furthermore, they review encrypted wireless networks, secure e-mail, encrypted messaging and videoconferencing, and privacy on social networks. Finally, we direct the interested reader to Horvitz and Mulligan 2015, which provides a thoughtful overview of these concerns in today's era of machine learning and large data; they note that informed discussions between technical experts, policy-makers, and the public will enable the design of programs and policies that balance privacy, fairness, and progress (Horvitz et al. 2015).

The US Food and Drug Administration (FDA) and the Federal Trade Commission (FTC) are the major governing agencies that evaluate safety and marketing claims of mobile health technologies, although neither currently provide comprehensive oversight of all apps and wearables. The approval of new devices and software requires evidence of safety and effectiveness, or the establishment of "substantial equivalence" to an already approved technology via the "510(k) pathway". Because thousands of new smartphone apps and several wearable devices with monitoring functionality are released every month, the FDA has chosen a tailored and risk-based approach. Specifically, "for many mobile apps that meet the regulatory definition of a device but pose minimal risk to patients and consumers, the FDA will exercise enforcement discretions and will not expect manufacturers to submit premarket review applications or to register and list their apps with the FDA" (U.S. Department of Health and Human Services Food and Drug Administration 2005). Notably, the FDA does not regulate the sale or general consumer use of smartphones or tablets, nor does their policy apply to mobile apps that function as an electronic health record or personal health record system (U.S. Department of Health and Human Services Food and Drug Administration 2015). Given the FDA's nascent yet rapidly evolving understanding of how smartphone, wearable devices, and software will be used for healthcare, it is important for clinicians, researchers, and software developers to consider potentially adverse outcomes

and subsequent liability risks that may occur downstream from their novel technologies (Armontrout et al. 2016). Such concerns are particularly important for technologies that can be used to inform and alter treatment recommendations such as pharmacotherapy dosage and clinical encounter scheduling.

7. Future potential

Currently, clinical trials have several limitations. Strict inclusion and exclusion criteria are employed to test interventions against a clean background, rather than a real-world scenario in which adherence to the intervention or data collection protocol can be more challenging. Data are collected from patients using long, paper-based questionnaires, journals, or web-based surveys. These tools are inconvenient and time-consuming to patients and do not reflect the context of their daily lives. Only 2% of the eligible population in the U.S. participate in clinical trials (U.S. Food & Drug Administration 2017). Those who do participate attend an average of 11 trial site visits over six months which can require traveling a significant distance. Finally, conducting trials for patients with serious neuropsychiatric illness can be especially challenging due to limited ability to adhere to study protocols.

Mobile and internet-connected technologies can help address some of these issues by enabling trials to be carried out at a participant's home or local physician's office – a "virtual" or "remote" trial – rather than at a central trial site (Seyfert-Margolis 2018). Virtual trials could also increase the rate of enrollment in exploratory or clinical studies (Savage 2015). For example, over eight months the MyHeart Counts app attracted over 48,000 people who consented to participate in a study of cardiovascular health; 40,000 people uploaded data including surveys on diet, well-being, risk perception, work-related and leisure-time physical activity, sleep, and cardiovascular health (McConnell et al. 2016). During the initial seven-day monitoring period, participants' motion was recorded via phone accelerometry. After one week, 4,990 people completed a six-minute walk test. Similarly, the mPower app, built using Apple's ResearchKit framework in a collaboration with the University of Rochester and Sage Bionetworks, aims to quantitatively assess symptoms of Parkinson's disease, and has been downloaded by 48,000 people with 9,520 subjects consenting to sharing their data (Bot et al. 2016). Novartis has worked with Science 37, a technology company that develops decentralized clinical trial technology and design, on virtual trials for cluster headache, acne and nonalcoholic steatohepatitis (NASH) (Novartis 2018). Recently, these two entities announced a strategic alliance to initiate up to 10 new decentralized and technology-driven remote clinical trials over the next three years. In addition to bolstering trial enrollment and retention, digital sensors could detect more subtle or nuanced effects of an intervention that could be missed by traditional outcome measures. The quantity and intrinsic speed of data gathering and processing afforded by sensors and software could also better enable adaptive trials, whereby investigators use accumulated data and modify or redesign the trial while the study is still ongoing (Chow 2014).

Digital sensor data is likely to complement rather than replace data obtained in current research trials such as blood biomarkers and imaging. The Emory Healthy Aging Study is an example of this multifaceted approach and will be the largest clinical research study ever

conducted in Atlanta, GA (Emory University 2016). The goal is to develop a midlife biomarker of Alzheimer's disease, since it is now well established that the disease begins about two decades prior to the onset of symptoms. Developing new ways to detect the disease in the asymptomatic phase is key for developing preventative treatments. To accomplish this goal, the Emory Healthy Aging Study first aims to recruit 100,000 individuals to participate in an online study to assess risk factors identified via health questionnaires, smartphones, and wearable devices. The second aim is to deeply phenotype a subpopulation of 3000 of these subjects every few years to assess risk factors by profiling genetics, cardiovascular physiology, blood and spinal fluid biomarkers, and brain and retinal imaging. Analyses of subjects' profiles, including their amyloid status, will facilitate discovery of new biomarkers with diagnostic and prognostic utility.

Mobile and wearable technologies have become dramatically cheaper over the past few decades and could help address the under-distribution of medications and personnel related to neuropsychiatric care in low-resource settings (Collins et al. 2011). Young males, ethnic minorities and people living in socioeconomically disadvantaged areas are more likely to experience "severe mental disorders including schizophrenia, bipolar affective disorder, and depression with psychotic symptoms such as hallucinations, delusions and cognitive disorganization" (Jongsma et al. 2018). Furthermore, even in a wealthy country such as the USA, ethnic minorities have significantly less access to care than do European Americans (Mcguire et al. 2008). Compounding this issue, the poorest countries spend the lowest percentages of their overall health budgets on mental health and have less relative availability of diagnostic encounters and interventions (Saxena et al. 2007). Telepsychiatry and teleneurology can extend the geographic reach of clinicians in regions with limited health resources, but this approach is still bottlenecked by the supply of trained professionals. To deliver interventions in a more scalable manner, smartphone and internet-based methods have been explored, including prerecorded video tutorials, self-help interventions, online communities or support groups, and guides to help patients navigate their healthcare system (Kazdin et al. 2013). Digital sensors could complement these approaches by enabling detection of early signs of illness relapse, medication adherence, or treatment efficacy. Although technology-based care delivery methods such as telemedicine are becoming increasingly available in health systems, passive monitoring has yet to become an established component of clinical workflow, especially in resource-poor regions. Many attempts at delivering affordable healthcare technologies into such environments have not achieved the intended levels of impact due to a focus only on cost or simplicity. Attention to sustainable business practices, local cultural dynamics, and integration with existing resource and workflow may enable the potential of these technologies to educate and assist patients and providers; Clifford 2016 provides a thorough review of these considerations and proposes structural ecosystem changes to help achieve empowerment.

Although much work has focused on demonstrating feasibility of passive sensing, the gap between data capture and meaningful improvements in patient outcomes has yet to be closed (Patel et al. 2015). A growing body of literature has shown that smartphones not only can monitor patients but can also send information to patients in a way that affects clinical outcomes. SMS can increase adherence to antiretroviral therapy and smoking cessation (Free et al. 2013), and smartphone delivery of cognitive behavioral therapy can reduce anxiety,

depression, stress, and substance use (Ehrenreich et al. 2017). Recently, Freeman et al. conducted the largest RCT of a psychological intervention for a mental health problem (Freeman et al. 2017). 3,755 students with insomnia from 26 UK universities were enrolled in the trial, with 1,891 receiving digital CBT for insomnia ("Sleepio"), and 1,864 receiving standard practice treatment. Digital CBT was accessible via web browser, and sleep diaries and relaxation audio was accessible via smartphone. At ten weeks, Sleepio significantly reduced insomnia, paranoia, and hallucinations compared to the usual practice. However, no large RCT focused on neuropsychiatric illness has reported a positive impact of passive monitoring on outcomes. A recent difference-in-differences random effects meta-analysis of RCTs of remote patient monitoring did not find statistically significant impacts on any of six outcomes including body mass index, weight, waist circumference, body fat percentage, systolic blood pressure, and diastolic blood pressure (Noah et al. 2017). Interventions based on health behavior models and personalized coaching – relevant to neuropsychiatric care – were most successful.

Remote sensing of patients is an important but early step in the iterative process by which data is used to improve patient management, e.g. revise parameters of CBT or other psychotherapy, adjust doses or selection of pharmacological agents, or modify recommended lifestyle and behavior changes. In turn, the effect of these interventions can be measured in near real-time. Thus, digital sensors could be an integral component of how neuropsychiatric care is delivered in the future: a feedback loop starting with data-driven insight about pathophysiology and/or treatment, that in turn optimizes therapy, and ultimately improves patient outcomes.

8. Conclusion

In closing, many studies have explored the use of smartphones, wearable accelerometers, Holter devices, and multimodal sensors for monitoring of patient physiology, psychology, and behavior. These technologies continue to decrease in cost and permeate other facets of daily life. Future work in this field must address numerous technical, cultural, and ethical considerations of data collection and analysis. Notably, digital sensor data needs to seamlessly integrate with existing clinical workflow, electronic medical record systems, and relevant personal information such as employment, social support systems, and financial access to healthcare. Together these data can paint a richer and more cohesive picture of a patient's health and enable more personalized, efficacious, and contextually appropriate care. Furthermore, the abundance of small-scale feasibility studies lack the sample size or properly randomized control groups to generate evidence necessary to translate technology-driven monitoring approaches into standard practice of care. More than a sufficient number of these exploratory studies have been published, and proof-of-concept is now well-established. It is time for the field to adopt a new standard of large, prospective, and multi-site RCTs that utilize machine learning and more sophisticated methods than univariate statistical significance testing to determine if monitoring with digital sensors can complement or even replace existing methods of diagnosing, monitoring, and treating patients with serious neuropsychiatric illness. Finally, digital sensor technologies will continue to rapidly advance due to tremendous non-healthcare enterprise and consumer impact, creating new opportunities to non-invasively measure other signals with clinical

relevance such as blood pressure, electroencephalography, skin conductance, and perhaps even biomolecular markers. The devices we carry in our pockets and wear on our wrists or elsewhere will transform how clinical trials are conducted, improve our understanding of day-to-day variability of neurological and mental illness, and most importantly facilitate tailored, dynamically responsive, and time-sensitive interventions.

Acknowledgments

Appendix A1

Table A1

Abbreviations used throughout review

Abbreviation	Definition
AD	Alzheimer's disease
ADHD	Attention deficit hyperactivity disorder
ADL	Activities of daily living
ANS	Autonomic nervous system
AUC	Area under the receiver operating characteristic curve
BD	Bipolar disorder
BLT	Bright light therapy
CDC	Centers for Disease Control and Prevention
DALY	Daily adjusted life year
ECG	Electrocardiogram
EMA	Ecological momentary assessment
GPS	Global positioning system
GSR	Galvanic skin response
HF	High frequency
HRV	Heart rate variability
IS	Interdaily stability
IV	Intradaily variability
L5	Mean activity level during the least active five hours
LF	Low frequency
M10	Mean activity level during the most active ten hours
MCI	Mild cognitive impairment
MDD	Major depressive disorder
PPG	Photoplethysmography
PTSD	Post traumatic stress disorder
RA	Relative amplitude
RMSSD	Root mean square of the successive differences
rCBF	Regional cerebral blood flow

Abbreviation	Definition
SAD	Seasonal affective disorder
SCID	Structured clinical interview for the DSM-IV
SDNN	Standard deviation of average normal-to-normal intervals
SMS	Short message service
VLF	Very low frequency power

Appendix A2

Relevant questionnaires, surveys, and scales

The self-reporting of symptoms is an extremely useful gauge of patient progress or acuity. Although such surveys have been traditionally administered via paper, or more recently via web pages, it is increasingly common to capture such data through an approach called Ecological Momentary Assessment (EMA), whereby questions can be delivered to the subject via smartphone in response to triggers, a certain time, or a pattern of interest in gathered data. The questions can be repeatedly administered if the user does not answer. While there is little evidence so far as to the effect this has on such scales, the flexibility this offers provides a new avenue for research into such systems, whereby timing of the response, and even corrections during the process could be analyzed to extract further information about the state of a patient. In this section we review a variety of the most relevant surveys for neuropsychiatric EMA and provide the evidence base for their traditional use.

The Perceived Stress Scale (PSS) was developed to measure psychological stress, defined as "the extent to which persons perceive that their life demands exceed their ability to cope" (Cohen et al. 1983). The PSS predicts both objective biological markers of aging (Epel et al. 2004), cortisol levels (Malarkey et al. 1995), immune markers (Maes et al. 1999), depression (Carpenter et al. 2004), and increased risk for disease among persons with higher perceived stress levels.

The Hamilton Rating Scale for Depression (HRSD, HAMD, or HAM-D) is a multiple item questionnaire used to quantify the results of an interview assessment of symptoms in an adult patient diagnosed with depression (Hamilton 1960). Severity of depression is assessed by probing mood, feelings of guilt, suicide ideation, insomnia, agitation or retardation, anxiety, weight loss, and somatic symptoms among 17 to 29 dimensions (depending on version; often referred to as the HAMD-17 or HAMD-29 respectively) with a score on a 3 or 5 point scale. A score of 0-7 is considered to be normal. Scores of 20 or higher indicate moderate, severe, or very severe depression, and are usually required for entry into a clinical trial. However, the HRSD has been criticized as a test because it places more emphasis on insomnia than on suicide ideas and gestures (Bagby et al. 2017). An antidepressant may show statistical efficacy even when thoughts of suicide increase but sleep is improved. Alternatively, even if a medication effectively reduces depressive symptoms, if sexual and gastrointestinal symptoms worsen as a side effect, efficacy can be underestimated. Results of a large meta-analysis suggest that HRSD achieves good overall levels of internal

consistency, inter-rater and test–retest reliability, but some HRSD items (e.g., "loss of insight") are not sufficiently reliable (Trajkovi et al. 2011).

The Quick Inventory of Depressive Symptomatology (QIDS-SR16) is a shortened 16-item version of the 30-item Inventory of Depressive Symptomatology (IDS), a structured interview that was constructed by selecting only items that assessed DSM-IV diagnostic criterion items for MDD (Rush et al. 2000). The research group that developed the IDS obtained feedback/critique from more than a dozen, largely US, clinical researchers who were experts in depression. The nine domains of the QIDS-SR16 comprise sad mood, concentration, self-criticism, suicidal ideation, interest, energy/fatigue, sleep disturbance (initial, middle, and late insomnia or hypersomnia), decrease/increase in appetite/weight, and psychomotor agitation/retardation. The total score ranges from 0 to 27. QIDS-SR16 has high internal consistency, as well as high correlation with the IDS and the HAMD (Rush et al. 2003).

The Primary Care Evaluation of Mental Disorders (PRIME-MD) Patient Health Questionnaire (PHQ) was designed by the PHQ Primary Care Study Group to be a fully self-administered survey; the original survey it was based upon was clinician-administered (Spitzer et al. 2012). There is an optional fourth page that includes questions about menstruation, pregnancy and child-birth, and recent psychosocial stressors. The original PHQ assessed 18 current mental disorders. By grouping several mood, anxiety, and somatoform categories together, the PHQ greatly simplifies the differential diagnosis by assessing only eight disorders: MDD, panic disorder, other anxiety disorder, bulimia nervosa, other depressive disorder, probable alcohol abuse or dependence, and somatoform and binge eating disorders. Patients indicate for each of the 9 depressive symptoms whether, during the previous 2 weeks, the symptom has bothered them "not at all," "several days," "more than half the days," or "nearly every day.". Patients also indicate for each of the 13 physical symptoms whether, during the previous month, they have been "not bothered," "bothered a little," or "bothered a lot" by the symptom. The PHQ Primary Care Study Group found agreement between PHQ diagnoses and those of independent mental health professionals; for the diagnosis of any 1 or more PHQ disorder, $\kappa = 0.65$; overall accuracy, 85%; sensitivity, 75%; specificity, 90%, similar to the original PRIME-MD questionnaire. Furthermore, in addition to making criteria-based diagnoses of depressive disorders, the PHQ-9 has also been show to be a reliable and valid measure of depression severity (Kroenke et al. 2001). A slightly shorter eight-question version of this survey, the PHQ-8, is also sometimes used.

The Center for Epidemiological Studies Depression (CESD) scale is a short self-report scale comprised of 20 questions that ask how often over the past week a person experienced symptoms associated with depression, such as restless sleep, poor appetite, or feeling lonely (Radloff 1977). Each item is scored 0 to 3: 0 = Rarely or None of the Time, 1 = Some or Little of the Time, 2 = Moderately or Much of the time, 3 = Most or Almost All the Time. Total scores range from 0 to 60, with higher scores indicating greater depressive symptoms. Cutoff scores identify individuals at risk for clinical depression with good sensitivity and specificity, and high internal consistency (Lewinsohn et al. 1997).

The Beck Depression Inventory (BDI) consists of 21 multiple-choice questions that ask how the subject has been feeling in the last week, and is a proxy for a structured clinical interview (Beck et al. 1961). Questions inquire about symptoms of depression such as hopelessness and irritability, cognitions such as guilt or feelings of being punished, physical symptoms such as fatigue, weight loss, and lack of interest in sex. Each question has a set of at least four possible responses, ranging in intensity. A value of 0 to 3 is assigned for each answer, and the values are summed to calculate a total sum up to 63. A higher total score indicates more severe depressive symptoms. The BDI is one of the most widely used psychometric tests for measuring the severity of depression; its successor is the BDI-II which is now more common. The BDI was revised 1996 to the BDI-II in response to the American Psychiatric Association's publication of the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, which changed many of the diagnostic criteria for MDD (Beck et al. 1996). The BDI-II is used to evaluate how the subject has been feeling over the past two weeks instead of one week, in order to be consistent with the DSM-IV time period for the assessment of MDD.

The Young Mania Rating Scale (YMRS) is an eleven-item clinician-administered scale to rate manic symptoms. This score correlated with the number of days of subsequent stay in hospital, and significantly differed in patients before versus after two weeks of treatment (Young et al. 1978). A parent report version of the YMRS (P-YMRS) was assessed in a cohort of 117 youths age 5-17 (Gracious et al. 2002). The P-YMRS demonstrated acceptable internal consistency. Logistic regressions discriminated bipolar mood disorder versus unipolar disorder, versus disruptive behavior disorder, and versus any other diagnosis. Classification rates exceeded 78%, and receiver operating characteristics analyses showed areas under the curve greater than 0.82.

The Altman Self-Rating Mania scale (ASRM) is a self-administered survey that was originally evaluated on a cohort of 22 schizophrenic, 13 schizoaffective, 36 depressed, and 34 manic patients (Altman et al. 1997). The Clinician Administered Rating Scale for Mania (CARS-M) and Mania Rating Scale (MRS) were completed at the same time to measure concurrent validity. Principal component analysis of ASRM items revealed three factors: mania, psychotic symptoms, and irritability. Baseline mania subscale scores were significantly higher for manic patients compared to all other diagnostic groups. Posttreatment scores were significantly decreased in manic patients for all three subscales. ASRM mania subscale scores significantly correlated with MRS total scores ($r = 0.72$) and CARS-M mania subscale scores ($r = 0.77$). Test-retest reliability for the ASRM was significant for all three subscales. Mania subscale scores of greater than 5 on the ASRM resulted in sensitivity of 85.5% and a specificity of 87.3%.

The General Behavior Inventory (GBI) is a 73-question self-administered survey that evaluates various aspects of mood and is designed to identify the presence and severity of manic and depressive moods in adults (Depue et al. 1981). It consists of two scales to assess depressive symptoms (46 items) and hypomanic / biphasic (mixed) symptoms (28 items) (Youngstrom et al. 2008). GBI items use a Likert scale from 0-3: 0 (never or hardly ever present), 1 (sometimes present), 2 (often present), and 3 (very often or almost constantly present). The GBI has high internal consistency and retest reliability because of its large

number of items. Retest reliability also is good over a week or two week period, although the required reading level and length make it challenging for some people to complete.

The Social Rhythm Metric (SRM) quantifies an individual's social zeitgebers (time givers, or circadian rhythm entrainment cues) (Monk et al. 1990). Social life may provide important social cues that entrain circadian rhythms, including sleep habits, eating times, and occupational routines. Disturbance of these social cues could result in dis-entrainment of circadian rhythms, which may increase the risk of developing mood disorders or other mental illnesses. The SRM score is determined from the timing of 15 specific and 2 built-in activities that constitute an individual's social rhythm. If the timing of an activity that occurs at least three times a week is within 45 minutes of the typical time, it is considered part of one's daily routine. The total number of these activities is divided by the total number of activities occurring at least three times a week. The result is the SRM score. A higher SRM score was found to relate to subjective better sleep, higher morning alertness and a deeper nocturnal temperature trough, whereas lower SRM-scores correlated with higher reports of depressive symptoms (Monk et al. 1994).

The Pittsburgh Sleep Quality Index (PSQI) is a self-administered survey which assesses sleep quality and disturbances over a one month time interval (Buysse et al. 1989). Nineteen individual items generate seven component scores: subjective sleep quality, sleep latency, sleep duration, habitual sleep efficiency, sleep disturbances, use of sleeping medication, and daytime dysfunction. The sum of scores for these seven components yields one global score.

The Positive and Negative Syndrome Scale (PANSS) is a 30-question clinician-administered survey that measures symptom severity of schizophrenia and has been widely used in the study of antipsychotic therapy (Kay et al. 1987). Seven questions assess positive symptoms, which refer to an excess or distortion of normal functions, e.g., hallucinations and delusions. Seven questions assess negative symptoms, which represent a decrease or loss of normal function, e.g. blunted affect and social withdrawal. 16 questions assess general psychopathology, e.g. feelings of guilt and poor attention. Each answer is rated 1 to 7 based on the interview as well as reports of family members or healthcare providers. The overall PANSS score thus ranges from 30 to 210. Kay's original publication reported a mean score of 77 for patients with schizophrenia.

The Calgary Depression Scale for Schizophrenia (CDSS) is a nine-item clinician-administered survey, in which each item has a four-point Likert scale measure (Addington et al. 1990). It was designed to assess depression specifically in psychotic populations, for whom previous depression instruments were not designed. Internal consistency is high, and significant and strong correlations have been found between scores on the CDSS, BDI, and HRSD (Addington et al. 1992; Addington et al. 1993). The CDSS depression score is obtained by adding each of the item scores. A score above 6 has an 82% specificity and 85% sensitivity for detecting a major depressive episode.

The Unified Parkinson's Disease Rating Scale (UPDRS) is a clinician-administered interview and exam that is used to describe the severity of Parkinson's Disease (Fahn et al. 1987). It is made up of the 1) Mentation, Behavior, and Mood, 2) ADL, and 3) Motor

sections. Some sections require multiple grades assigned to each extremity. The score ranges from 0 to 199; 0 represents no disability, and 199 represents total disability. Strengths of the UPDRS include its wide utilization, its application across the clinical spectrum of PD, its nearly comprehensive coverage of motor symptoms, and its clinimetric properties, including reliability and validity. Weaknesses include several ambiguities in the written text, inadequate instructions for raters, some metric flaws, and the absence of screening questions on several important non-motor aspects of PD (Goetz 2003). The motor section of the UPDRS (UPDRS-III) is often used in lieu of the entire UPDRS for patients with PD, and the exam is ideally performed by a movement disorder specialist. In 2007 the Movement Disorder Society (MDS) revised the UPDRS which originally placed nonmotor elements in PD throughout the subscales, with mental features captured in Part I, pain in Part II, and sleep disorders and dysautonomia in Part IV (Goetz et al. 2007). The scale was reorganized so that Part I of the MDS-UPDRS is now titled "Nonmotor Experiences of Daily Living" and encompasses questions requiring medical expertise to answer (cognitive impairment, hallucinations, depressed mood, anxious mood, apathy, and dopamine dysregulation) as well as simpler questions that were considered better suited for a patient or caregiver questionnaire (sleep, staying awake, pain and abnormal sensory sensations, urinary function, constipation, lightheadedness on standing, and fatigue). Part II was retitled to "Motor Experiences of Daily Living", Part III remains "Motor Examination" to be completed by the rater, and Part IV was restricted to "Motor Complications" which include dyskinesias and motor fluctuations. This revised MDS-UPDRS is now commonly used in PD research.

The Hoehn and Yahr (HY) scale was originally designed to be a descriptive, clinician-administered structured interview and staging scale that estimates clinical function in PD, combining functional disability and objective signs of impairment (Hoehn et al. 1967). Strengths of the HY scale include its wide utilization and acceptance. Higher stages correlate with dopaminergic loss as confirmed via neuroimaging studies, and the HY scale has been shown to highly correlate with some standardized scales of motor impairment, disability, and quality of life (Goetz et al. 2004). Weaknesses include the scale's mixing of impairment and disability. Because the HY scale is weighted heavily toward postural instability in determining disease severity, it does not capture impairments or disability from other motor features of PD, and gives no information on nonmotor problems which are also features of the illness that contribute to decreased quality of life. The UPDRS has largely supplanted the HY scale in clinical and research use.

The Short Form-36 (SF-36) is the most widely used health-related quality-of-life measure in research to date, and can be either self-administered or administered by a trained interviewer over the phone or in person (Ware Jr et al. 1992). The SF-36 yields eight scale scores and two summary scores: a physical component summary (PCS), and mental component summary (MCS). The physical and mental components were designed to be uncorrelated. The eight scale scores represent physical functioning, bodily pain, role limitations due to physical health problems, role limitations due to personal or emotional problems, general mental health, social functioning, energy/fatigue or vitality, and general health perceptions. A higher score represents better health. The PCS and MCS scores are calculated by z-scoring each of the eight scores across the general U.S. population, then multiplying by the corresponding factor scoring coefficient for each scale (Taft et al. 2001).

The Instrumental Activities of Daily Living (IADL) scale assesses independent living skills, identifies how a person is functioning at the present time, and determines improvement or deterioration over time (Lawton et al. 1969). In the original study, the survey was administered by a social worker who gathered information from the subjects, family members, employees, etc. Eight domains of function are measured: ability to use the telephone, shopping, food preparation, housekeeping, laundry, mode of transportation, responsibility for own medications, and ability to handle finances. The IADL Scale is intended to be used among older adults, and may be used in community, clinic, or hospital settings, but is not useful for institutionalized older adults. Although the IADL Scale is easy to administer and focuses on practical functionality related to daily living, it relies on self-report or surrogate report rather than a demonstration of the functional task.

The State Trait Anxiety Inventory (STAI) is a 40-item self-administered survey designed to measure anxiety at two ends of the "affect curve", e.g. feelings of anxiety due to a stressful state or situation, versus enduring personality traits (Spielberger et al. 1983). Each item has a four-point Likert scale measure. Overall scores thus range from 20 to 80, with higher scores suggesting more severe anxiety.

The Generalized Anxiety Disorder (GAD-7) is a 7-item self-administered survey used to identify GAD (Spitzer et al. 2006). It was constructed from 965 adult primary care patients who completed a questionnaire and telephone interview with a mental health professional within a week, and achieved a sensitivity of 89% and specificity of 82% in assessing generalized anxiety disorder, with good agreement between self-report and interviewer-administered versions of the scale.

The Apathy Evaluation Scale (AES) is used to evaluate apathy – the lack of will to act and the inability to care about the consequences – in a patient based on interview of a person familiar with the patient (Marin 1996). The scale consists of 18 questions that each use a four point Likert scale measure ranging from 0 to 3. Overall scores thus range from 0 to 54; the higher the score the greater the level of apathy.

The Brief Psychiatric Rating Scale (BPRS) is used for measuring general psychiatric symptoms such as depression, anxiety, hallucinations and unusual behavior (Overall et al. 1962). During a structured clinical interview, 18-24 symptoms are scored, and each symptom is rated 1-7 where 1 indicates absence of symptomatology or concern, and 7 indicates extreme severity. The BPRS is one of the oldest and most widely used scales to measure psychotic symptoms.

Table A2

Questionnaires, surveys, and scales

Reference	Survey (acronym)	Indication
Cohen et al. 1983	Perceived Stress Scale (PSS)	Stress
Hamilton 1960	Hamilton Rating Scale for Depression (HRSD or HAMD)	Depression

Reference	Survey (acronym)	Indication
Rush et al. 2000	16-item Quick Inventory of Depressive Symptomatology (QIDS-SR16)	Depression
Spitzer et al. 1999	Patient Health Questionnaire (PHQ)	Depressive disorders
Radloff 1977	Center for Epidemiological Studies Depression (CESD)	Depression
Beck et al. 1961	Beck Depression Inventory (BDI)	Depression
Young et al. 1978	Young Mania Rating Scale (YMRS)	Mania
Altman et al. 1997	Altman Self-Rating Mania scale (ASRM)	Mania
Depue et al. 1981	General Behavior Inventory (GBI)	Mania and depression
Monk et al. 1990	Social Rhythm Metric (SRM)	Circadian entrainment
Buysse et al. 1989	Pittsburgh Sleep Quality Index (PSQI)	Sleep
Kay et al. 1987	Positive and Negative Syndrome Scale (PANSS)	Schizophrenia
Addington et al. 1990	Calgary Depression Scale for Schizophrenia (CDSS)	Depression in schizophrenia
Fahn & Elton 1987	Unified Parkinson's Disease Rating Scale (UPDRS)	Parkinson's disease
Hoehn & Yahr 1967	Hoehn and Yahr (HY) scale	Parkinson's disease
Ware & Sherbourne 1992	Short Form-36 (SF-36)	Quality of life
Lawton & Brody 1969	Lawton Instrumental Activities of Daily Living (IADL)	–
Spielberger et al. 1983	State Trait Anxiety Inventory (STAI)	Anxiety
Spitzer et al. 2006	7-item Generalized Anxiety Disorder (GAD-7) scale	Anxiety
Marin et al. 1996	Apathy Evaluation Scale (AES)	Apathy
Overall & Gorham 1962	Brief Psychiatric Rating Scale (BPRS)	General psychiatric symptoms

Table A3

Studies of smartphones and wearables for monitoring neuropsychiatric illness

Reference	Key aim	Population	Sensors	Design
Abdullah et al. 2016	Estimate social rhythms (assessed via SRM questionnaires) using smartphone data	Seven subjects with BD	Smartphones recorded GPS data, accelerometry, microphone audio, and social communication	Offline retrospective
Aguilera et al. 2015	Assess relationship between daily / weekly mood scores and PHQ-9 scores	33 subjects	Smartphone administered PHQ-9 surveys	Offline retrospective
Albert et al. 2017	Distinguish subjects with PD from controls using accelerometry of hand tremor	Eight subjects with PD and 18 controls	Smartphones recorded accelerometry of hand tremor during motor tasks	Offline retrospective
AlHanai et al. 2017	Classify subject mood while reading happy or sad stories using wearable data	Ten healthy subjects	Audio was recorded using Apple iPhones. Samsung Simband smartwatches recorded PPG, ECG, accelerometry, skin impedance, galvanic skin response, and skin temperature	Online real-time

Reference	Key aim	Population	Sensors	Design
Apiquian et al. 2017	Assess motor activity and sleep time before and after antipsychotic treatment	20 subjects with schizophrenia and 20 controls	Wrist-worn devices recorded accelerometry	Offline retrospective
Barnett et al. 2018	Predict clinical relapse from behavioral anomalies in two-week window prior to event	17 subjects with schizophrenia	Smartphones recorded mobility, social activity, and questionnaires	Offline prospective
Beiwinkel et al. 2016	Depressive and manic symptoms (assessed via HAMD and YMRS questionnaires administered every three weeks) were classified using smartphone data	13 subjects with BD	Smartphones recorded GPS, accelerometery, and cell tower data; mood states were assessed via a self-reported two-item questionnaire	Offline retrospective
Ben-Zeev et al. 2015	Correlate smartphone features with daily stress ratings, PHQ-9, PSS, and Revised UCLA Loneliness Scale scores	47 healthy subjects	Smartphones recorded GPS, accelerometery, sleep duration, and time proximal to human speech	Offline retrospective
Berle et al. 2017	Assess motor activity and rest-activity characteristics	46 subjects with schizophrenia and 32 controls	Wrist-worn devices recorded actigraphy	Offline retrospective
Bullock et al. 2014	Assess rest-activity metrics in BD patients with low and high trait vulnerability (assessed via the GBI questionnaire)	72 subjects with BD	Wrist-worn devices recorded accelerometry	Offline retrospective
Burns et al. 2011	Correlate EMA survey scores with smartphone features	Eight subjects with MDD	Smartphones recorded GPS, accelerometry, ambient light, and recent calls	Offline retrospective
Canzian et al. 2015	Correlate and predict PHQ score deviations with smartphone features	28 healthy subjects	Smartphones recorded GPS and accelerometry	Offline prospective
Capecci et al. 2016	Identify freezing of gait events using accelerometry	20 subjects with PD	Smartphones recorded accelerometry while subjects walked and were video recorded	Offline retrospective
Cella et al. 2017	Assess autonomic dysfunction in schizophrenia using wearable device data	30 subjects with schizophrenia and 25 controls	Empatica E4 devices recorded skin conductance, HRV, and accelerometry	Offline retrospective
Ellis et al. 2015	Compare outcome measures of gait and gait variability in subjects with PD versus controls	12 subjects with PD and 12 controls	Steps were captured via a smartphone, heel-mounted sensors, and a sensor mat	Offline retrospective
Kamdar et al. 2016	Estimate variance of emotional state	13 healthy subjects	Samsung Gear S smartwatches recorded accelerometry,	Offline retrospective

Reference	Key aim	Population	Sensors	Design
	from wearable data via random forest		ambient light, heart rate; web app administered mood surveys	
Moore et al. 2012	Forecast mood time series using previous week's self-rated mood data via exponential smoothing and Gaussian process regression	100 subjects with BD	Mood surveys recorded via SMS	Offline prospective
Faedda et al. 2016	Distinguish BD from ADHD using wearables data	48 subjects with BD, 65 subjects with ADHD, and 42 controls	Belt-worn devices recorded accelerometry for five minutes	Offline retrospective
Faurholt-Jepsen et al. 2015	Correlate smartphone data with depressive and manic symptoms via HDRS-17 and YMRS scores assessed monthly	61 subjects with BD	Smartphones recorded speech duration, social activity, and accelerometry	Offline retrospective
Maria et al. 2016	Classify depressive and manic states (via HDRS-17 and YMRS scores) using smartphone data and voice features	28 subjects with BD	Smartphones recorded voice features (pitch, duration, etc.), speech duration, social activity, and accelerometry	Offline retrospective
Fasmer et al. 2015	Fit resting and active periods to power law distributions and assess differences in MDD	47 subjects with MDD and 29 controls	Wrist-worn devices recorded accelerometry	Offline retrospective
Griffiths et al. 2012	Assess features of dyskinesia and akinesia from wearable data, and identify improvements in UPDRS scores after medication	34 subjects with PD and 10 controls	Wrist-worn devices recorded accelerometry	Offline retrospective
Grünerbl et al. 2015	Depressive and manic symptoms (assessed via HAMD and YMRS questionnaires administered every three weeks) were classified using smartphone data	Ten subjects with BD	Smartphones recorded GPS, accelerometry, number and length of phone calls, and speech and voice features	Offline retrospective
Hauge et al. 2011	Assess motor activity and rest-activity characteristics	24 subjects with schizophrenia, 25 subjects with depression, and 32 controls	Wrist-worn devices recorded actigraphy	Offline retrospective
Kassavetis et al. 2016	Correlate UPDRS scores with smartphone data	14 subjects with PD	Smartphones recorded accelerometry while subjects performed motor tasks	Offline retrospective
Kheirkhahan et al. 2016	Correlate impaired mobility from wearable data	1,135 subjects	Hip-worn devices recorded accelometry	Offline retrospective

Reference	Key aim	Population	Sensors	Design
Kim et al. 2015	Classify freezing episodes from normal walking using accelerometry	15 subjects with PD	Smartphones recorded accelerometry while subjects walked and were video recorded	Offline retrospective
Kostikis et al. 2014	Correlate accelerometry features with UPDRS hand tremor scores	23 subjects with PD	Smartphones recorded accelerometry of hand tremor during motor tasks	Offline retrospective
Kostikis et al. 2015	Distinguish subjects with PD from controls using accelerometry of hand tremor	25 subjects and 20 controls	Smartphones recorded accelerometry of hand tremor during motor tasks	Offline retrospective
Krane-gartiser et al. 2014	Assess mean activity, variance, symbolic dynamics, and power spectral features	18 subjects with mania and 12 subjects with BD	Wrist-worn devices recorded accelerometry	Offline retrospective
Kuhlmei et al. 2013	Associate activity with apathy and depression (assessed via AES and BDI questionnaires)	32 subjects with dementia, 21 subjects with MCI, and 23 controls	Wrist-worn devices recorded accelerometry during motor tasks	Offline retrospective
Lee et al. 2015	Compare RR peak detection, HRV measures, and stress detection from wearable versus Holter monitor	17 subjects	Custom ECG patch was developed to record cardiac activity	Offline retrospective
Lee et al. 2016	Correlate UPDRS scores with smartphone data	103 subjects with PD	Smartphones recorded hand dexterity via timed tapping test, rapid alternating movements, tremor tracker via tracing between two parallel lines, and a cognitive interference test	Offline retrospective
Martin et al. 2006	Assess time in bed, sleep consistency, daytime sleeping, and circadian rhythm regularity	28 subjects with schizophrenia and 28 controls	Wrist-worn devices recorded accelerometry and light exposure	Offline retrospective
Nakamura et al. 2007	Fit resting and active periods to power law distributions and assess differences in MDD	14 subjects with MDD and 11 controls	Wrist-worn devices recorded accelerometry	Offline retrospective
Nero et al. 2015	Define accelerometer cut points for different walking speeds in adults with PD	30 subjects with PD	Waist-worn devices recorded accelerometry	Offline retrospective
Niwa et al. 2011	Assess if medication status, MMSE scores, activity, and HRV features differed by disease severity (assessed via	27 subjects with PD and 30 controls	Wrist-worn devices recorded accelerometry and Holter monitors recorded ambulatory ECG	Offline retrospective

Reference	Key aim	Population	Sensors	Design
	UPDRS scores) or disease duration			
O'Brien et al. 2016	Assess relationship between quality of life, ADLs, learning, and depression (assessed via SF-36 and IADLS questionnaires) and smartphone data	29 subjects with MDD and 30 controls	Wrist-worn devices recorded accelerometry. Quality of life, ADLs, learning, and depression were assessed via SF-36 and IADLS questionnaires	Offline retrospective
Osipov et al. 2015	Classify schizophrenic subjects from controls using rest-activity characteristics and HRV features	16 subjects with schizophrenia and 19 controls	Adhesive patches recorded locomotor activity and ECG	Offline retrospective
Palmius et al. 2017	Estimate depressive symptoms (assessed via QIDS-SR16 questionnaires administered weekly) and detect depression using smartphone data	22 subjects with BD and 14 controls	Smartphones recorded GPS data	Offline retrospective
Pan et al. 2015	Correlate accelerometry features with UPDRS scores, and use features to detect hand resting tremor and gait difficulty	40 subjects with PD	Smartphones recorded accelerometry of hand tremor and gait during motor and walking tasks	Offline retrospective
Patel et al. 2009	Estimate UPDRS scores using wearable data	12 subjects with PD	Arm and leg-worn devices recorded accelerometry	Offline retrospective
Place et al. 2017	Estimate depression and PTSD symptoms (assessed via SCID questionnaires) using smartphone data	73 subjects with at least one symptom of PTSD or depression	Smartphones recorded GPS, accelerometry, calls and SMS activity, device use, and voice audio	Offline retrospective
Reinertsen et al. 2017a	Classify patients with PTSD using time-domain, frequency-domain, and complexity features from RR interval time series	23 subjects with PTSD and 25 controls	A Holter monitor recorded RR intervals for 24 hours	Offline retrospective
Reinertsen et al. 2017b	Classify schizophrenic subjects from controls using rest-activity characteristics and HRV features, and evaluate relationship between number of days of data and classifier accuracy	16 subjects with schizophrenia and 19 controls	Adhesive patches recorded locomotor activity and ECG	Offline retrospective

Reference	Key aim	Population	Sensors	Design
Roh et al. 2014	Compare RR peak detection, signal-to-noise, and HRV measures from wearable versus Holter monitor	12-41 subjects (varied by test)	Custom ECG patch was developed to record cardiac activity	Offline retrospective
Roy et al. 2011	Classify tremor and dyskinesia from wearable data	11 subjects with PD	Arm and leg-worn devices recorded accelerometry	Offline retrospective
Saeb et al. 2015	Classify low from high PHQ-9 scores using smartphone features	28 healthy subjects	Smartphones recorded GPS and phone usage	Offline retrospective
Saeb et al. 2016a	Correlate PHQ-9 scores with smartphone features from weekend vs. weekday data	48 healthy subjects	Smartphones recorded GPS and phone usage	Offline retrospective
Sano et al. 2012	Fit resting and active periods to power law distributions and assess differences in schizophrenia	19 subjects with schizophrenia and 11 controls	Wrist-worn devices recorded accelerometry	Offline retrospective
Sano et al. 2013	Distinguish stressed from non-stressed states using wearable data	18 subjects	Wrist-worn devices recorded accelerometry and skin conductance. Smartphones recorded call and SMS activity. Surveys assessed stress, mood, sleep, tiredness, general health, alcohol or caffeine intake, and electronics usage.	Offline retrospective
Sano et al. 2015	Estimate PSQI, PSS, and MCS questionnaire scores from wearable data	66 subjects	Wrist-worn devices recorded accelerometry and skin conductance. Smartphones recorded call and SMS activity. Sleep, stress, and mental health were assessed via PSQI, PSS, and MCS questionnaires respectively	Offline retrospective
Shin et al. 2016	Correlate symptom severity (assessed via the PANSS questionnaire) with activity levels	61 subjects with schizophrenia	Wrist-worn devices recorded accelerometry	Offline retrospective
Stamatakis et al. 2013	Classify UPDRS score categories from wearable data	36 subjects with PD and 10 controls	Finger-worn sensors recorded accelerometry during a tapping test	Offline retrospective
Tung et al. 2014	Compare area, perimeter, and mean distance from home in subjects with AD versus controls using smartphone data	19 subjects with AD and 33 controls	Smartphones recorded GPS	Offline retrospective
Walther et al. 2009b	Assess if motor symptoms (assessed via PANSS questionnaires) correlate with wearables data	55 subjects with schizophrenia	Wrist-worn devices recorded actigraphy	Offline retrospective

Reference	Key aim	Population	Sensors	Design
Walther et al. 2009a	Assess if activity differs by schizophrenia subtype	60 subjects with schizophrenia	Wrist-worn devices recorded actigraphy	Offline retrospective
Wang et al. 2014	Correlate smartphone data with PHQ-9, PSS, flourishing scale, and UCLA loneliness scale scores	48 healthy subjects	Smartphones recorded accelerometry, conversations, sleep, and location	Offline retrospective
Wang et al. 2016	Determine associations between EMA survey scores and smartphone data via generalized estimating equations	21 subjects with schizophrenia	Smartphones recorded accelerometry, voice audio, light sensor readings, GPS data, and application usage	Offline retrospective
Weenk et al. 2017	Evaluate association between changes in HRV measures and stress in surgeons	20 subjects	Adhesive patch measured single-lead ECG, respiratory rate, skin temperature, body posture, activity, and steps	Offline retrospective
Wichniak et al. 2011	Measure association between activity levels and mental status (measured via PANSS and CDSS questionnaires)	73 subjects with schizophrenia and 36 controls	Wrist-worn devices recorded accelerometry	Offline retrospective
Winkler et al. 2005	Assess if light therapy can improve sleep efficiency and stability in people with seasonal affective disorder (SAD)	17 subjects with SAD and 17 controls	Wrist actigraphy was recorded from which sleep-wake amplitude, phase, and sleep efficiency was estimated	Offline retrospective
Woods et al. 2014	Distinguish PD from essential tremor using accelerometry	14 subjects with PD and 18 subjects with essential tremor	Smartphones recorded accelerometry of hand tremor during motor tasks	Offline retrospective
Vallance et al. 2011	Assess relationship between depression (assessed via PHQ-9 questionnaires) and activity	2,862 subjects	Wrist-worn devices recorded accelerometry	Offline retrospective

Table A4

Platforms, pilots, and ongoing studies.

Reference or study	Sample size	Methods
Faurholt-Jepsen, M. et al. 2017	400 subjects with BD	Patients will be randomized to either 1) a smartphone-based monitoring system including a feedback loop between patients and clinicians, and cognitive behavioral therapy, or 2) standard treaatment. The outcomes are number and duration of re-admissions,

Reference or study	Sample size	Methods
		2) severity of depressive and manic symptoms, and 3) perceived stress, quality of life, symptomatology, etc.
AURORA	5,000 subjects with trauma	Verily, University of North Carolina, and Harvard University are leading a 19-institution five-year endeavor to perform the most comprehensive observational study of trauma to date. Investigators will examine passive data collection methods using smartphone apps, as well as in-person visits, genomic measurements, neurocognitive tests, patient surveys, and medical record reviews. This collaboration presents a unique opportunity to discover new insights that could translate into fundamental advances in our understanding of post-traumatic conditions. See https://www.nimh.nih.gov/news/science-news/2016/nimh-funded-study-to-track-the-effects-of-trauma.shtml.
Healthy Aging Study	100,000 subjects	The overarching goal is to develop a midlife biomarker of Alzheimer's disease, since it is now well established that the disease begins about 2 decades prior to the onset of clinical symptoms. It is critical to develop new ways to detect the disease in the silent asymptomatic phase in order to develop preventative treatments. To accomplish this goal, the Emory Healthy Aging Study first aims to recruit 100,000 individuals to participate in an online study to assess risk factors identified in health questionnaires and by apps to measure cognition. The second aim is to deeply phenotype a subpopulation of about 3000 or more of these subjects every few years to assess a variety of risk factors by profiling genetics, cardiovascular physiology, blood and spinal fluid biomarkers, brain and retinal imaging. Multi-level longitudinal analyses of subjects profiles, including their amyloid status, will facilitate discovery of new biomarkers. See https://healthyaging.emory.edu/about-the-study/.
Batista, E. et al. 2015	16 subjects	Study of AD and MCI. The System for the Private and Autonomous Surveillance based on Information and Communication Technologies (SIMPATIC) project is a smartphone app-based system for monitoring people with MCI. The smartphone app raises alarms under certain conditions, such as an AD patient leaving a defined geographic zone (e.g. home), not moving after a certain amount of time, moving at too high a speed (suggesting they are utilizing transportation), or the phone battery level reaching too low a level.
Faurholt-Jepsen, M. et al. 2013.	78 subjects	Six month study of BD. The "MONARCA" smartphone app administered subjective questionnaires assessing mood, sleep, medicine intake, etc., and monitored speech duration, social activity, and accelerometry.
RADAR-CNS: Remote Assessment of Disease and Relapse - Central Nervous System	Unknown	A collaborative research program exploring the potential of wearable devices to help prevent and treat depression, multiple sclerosis and epilepsy. Jointly led by King's College London and Janssen Pharmaceutica NV, funded by the Innovative Medicines Initiative, and includes 23 organizations from across Europe and the US.
UCLA Depression Grand Challenge	Study aims to enroll 100,000 people	10-year study with aim of identifying will screen for depression, analyze participants' genetics, measure early adversity and life stress and assess symptoms through remote monitoring using cell phones and wearable devices.
mPower: Mobile Parkinson Disease Study	48,000 people downloaded the app; 9,520 people consented to share data	This study will monitor individual's health and symptoms of PD progression like dexterity, balance and gait using questionnaires and sensors via the Parkinson mPower mobile phone application and wearable devices if available.

References

Abdullah S, Matthews M, Frank E. Automatic detection of social rhythms in bipolar disorder. Journal of the American Medical Informatics Association. 2016; 23(3):538–543. [PubMed: 26977102]

Addington D, Addington J, Eleanor MT, et al. Reliability and validity of a depression rating scale for schizophrenics. Schizophrenia Research. 1992; 6(3):201–208. [PubMed: 1571313]

Addington D, Addington J, Maticka-Tyndale E. Assessing depression in schizophrenia: the Calgary Depression Scale. The British Journal of Psychiatry. 1993; 22:39–44.

Addington D, Addington J, Schissel B. A depression rating scale for schizophrenics. Schizophrenia research. 1990; 3(4):247–51. [PubMed: 2278986]

Aguilera A, Schueller SM, Leykin Y. Daily mood ratings via text message as a proxy for clinic based depression assessment. Journal of Affective Disorders. 2015; 175:471–474. [PubMed: 25679202]

Akselrod S, Gordon D, Ubel F, et al. Power spectrum analysis of heart rate fluctuation: a quantitative probe of beat-to-beat cardiovascular control. Science. 1981; 213(4504):220–22. [PubMed: 6166045]

Albert MV, Toledo S, Shapiro M, et al. Using Mobile Phones for Activity Recognition in Parkinson's Patients. Frontiers in Neurology. 2017; 3:158.

AlHanai T, Ghassem MM. Predicting Latent Narrative Mood using Audio and Physiologic Data. AAAI. 2017

Allan LM, Kerr SRJ, Ballard CG. Autonomic function assessed by heart rate variability is normal in Alzheimer's disease and vascular dementia. Dementia and Geriatric Cognitive Disorders. 2005

Allen J. Photoplethysmography and its application in clinical physiological measurement. Physiological Measurement. 2017

Altman EG, Hedeker D, Peterson JL, et al. The Altman Self-Rating Mania Scale. Biological psychiatry. 1997; 42(10):948–955. [PubMed: 9359982]

Alzheimer's Association. 2016 Alzheimer's disease facts and figures. Alzheimer's Association; 2016. p. 459-509.Tech. rep. 4

Andrews S, Ellis DA, Shaw H, et al. Beyond Self-Report: Tools to Compare Estimated and Real-World Smartphone Use. PLoS ONE. 2015; 10(10):e0139004. [PubMed: 26509895]

Apiquian R, Fresán A, Jairo MD, et al. Variations of rest-activity rhythm and sleep-wake in schizophrenic patients versus healthy subjects: An actigraphic comparative study. Biological Rhythm Research. 2017; 39(1):69–78.

Armontrout J, Torous J, Fisher M, et al. Mobile Mental Health: Navigating New Rules and Regulations for Digital Tools. Current Psychiatry Reports. 2016; 18:10. [PubMed: 26769199]

Asch DA, Muller RW, Volpp KG. Automated Hovering in Health Care - Watching Over the 5000 Hours. New England Journal of Medicine. 2012; 367(1):1–3. [PubMed: 22716935]

Aung M, Matthews M, Choudhury T. Sensing behavioral symptoms of mental health and delivering personalized interventions using mobile technologies. Depression and Anxiety. 2017; 34(7):603– 609. [PubMed: 28661072]

Bagby R, Ryder A, Schuller D, et al. The Hamilton Depression Rating Scale: Has the Gold Standard Become a Lead Weight? American Journal of Psychiatry. 2017; 161(12):2163–2177.

Bär KJ, Boettger M, Koschke M, et al. Non-linear complexity measures of heart rate variability in acute schizophrenia. Clinical Neurophysiology. 2017; 118(9):2009–2015.

Barnett I, Torous J, Staples P, et al. Relapse prediction in schizophrenia through digital phenotyping: a pilot study. Neuropsychopharmacology. 2018 Jan.:1. [PubMed: 29192667]

Barrett PM, Steinhubl SR, Muse ED, et al. Digitising the mind. Lancet. 2017; 389(10082):1877. [PubMed: 28513442]

Batista E, Borràs F, Antoni MB. Monitoring People with MCI: Deployment in a Real Scenario for Low-Budget Smartphones. 2015 6th International Conference on Information, Intelligence, Systems and Applications. 2015:1–6.

Bauer A, Kantelhardt JW, Bunde A, et al. Phase-rectified signal averaging detects quasi-periodicities in non-stationary data. Physica A: Statistical Mechanics and its Applications. 2017; 364:423–434.

Beck AT, Steer RA, Antonio BGK. Beck depression inventory-II. San Antonio. 1996; 78(2):490–498.

Beck AT, Ward CH, Mendelson M, et al. An inventory for measuring depression. Archives of General Psychiatry. 1961; 4(6):561–571. [PubMed: 13688369]

Behar J, Roebuck A, Domingos JS, et al. A review of current sleep screening applications for smartphones. Physiological Measurement. 2013; 34(7):R29–R46. [PubMed: 23771214]

Beiwinkel T, Kindermann S, Maier A, et al. Using Smartphones to Monitor Bipolar Disorder Symptoms: A Pilot Study. JMIR Mental Health. 2016; 3(1):e2. [PubMed: 26740354]

Belmaker RH, Agam G. Major depressive disorder. New England Journal of Medicine. 2008; 358(1): 55–68. [PubMed: 18172175]

Ben-Zeev D, Scherer EA, Wang R, et al. Next-Generation Psychiatric Assessment: Using Smartphone Sensors to Monitor Behavior and Mental Health. Psychiatric Rehabilitation Journal. 2015; 3:218–226.

Berle JO, Hauge ER, Oedegaard KJ, et al. Actigraphic registration of motor activity reveals a more structured behavioural pattern in schizophrenia than in major depression. BMC Research Notes. 2017; 3(1):1–7.

Bot BM, Suver C, Neto EC, et al. The mPower study, Parkinson disease mobile data collected using ResearchKit. Scientific Data. 2016; 3 url: http://www.nature.com/articles/sdata201611.

Brown R, Duma S, Piguet O, et al. Cardiovascular variability in Parkinson's disease and extrapyramidal motor slowing. Clinical Autonomic Research. 2012; 22(4):191–196. [PubMed: 22532273]

Bullock B, Murray G. Reduced amplitude of the 24 hour activity rhythm: a biomarker of vulnerability to bipolar disorder? Clinical Psychological Science. 2014; 2(1):86–96.

Burns MN, Begale M, Duffecy J, et al. Harnessing context sensing to develop a mobile intervention for depression. Journal of Medical Internet Research. 2011; 13(3):e55. [PubMed: 21840837]

Burton C, McKinstry B, Szentagotai T tar A, et al. Activity monitoring in patients with depression: a systematic review. Journal of Affective Disorders. 2013; 145(1):21–28. [PubMed: 22868056]

Buysse D, Reynolds CF, Monk TH, et al. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. Psychiatry Research. 1989

Canzian L, Musolesi M. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. UbiComp 2015. 2015

Capecci M, Pepa L, Verdini F, et al. A smartphone-based architecture to detect and quantify freezing of gait in Parkinson's disease. Gait & Posture. 2016; 50:28–33. [PubMed: 27567449]

Carpenter LL, Tyrka AR, McDougle CJ, et al. Cerebrospinal fluid corticotropin-releasing factor and perceived early-life stress in depressed patients and healthy control subjects. Neuropsychopharmacology. 2004; 29(4):777. [PubMed: 14702025]

Cella M, Okruszek L, Lawrence M, et al. Using wearable technology to detect the autonomic signature of illness severity in schizophrenia. Schizophrenia Research. 2017

Cemer I. Noise Measurement. 2011 Visited on 03/05/2018.

Chang JS, Yoo CS, Yi SH, et al. Differential pattern of heart rate variability in patients with schizophrenia. Progress in Neuro-Psychopharmacology & Biological Psychiatry. 2009; 33(6):991–995. [PubMed: 19427888]

Chow SC. Adaptive Clinical Trial Design. Annual Review of Medicine. 2014; 65(1):405–415.

Clifford, GD. PhD thesis. University of Oxford; 2002. Signal processing methods for heart rate variability.

Clifford GD. The use of sustainable and scalable health care technologies in developing countries. Innovation and Entrepreneurship in Health. 2016; 3:35–46.

Clifford GD, Behar J, Li Q, et al. Signal quality indices and data fusion for determining clinical acceptability of electrocardiograms. Physiological Measurement. 2012a; 33(9):1419–1433. [PubMed: 22902749]

Clifford GD, Clifton DA. Wireless Technology in Disease Management and Medicine. Annual Review of Medicine. 2012b; 63(1):479–492.

Clifford GD, Lopez D, Li Q, et al. Signal quality indices and data fusion for determining acceptability of electrocardiograms collected in noisy ambulatory environments. Computing in Cardiology. 2011; 1419:285–288.

Cohen H, Benjamin J, Geva AB, et al. Autonomic dysregulation in panic disorder and in post-traumatic stress disorder: application of power spectrum analysis of heart rate variability at rest and in response to recollection of trauma or panic attacks. Psychiatry Research. 2000; 96(1):1–13. [PubMed: 10980322]

Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. Journal of Health and Social Behavior. 1983; 24(4):385–396. [PubMed: 6668417]

Collins PY, Patel V, Joestl SS, et al. Grand challenges in global mental health. Nature. 2011; 475(7354):27–30. [PubMed: 21734685]

Copeland LA, Zeber JE, Salloum IM, et al. Treatment Adherence and Illness Insight in Veterans With Bipolar Disorder. The Journal of Nervous and Mental Disease. 2017; 196(1):16.

Costa M, Goldberger A, Peng CK. Multiscale entropy analysis of complex physiologic time series. Physical Review Letters. 2002

Delaney JPA, Brodie DA. Effects of short-term psychological stress on the time and frequency domains of heart-rate variability. Perceptual and Motor Skills. 2000; 91:515–524. [PubMed: 11065312]

Depue RA, Slater JF, Heidi WK, et al. A behavioral paradigm for identifying persons at risk for bipolar depressive disorder: a conceptual framework and five validation studies. Journal of Abnormal Psychology. 1981; 90(5):381. [PubMed: 7298991]

Dodel RC, Singer M, Rudolf KV, et al. The Economic Impact of Parkinson's Disease. Pharmacoeconomics. 1998; 14(3):299–312. [PubMed: 10186468]

Draghici AE, Taylor JA. The Physiological Basis and Measurement of Heart Rate Variability in Humans. Journal of Physiological Anthropology. 2016; 35(1):22–29. [PubMed: 27680542]

Eadicicco L. Americans Check Their Phones 8 Billion Times a Day. Time. 2016

Ebner-Priemer UW, Kuo J, Welch S, et al. A valence-dependent group-specific recall bias of retrospective self-reports: A study of borderline personality disorder in everyday life. The Journal of Nervous and Mental Disease. 2006; 194(10):774–779. [PubMed: 17041290]

Ehrenreich B, Righter B, Rocke D, et al. Are Mobile Phones and Handheld Computers Being Used to Enhance Delivery of Psychiatric Treatment?: A Systematic Review. The Journal of Nervous and Mental Disease. 2017; 199(11):886.

Elhai JD, Frueh BC. Security of electronic mental health communication and record-keeping in the digital age. Journal of Clinical Psychiatry. 2015; 77:2.

Ellis RJ, Ng Y, Zhu S, et al. A validated smartphone-based assessment of gait and gait variability in Parkinson's disease. PLoS ONE. 2015; 10(10):e0141694. [PubMed: 26517720]

Emam KE, Rodgers S, Malin B. Anonymising and sharing individual patient data. BMJ. 2015; 350

Emory University. Emory Healthy Aging Study. 2016. url: https://healthyaging.emory.edu (visited on 03/30/2016).

Epel ES, Blackburn EH, Lin J, et al. Accelerated telomere shortening in response to life stress. Proceedings of the National Academy of Sciences. 2004; 101:49.

Eyben F, Wöllmer M, Schuller B. Opensmile: the munich versatile and fast open-source audio feature extractor. Proceedings of the 18th ACM international conference on Multimedia. 2010:1459–1462.

Faedda GL, Ohashi K, Hernandez M, et al. Actigraph measures discriminate pediatric bipolar disorder from attention-deficit/hyperactivity disorder and typically developing controls. Journal of Child Psychology and Psychiatry. 2016; 57(6):706–716. [PubMed: 26799153]

Fahn, S., Elton, RL. Recent developments in Parkinson's disease. Vol. 2. Macmillan Healthcare Information; 1987. Unified Parkinsons Disease Rating Scale; p. 153-163.

Fasmer OB, Hauge E, Berle J, et al. Distribution of Active and Resting Periods in the Motor Activity of Patients with Depression and Schizophrenia. Psychiatry Investigation. 2015; 13(1):112–120. [PubMed: 26766953]

Faurholt-Jepsen M, Vinberg M, Frost M, et al. Smartphone data as an electronic biomarker of illness activity in bipolar disorder. Bipolar Disorders. 2015; 17(7):715–728. [PubMed: 26395972]

Femminella GD, Rengo G, Kimici K, et al. Autonomic dysfunction in Alzheimer's disease: tools for assessment and review of the literature. Journal of Alzheimer's Disease. 2014; 42:369–377.

Firth J, Cotter J, Torous J, et al. Mobile Phone Ownership and Endorsement of "mHealth" Among People With Psychosis: A Meta-analysis of Cross-sectional Studies. Schizophrenia Bulletin. 2016; 42(2):448–455. [PubMed: 26400871]

Free C, Phillips G, Galli L, et al. The effectiveness of mobile-health technology-based health behaviour change or disease management interventions for health care consumers: a systematic review. PLoS Medicine. 2013; 10(1):e1001362. [PubMed: 23349621]

Freeman D, Sheaves B, Goodwin GM, et al. The effects of improving sleep on mental health (OASIS): a randomised controlled trial with mediation analysis. Lancet Psychiatry. 2017; 4(10):749–758. [PubMed: 28888927]

Gandhi M, Wang T. The Future of Biosensing Wearables. 2014Tech rep

Goetz CG. The Unified Parkinson's Disease Rating Scale (UPDRS): Status and recommendations. Movement Disorders. 2003; 18(7):738–750. [PubMed: 12815652]

Goetz CG, Fahn S, Martinez-Martin P, et al. Movement disorder society-sponsored revision of the unified Parkinson's disease rating scale (MDS-UPDRS): Process, format, and clinimetric testing plan. Movement Disorders. 2007; 22(1):41–47. [PubMed: 17115387]

Goetz CG, Poewe W, Rascol O, et al. Movement Disorder Society Task Force report on the Hoehn and Yahr staging scale: Status and recommendations. Movement Disord. 2004; 19(9):1020–1028. [PubMed: 15372591]

Gracious B, Youngstrom EA, Findling R, et al. Discriminative Validity of a Parent Version of the Young Mania Rating Scale. Journal of the American Academy of Child & Adolescent Psychiatry. 2002; 41(11):1350–1359. [PubMed: 12410078]

Griffiths RI, Kotschet K, Arfon S, et al. Automated assessment of bradykinesia and dyskinesia in Parkinson's disease. Journal of Parkinson's Disease. 2012; 2(1):47–55.

Grippo AJ, Johnson AK. Stress, depression and cardiovascular dysregulation: a review of neurobiological mechanisms and the integration of research from preclinical disease models. Stress. 2009; 12(1):1–21. [PubMed: 19116888]

Grünerbl A, Muaremi A, Osmani V, et al. Smartphone-based recognition of states and state changes in bipolar disorder patients. IEEE Journal of Biomedical and Health Informatics. 2015; 19(1):140–148. [PubMed: 25073181]

Guzzetti S, Borroni E, Garbelli PE, et al. Symbolic Dynamics of Heart Rate Variability. Circulation. 2005; 112(4):465–470. [PubMed: 16027252]

Haapaniemi TH, Pursiainen V, Korpelainen JT, et al. Ambulatory ECG and analysis of heart rate variability in Parkinson's disease. Journal of Neurology, Neurosurgery, and Psychiatry. 2001; 70(3):305–310.

Hamilton M. A rating scale for depression. Journal of neurology, neurosurgery, and psychiatry. 1960; 23:56–62.

Hauge ER, Berle JØ, Oedegaard KJ, et al. Nonlinear analysis of motor activity shows differences between schizophrenia and depression: a study using Fourier analysis and sample entropy. PLOS ONE. 2011; 6(1):e16291. [PubMed: 21297977]

Henry BL, Minassian A, Paulus MP, et al. Heart rate variability in bipolar mania and schizophrenia. Journal of Psychiatric Research. 2010; 44(3):168–176. [PubMed: 19700172]

Hoehn MM, Yahr MD. Parkinsonism: onset, progression and mortality. Neurology. 1967; 17(5):427–442. [PubMed: 6067254]

Horvitz E, Mulligan D. Data, privacy, and the greater good. Science. 2015; 349(6245):253–255. [PubMed: 26185242]

Huse DM, Schulman K, Orsini L, et al. Burden of illness in Parkinson's disease. Movement Disorders. 2005; 20(11):1449–1454. [PubMed: 16007641]

Insel TR. Digital Phenotyping: Technology for a New Science of Behavior. JAMA. 2017

Johnson AE, Ghassemi MM, Nemati S, et al. Machine Learning and Decision Support in Critical Care. Proceedings of the IEEE. 2016; 104:2.

Jongsma HE, Gayer-Anderson C, Lasalvia A, et al. Treated incidence of psychotic disorders in the multinational EU-GEI study. JAMA Psychiatry. 2018; 75(1):36–46. [PubMed: 29214289]

Kahn R, Sommer I, Murray R, et al. Schizophrenia. Nature Reviews Disease Primers. 2015; 1:nrdp201567.

Kallio M, Haapaniemi TH. Heart rate variability in patients with untreated Parkinson's disease. European Journal of Neurology. 2000; 7(6):667–672. [PubMed: 11136353]

Kamdar MR, Wu MJ. Prism: a data-driven platform for monitoring mental health. Pacific Symposium on Biocomputing. 2016; 21:333–344. [PubMed: 26776198]

Karemaker J. An introduction into autonomic nervous function. Physiological Measurement. 2017; 38:aa6782.

Karow A, Pajonk FG, Reimer J, et al. The dilemma of insight into illness in schizophrenia: self- and expert-rated insight and quality of life. European Archives of Psychiatry and Clinical Neuroscience. 2008; 258(3):152–159. [PubMed: 18000637]

Kassavetis P, Saifee TA, Roussos G, et al. Developing a Tool for Remote Digital Assessment of Parkinson's Disease. Movement Disorders Clinical Practice. 2016; 3(1):59–64.

Kay SR, Fiszbein A, Opler LA. The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia. Schizophrenia Bulletin. 1987; 13(2):261–276. [PubMed: 3616518]

Kazdin AE, Rabbitt SM. Novel models for delivering mental health services and reducing the burdens of mental illness. Clinical Psychological Science. 2013; 1(2):170–191.

Kemp AH, Quintana DS, Gray MA, et al. Impact of Depression and Antidepressant Treatment on Heart Rate Variability: A Review and Meta-Analysis. Biological Psychiatry. 2010; 67(11):1067–1074. [PubMed: 20138254]

Kheirkhahan M, Catrine TL, Axtell R, et al. Actigraphy features for predicting mobility disability in older adults. Physiological Measurement. 2016; 37(10):1813–1833. [PubMed: 27653966]

Kim H, Lee H, Lee W, et al. Unconstrained detection of freezing of gait in Parkinson's disease patients using smartphone. 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2015:3751–3754.

Kleiger RE, Stein PK, Bigger JT. Heart Rate Variability: Measurement and Clinical Utility. Annals of Noninvasive Electrocardiology. 2005; 10(1):88–101. [PubMed: 15649244]

Kostikis N, Hristu-Varsakelis D, Arnaoutoglou M, et al. A smartphone-based tool for assessing parkinsonian hand tremor. IEEE Journal of Biomedical and Health Informatics. 2015; 19(6):1835–1842. [PubMed: 26302523]

Kostikis N, Hristu-Varsakelis M, Arnaoutoglou M, et al. Smartphone-based Evaluation of Parkinsonian hand tremor: Quantitative Measurements vs Clinical Assessment Scores. 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2014:906–909.

Krane-gartiser K, Henriksen T, Morken G, et al. Actigraphic assessment of motor activity in acutely admitted inpatients with bipolar disorder. PLoS ONE. 2014; 9(2):e89574. [PubMed: 24586883]

Kroenke K, Spitzer RL, Williams JBW. The PHQ-9 validity of a brief depression severity measure. Journal of General Internal Medicine. 2001

Krystal AD. Psychiatric Disorders and Sleep. Neurologic Clinics. 2012; 30(4):1389–1413. [PubMed: 23099143]

Kubota KJ, Chen JA, Little MA. Machine learning for large-scale wearable sensor data in Parkinson's disease: Concepts, promises, pitfalls, and futures. Movement Disorders. 2016; 31(9):1314–1326. [PubMed: 27501026]

Kuhlmei A, Walther B, Becker T, et al. Actigraphic daytime activity is reduced in patients with cognitive impairment and apathy. European Psychiatry. 2013; 28(2):94–97. [PubMed: 21696925]

Kuhn E, Greene C, Hoffman J, et al. Preliminary Evaluation of PTSD Coach, a Smartphone App for Post-Traumatic Stress Symptoms. Military Medicine. 2014; 179(1):12–18. [PubMed: 24402979]

Kumar RB, Goren ND, Stark DE, et al. Automated integration of continuous glucose monitor data in the electronic health record using consumer technology. Journal of the American Medical Informatics Association. 2016; 23(3):532–537. [PubMed: 27018263]

Lane RD, McRae K, Reiman EM, et al. Neural correlates of heart rate variability during emotion. NeuroImage. 2009; 44(1):213–222. [PubMed: 18778779]

Lawton MP, Brody EM. Assessment of older people: self-maintaining and instrumental activities of daily living. The Gerontologist. 1969; 9(3):179–186. [PubMed: 5349366]

Lee WK, Yoon H, Park KS. Smart ECG Monitoring Patch with Built-in R-Peak Detection for Long-Term HRV Analysis. Annals of Biomedical Engineering. 2015; 44(7):2292–2301. [PubMed: 26558395]

Lee W, Evans A, Williams DR. Validation of a Smartphone Application Measuring Motor Function in Parkinson's Disease. Journal of Parkinson's Disease. 2016; 6(2):371–382.

Lewinsohn PM, Seeley JR, Roberts RE, et al. Center for Epidemiologic Studies Depression Scale (CES-D) as a screening instrument for depression among community-residing older adults. Psychology and Aging. 1997; 12(2):277. [PubMed: 9189988]

Liddell BJ, Kemp AH, Steel Z, et al. Heart rate variability and the relationship between trauma exposure age, and psychopathology in a post-conflict setting. BMC Psychiatry. 2016; 16:133. [PubMed: 27165346]

Lo A, Chernoff H, Zheng T, et al. Why significant variables aren't automatically good predictors. Proceedings of the National Academy of Sciences. 2015; 112(45):13892–13897.

Maes M, Van Bockstaele DR, Gastel AV, et al. The effects of psychological stress on leukocyte subset distribution in humans: evidence of immune activation. Biological Psychiatry. 1999; 39(1):1–9.

Maetzler W, Domingos J, Srulijes K, et al. Quantitative wearable sensors for objective assessment of Parkinson's disease. Movement Disorders. 2013; 28(12):1628–1637. [PubMed: 24030855]

Malarkey WB, Pearl DK, Demers LM, et al. Influence of academic stress and season on 24-hour mean concentrations of ACTH, cortisol, and beta-endorphin. Psychoneuroendocrinology. 1995; 20(5): 499–508. [PubMed: 7675934]

Maria FJ, Busk J, Frost M, et al. Voice analysis as an objective state marker in bipolar disorder. Translational Psychiatry. 2016; 6:7.

Maria FJ, Frost M, Martiny K, et al. Reducing the rate and duration of Re-ADMISsions among patients with unipolar disorder and bipolar disorder using smartphone-based monitoring and treatment – the RADMIS trials: study protocol for two randomized controlled trials. Trials. 2017; 18(1):277. [PubMed: 28619114]

Maria FJ, Frost M, Vinberg M, et al. Smartphone data as objective measures of bipolar disorder symptoms. Psychiatry Research. 2014; 217(1–2):124–127. [PubMed: 24679993]

Maria FJ, Vinberg M, Christensen EM, et al. Daily electronic self-monitoring of subjective and objective symptoms in bipolar disorder - the MONARCA trial protocol (monitoring, treatment and prediction of bipolar disorder episodes): a randomised controlled single-blind trial. BMJ Open. 2013; 3:e003353.

Marin RS. Apathy: Concept, Syndrome, Neural Mechanisms, and Treatment. Seminars in Clininical Neuropsychiatry. 1996; 1(4):304–314.

van de Martel TF. Faking it: social desirability response bias in self-report research. Australian Journal of Advanced Nursing. 2008; 25(4):40–48.

Martin A, Rief W, Klaiberg A, et al. Validity of the Brief Patient Health Questionnaire Mood Scale (PHQ-9) in the general population. General Hospital Psychiatry. 2006; 28(1):71–77. [PubMed: 16377369]

Martin JL, Jeste DV, Sonia AI. Older schizophrenia patients have more disrupted sleep and circadian rhythms than age-matched comparison subjects. Journal of Psychiatric Research. 2005; 39(3): 251–259. [PubMed: 15725423]

Masters CL, Bateman R, Blennow K, et al. Alzheimer's disease. Nature Reviews Disease Primers. 2015; 1:15056.

McConnell MV, Shcherbina A, Pavlovic A, et al. Feasibility of Obtaining Measures of Lifestyle From a Smartphone App: The MyHeart Counts Cardiovascular Health Study. JAMA Cardiology. 2016

Mcguire TG, Miranda J. New Evidence Regarding Racial And Ethnic Disparities In Mental Health: Policy Implications. Health Affairs. 2008; 27:2, 393–403.

Mohr DC, Zhang M, Schueller SM. Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning. Annual Review of Clinical Psychology. 2016; 13:1.

Monk TH, Petrie SR, Hayes AJ, et al. Regularity of daily life in relation to personality, age, gender, sleep quality and circadian rhythms. Journal of Sleep Research. 1994; 3(4):196–205. [PubMed: 10607126]

Monk TK, Flaherty JF, Frank E, et al. The Social Rhythm Metric: an instrument to quantify the daily rhythms of life. Journal of Nervous and Mental Disease. 1990; 178:2.

Montano N, Porta A, Cogliati C, et al. Heart rate variability explored in the frequency domain: A tool to investigate the link between heart and behavior. Neuroscience and Biobehavioral Reviews. 2009; 33(2):71–80. [PubMed: 18706440]

Moore PJ, Little MA, McSharry PE, et al. Forecasting depression in bipolar disorder. IEEE Transactions on Biomedical Engineering. 2012; 59:10, 2801–2807.

Nahshoni E, Aravot D, Aizenberg D, et al. Heart rate variability in patients with major depression. Psychosomatics. 2004; 45(2):129–134. [PubMed: 15016926]

Nakamura T, Kiyono K, Yoshiuchi K, et al. Universal scaling law in human behavioral organization. Physical Review Letters. 2007; 99(13):138103. [PubMed: 17930642]

National Institute of Mental Health. NIMH-Funded Study to Track the Effects of Trauma. 2016. url: https://www.nimh.nih.gov/news/science-news/2016/nimh-funded-study-to-track-the-effects-of-trauma.shtml (visited on 10/17/2016).

akan Nero H, Wallén M, Franzén E, et al. Accelerometer cut points for physical activity assessment of older adults with Parkinson's disease. PLoS ONE. 2015; 10(9):e0135899. [PubMed: 26332765]

Newcomer JW, Hennekens CH. Severe mental illness and risk of cardiovascular disease. JAMA. 2007; 298(15):1794–1796. [PubMed: 17940236]

Niwa F, Kuriyama N, Nakagawa M, et al. Circadian rhythm of rest activity and autonomic nervous system activity at different stages in Parkinson's disease. Autonomic Neuroscience. 2011; 165(2):195–200. [PubMed: 21871844]

Noah B, Keller MS, Mosadeghi S, et al. Impact of remote patient monitoring on clinical outcomes: an updated meta-analysis of randomized controlled trials. npj Digital Medicine. 2017; 1(1):2.

Novartis. Novartis expands alliance with Science 37 to advance virtual clinical trials program. 2018. url: https://www.novartis.com/news/media-releases/novartis-expands-alliance-science-37-advance-virtual-clinical-trials-program (visited on 03/07/2018).

Nutt JG, Wooten G. Diagnosis and initial management of Parkinson's disease. New England Journal of Medicine. 2005; 353:1021–1027. [PubMed: 16148287]

Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. New England Journal of Medicine. 2016; 375(13):1216–1219. [PubMed: 27682033]

O'Brien J, Gallagher P, Stow D, et al. A study of wrist-worn activity measurement as a potential real-world biomarker for late-life depression. Psychological Medicine. 2016; 47(1):93–102. [PubMed: 27667663]

Oh YS, Kim JS, Kim YI, et al. Circadian blood pressure and heart rate variations in de novo Parkinson's disease. Biological Rhythm Research. 2014; 45(3):335–343.

Osipov M, Behzadi Y, Kane JM, et al. Objective identification and analysis of physiological and behavioral signs of schizophrenia. Journal of Mental Health. 2015; 24(5):276–282. [PubMed: 26193048]

Otte C, Gold SM, Penninx BW, et al. Major depressive disorder. Nature Reviews Disease Primers. 2016; 2:16065.

Overall JE, Gorham DR. The Brief Psychiatric Rating Scale. Psychological Reports. 1962; 10(3):799–812.

Palmius N, Osipov M, Bilderbeck AC, et al. A multi-sensor monitoring system for objective mental health management in resource constrained environments. Appropriate Healthcare Technologies for Low Resource Settings. 2014:1–4.

Palmius N, Tsanas A, Saunders KEA, et al. Detecting Bipolar Depression From Geographic Location Data. IEEE Transactions on Biomedical Engineering. 2017; 64(8):1761–1771. [PubMed: 28113247]

Pan D, Dhall R, Lieberman A, et al. A mobile cloud-based Parkinson's disease assessment system for home-based monitoring. JMIR mHealth and uHealth. 2015; 3(1):e29. [PubMed: 25830687]

Patel MS, Asch DA, Volpp KG. Wearable Devices as Facilitators, Not Drivers, of Health Behavior Change. JAMA. 2015; 313(5):459–460. [PubMed: 25569175]

Patel S, Lorincz K, Hughes R, et al. Monitoring Motor Fluctuations in Patients with Parkinson's Disease Using Wearable Sensors. IEEE Transactions on Information Technology in Biomedicine. 2009; 13(6):864–873. [PubMed: 19846382]

Peeples MM, Iyer AK, Cohen JL. Integration of a mobile-integrated therapy with electronic health records: Lessons learned. Journal of Diabetes Science and Technology. 2013; 7(3):602–611. [PubMed: 23759392]

Place S, Blanch-Hartigan D, Rubin C, et al. Behavioral indicators on a mobile sensing platform predict clinically validated psychiatric symptoms of mood and anxiety disorders. Journal of Medical Internet Research. 2017; 19(3):1–9.

Poewe W, Seppi K, Tanner CM, et al. Parkinson disease. Nature Reviews Disease Primers. 2017; 3:17013.

Poushter, J. Smartphone Ownership and Internet Usage Continues to Climb in Emerging Economies. Pew Research Center; 2016. Tech repurl: http://www.pewglobal.org/2016/02/22/smartphone-ownership-and-internet-usage-continues-to-climb-in-emerging-economies/.

Radloff LS. A Self-Report Depression Scale for Research in the General Population. Applied Psychological Measurement. 1977; 1(3):385–401.

Reinertsen E, Nemati S, Vest AN, et al. Heart rate-based window segmentation improves accuracy of classifying posttraumatic stress disorder using heart rate variability measures. Physiological Measurement. 2017a; 38:6.

Reinertsen E, Osipov M, Liu C, et al. Continuous assessment of schizophrenia using heart rate and accelerometer data. Physiological Measurement. 2017b; 38(7):1456–1471. [PubMed: 28653659]

Rodgers MM, Pai VM, Conroy RS. Recent Advances in Wearable Sensors for Health Monitoring. IEEE Sensors Journal. 2015; 15(6):3119–3126.

Roebuck A, Monasterio V, Gederi E, et al. A review of signals used in sleep analysis. Physiological Measurement. 2014; 35(1):R1–R57. [PubMed: 24346125]

Roh T, Hong S, Yoo HJ. Wearable depression monitoring system with heart-rate variability. 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2014:562–565.

Roy SH, Cole BT, Gilmore DL, et al. Resolving Signal Complexities for Ambulatory Monitoring of Motor Function in Parkinson's Disease. 2011 33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2011; 2011:4836–4839.

Rush AJ, Carmody TJ, Reimitz PEP. The Inventory of Depressive Symptomatology (IDS): Clinician (IDS-C) and Self-Report (IDS-SR) ratings of depressive symptoms. International Journal of Methods in Psychiatric Research. 2000; 9(2):45–59.

Rush AJ, Trivedi MH, Ibrahim HM, et al. The 16-Item quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. Biological Psychiatry. 2003; 54(5):573–583. [PubMed: 12946886]

Sadeh A, Acebo C. The role of actigraphy in sleep medicine. Sleep Medicine Reviews. 2002; 6(2): 113–124. [PubMed: 12531147]

Saeb S, Lattie EG, Schueller SM, et al. The relationship between mobile phone location sensor data and depressive symptom severity. PeerJ. 2016a; 4:e2537. [PubMed: 28344895]

Saeb S, Lonini L, Jayaraman A, et al. Voodoo Machine Learning for Clinical Predictions. bioRxiv. 2016b:059774.

Saeb S, Zhang M, Karr CJ, et al. Mobile Phone Sensor Correlates of Depressive Symptom Severity in Daily-Life Behavior: An Exploratory Study. Journal of Medical Internet Research. 2015; 17(7):e175. [PubMed: 26180009]

Sano A, Phillips AJ, Yu AZ, et al. Recognizing academic performance, sleep quality, stress level, and mental health using personality traits, wearable sensors and mobile phones. 2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks. 2015:1–6.

Sano A, Picard RW. Stress recognition using wearable sensors and mobile phones. 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction Stress. 2013

Sano W, Nakamura T, Yoshiuchi K, et al. Enhanced persistency of resting and active periods of locomotor activity in schizophrenia. PLoS ONE. 2012; 7(8):e43539. [PubMed: 22952701]

Savage N. Mobile data: Made to measure. Nature. 2015; 527(7576):S12–S13. [PubMed: 26536217]

Saxena S, Thornicroft G, Knapp M, et al. Resources for mental health: scarcity, inequity, and inefficiency. Lancet. 2007; 370(9590):878–889. [PubMed: 17804062]

Sayers J. The world health report 2001 - Mental health: new understanding, new hope. Bulletin of the World Health Organization. 2001; 79(11):1085.

Schreiber T. Measuring Information Transfer. Physical Review Letters. 2000; 85(2):1–3. [PubMed: 10991144]

Schueller SM, Begale M, Penedo FJ, et al. Purple: A Modular System for Developing and Deploying Behavioral Intervention Technologies. Journal of Medical Internet Research. 2014; 16(7):e181. [PubMed: 25079298]

Seyfert-Margolis V. The evidence gap. Nature Biotechnology. 2018; 36(3):228–232.

Shalev A, Liberzon I, Marmar C. Post-Traumatic Stress Disorder. New England Journal of Medicine. 2017; 376:2459–2469. [PubMed: 28636846]

Shin S, Yeom CW, Shin C, et al. Activity monitoring using a mHealth device and correlations with psychopathology in patients with chronic schizophrenia. Psychiatry Research. 2016; 246:712–718. [PubMed: 27836243]

Solhan MB, Trull TJ, Jahng S, et al. Clinical assessment of affective instability: comparing EMA indices, questionnaire reports, and retrospective recall. Psychological Assessment. 2009; 21(3): 425–436. [PubMed: 19719353]

Sowden G, Huffman JC. The impact of mental illness on cardiac outcomes: a review for the cardiologist. International Journal of Cardiology. 2009

Spielberger, CD., Gorsuch, RL., Lushene, R., et al. Manual for the State-Trait Anxiety Inventory. Consulting Psychologists Press; 1983.

Spitzer RL, Kroenke K, Williams JBW, et al. A Brief Measure for Assessing Generalized Anxiety Disorder. Archives of Internal Medicine. 2006; 166(10):1092. [PubMed: 16717171]

Spitzer RL, Kroenke K, Williams JBW, et al. Validation and Utility of a Self-report Version of PRIME-MD. JAMA. 2012; 282(18):1737–1744.

Stamatakis J, Ambroise J, Crémers J, et al. Finger Tapping Clinimetric Score Prediction in Parkinson's Disease Using Low-Cost Accelerometers. Computational Intelligence and Neuroscience. 2013:717853. [PubMed: 23690760]

Staples P, Torous J, Barnett I, et al. A comparison of passive and active estimates of sleep in a cohort with schizophrenia. npj Schizophrenia. 2017; 3(1):37.

Statista. Wearable device sales revenue worldwide from 2016 to 2022 (in billion US dollars). 2017

Stein PK, Bosner MS, Kleiger RE, et al. Heart rate variability: A measure of cardiac autonomic tone. American Heart Journal. 1994; 127(5):1376–1381. [PubMed: 8172068]

Taft C, Karlsson J, Sullivan M. Do SF-36 summary component scores accurately summarize subscale scores? Quality of Life Research. 2001; 10(5):395–404. [PubMed: 11763202]

Teicher MH. Actigraphy and Motion Analysis: New Tools for Psychiatry. Harvard Review of Psychiatry. 1995; 3:18–35. [PubMed: 9384925]

Thayer JF, Åhs F, Fredrikson M, et al. A meta-analysis of heart rate variability and neuroimaging studies: Implications for heart rate variability as a marker of stress and health. Neuroscience and Biobehavioral Reviews. 2012; 36(2):747–756. [PubMed: 22178086]

Torous J, Chan S, Tan S, et al. Patient Smartphone Ownership and Interest in Mobile Apps to Monitor Symptoms of Mental Health Conditions: A Survey in Four Geographically Distinct Psychiatric Clinics. JMIR Mental Health. 2014; 1(1):e5. [PubMed: 26543905]

Torous J, Kiang MV, Lorme J, et al. New Tools for New Research in Psychiatry: A Scalable and Customizable Platform to Empower Data Driven Smartphone Research. JMIR Mental Health. 2016; 3(2):e16. [PubMed: 27150677]

Trajkovi G, Star evi V, Latas M, et al. Reliability of the Hamilton Rating Scale for Depression: A meta-analysis over a period of 49 years. Psychiatry Research. 2011; 189(1):1–9. [PubMed: 21276619]

Tsanas A, Little MA, McSharry PE, et al. Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity. Journal of The Royal Society Interface. 2011; 8(59):842–855.

Tsanas A, Little MA, McSharry PE, et al. Novel speech signal processing algorithms for high-accuracy classification of Parkinsons disease. IEEE Transactions on Biomedical Engineering. 2012; 59(5): 1264–1271. [PubMed: 22249592]

Tsanas A, Saunders KEA, Bilderbeck AC, et al. Daily longitudinal self-monitoring of mood variability in bipolar disorder and borderline personality disorder. Journal of Affective Disorders. 2016; 205:225–233. [PubMed: 27449555]

Tung JY, Rose RV, Gammada E, et al. Measuring life space in older adults with mild-to-moderate Alzheimer's disease using mobile phone GPS. Gerontology. 2014; 60(2):154–162. [PubMed: 24356464]

U.S. Department of Health and Human Services Food and Drug Administration. Guidance for Industry and FDA Staff: Guidance for the Content of Premarket Submissions for Software Contained in Medical Devices. 2005. url: https://www.fda.gov/downloads/MedicalDevices/.../ucm089593.pdf.

U.S. Department of Health and Human Services Food and Drug Administration. Mobile Medical Applications. 2015. url: https://www.fda.gov/MedicalDevices/DigitalHealth/ MobileMedicalApplications/default.htm.

U.S. Food & Drug Administration. 2015–2016 Global Participation in Clinical Trials Report. 2017Tech rep

Validic. Partners Connected Health And Validic Announce Collaboration To Bring Patient-Generated Data To Patient Care. 2017. url: https://validic.com/news/partners-connected-health-and-validic- announce-collaboration-to-bring-patient-generated-data-to-patient-care/.

Vallance JK, Winkler E, Gardiner PA, et al. Associations of objectively-assessed physical activity and sedentary time with depression: NHANES 2005–2006. Preventative Medicine. 2011; 53(4–5): 284–288.

Van Someren EJW, Swaab D, Colenda C, et al. Bright Light Therapy: Improved Sensitivity to Its Effects on Rest-Activity Rhythms in Alzheimer Patients by Application of Nonparametric Methods. Chronobiology International. 2009; 16(4):505–518.

Vancampfort D, Firth J, Schuch F, et al. Sedentary behavior and physical activity levels in people with schizophrenia, bipolar disorder and major depressive disorder: a global systematic review and meta-analysis. World Psychiatry. 2017; 16(3):308–315. [PubMed: 28941119]

Vigo D, Thornicroft G, Atun R. Estimating the true global burden of mental illness. The Lancet Psychiatry. 2016; 3(2):171–178. [PubMed: 26851330]

Walther S, Horn H, Razavi N, et al. Quantitative motor activity differentiates schizophrenia subtypes. Neuropsychobiology. 2009a; 60(2):80–86. [PubMed: 19752582]

Walther S, Koschorke P, Horn H, et al. Objectively measured motor activity in schizophrenia challenges the validity of expert ratings. Psychiatry Research. 2009b; 169(3):187–190. [PubMed: 19729205]

Wang R, Aung MSH, Abdullah S, et al. 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. 2016

Wang R, Chen F, Chen Z, et al. 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. 2014:3–14.

Ware JE Jr, Sherbourne CD. The MOS 36-item Short-Form Health Survey (SF-36). I. Conceptual framework and item selection. Medical Care. 1992; 90(6):473–483.

Wasserstein RL, Lazar NA. The ASA's Statement on p-Values: Context, Process, and Purpose. The American Statistician. 2016; 70(2):129–133.

Weenk M, Alken A, Engelen L, et al. Stress measurement in surgeons and residents using a smart patch. American Journal of Surgery. 2017

Whiteford HA, Degenhardt L, Rehm J, et al. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. Lancet. 2013; 382(9904):1575–1586. [PubMed: 23993280]

Wichniak A, Skowerska A, Jolanta CW, et al. Actigraphic monitoring of activity and rest in schizophrenic patients treated with olanzapine or risperidone. Journal of Psychiatric Research. 2011; 45(10):1381–1386. [PubMed: 21679968]

Winkler D, Pjrek E, Nicole PR, et al. Actigraphy in patients with seasonal affective disorder and healthy control subjects treated with light therapy. Biological Psychiatry. 2005; 58(4):331–336. [PubMed: 16102546]

Winter DA, Quanbury AO, Reimer GD. Analysis of instantaneous energy of normal gait. Journal of Biomechanics. 1972

Witting W, Kwa IH, Eikelenboom P, et al. Alterations in the circadian rest-activity rhythm in aging and Alzheimer's disease. Biological Psychiatry. 1990; 27(6):563–572. [PubMed: 2322616]

Wittkampf KA, Naeije L, Schene AH, et al. Diagnostic accuracy of the mood module of the Patient Health Questionnaire: a systematic review. General hospital psychiatry. 2007; 29(5):388–395. [PubMed: 17888804]

Wolk, DA., Dickerson, BC. Clinical features and diagnosis of Alzheimer disease. In: Post, TW., editor. UpToDate. 2017.

Woods AM, Nowostawski M, Franz EA, et al. Parkinson's disease and essential tremor classification on mobile device. Pervasive and Mobile Computing. 2014; 13:1–12.

Young RC, Biggs JT, Ziegler VE, et al. A rating scale for mania: reliability, validity and sensitivity. The British Journal of Psychiatry. 1978; 133:429–435. [PubMed: 728692]

Youngstrom EA, Frazier TW, Demeter C, et al. Developing a ten item mania scale from the Parent General Behavior Inventory for children and adolescents. Journal of Clinical Psychiatry. 2008; 69(5):831–839. [PubMed: 18452343]

Zinkhan M, Kantelhardt JW. Sleep Assessment in Large Cohort Studies with High-Resolution Accelerometers. Sleep Medicine Clinics. 2016; 11(4):469–488. [PubMed: 28118871]

Zulli R, Nicosia F, Borroni B, et al. QT Dispersion and Heart Rate Variability Abnormalities in Alzheimer's Disease and in Mild Cognitive Impairment. Journal of the American Geriatrics Society. 2005; 53(12):2135–2139. [PubMed: 16398898]
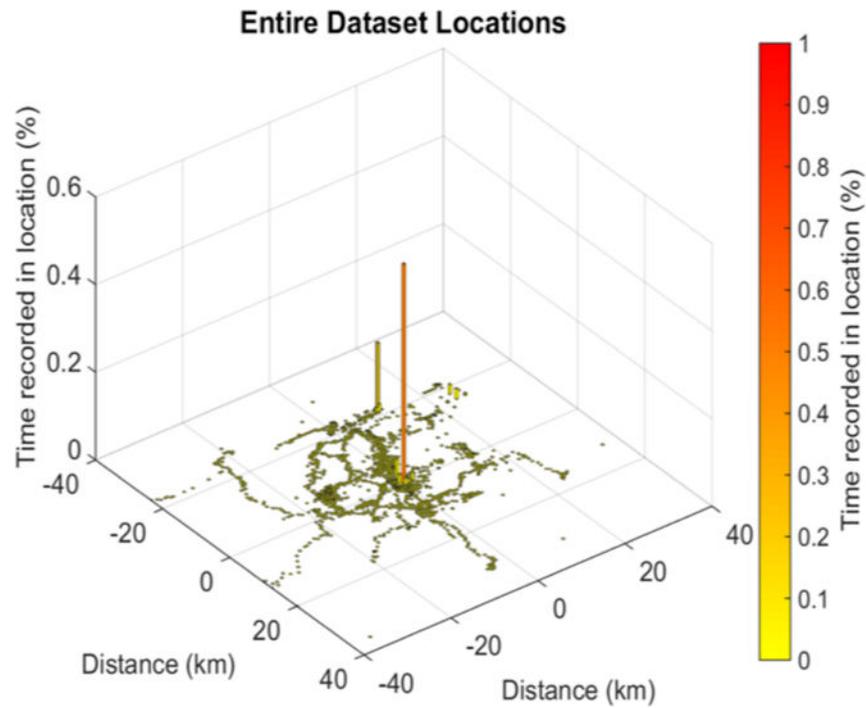
Entire Dataset Locations

Figure 1.
Geolocation data measured via smartphone can track time spent at modal locations. The x- and y-axes are distance from the most commonly visited location. The z-axis is the percentage of total time spent in a given location, with darker orange encoding a higher percentage and a lighter yellow encoding a lower percentage. The dark orange peak at the origin where the individual spends the most time is assumed to be home, and the second-largest peak (z-axis value) where the individual spends the next most time is assumed to be work, or vice-versa if the individual spends more time at work than home.
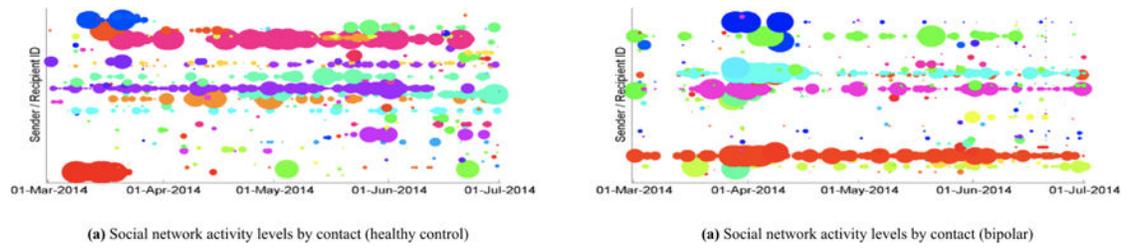
(a) Social network activity levels by contact (healthy control)

(a) Social network activity levels by contact (bipolar)

Figure 2.
Social network activity measured via smartphone can identify mood and illness. The y-axis encodes unique pairings of sender and recipient IDs. The x-axis encodes time. The radius of each colored dot is proportional to the number of calls and text messages in one day. Interactions from a sender-recipient pairing have the same color over time, i.e. all red dots with the same height on the y-axis represent interactions between the same two unique individuals. Qualitatively, (a) healthy controls demonstrate more regular amounts of interaction over time with their social contacts compared to (b) subjects with bipolar disorder who alternate bouts of high and low levels of interaction.
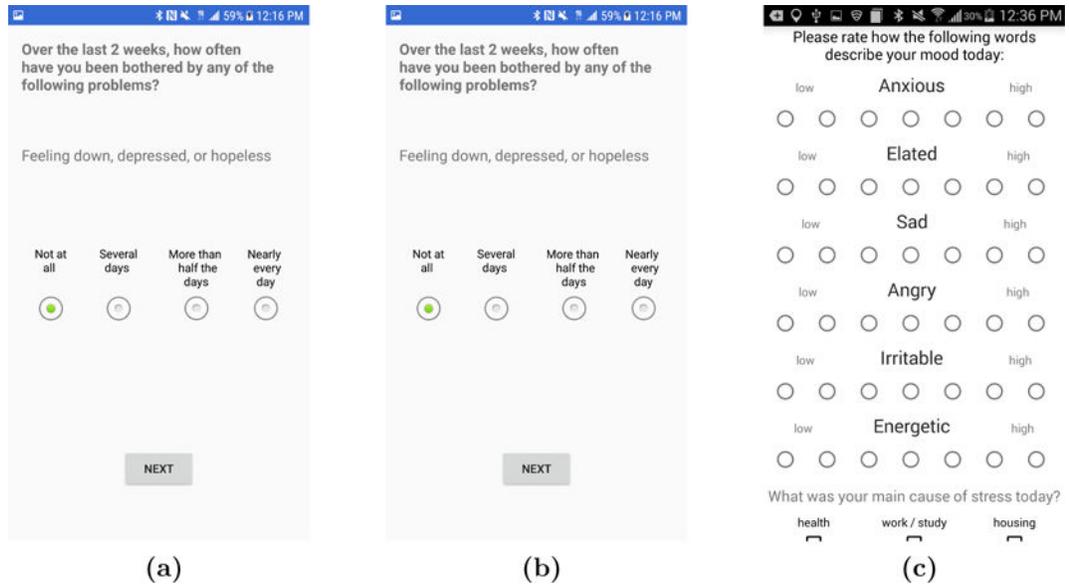
Figure 3.
Screenshots of the AMoSS app for the Android operating system: (a) PHQ-9 questionnaire, (b) Simple mood assessment tool, and (c) "MoodZoom" survey to assess emotional status and mood.
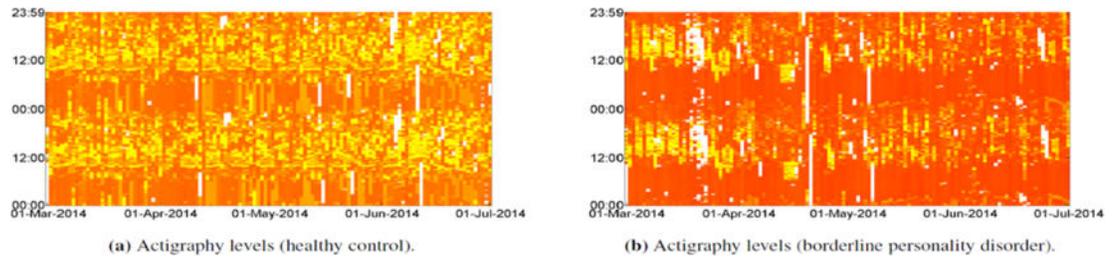
(a) Actigraphy levels (healthy control).

(b) Actigraphy levels (borderline personality disorder).

Figure 4.
A "double-plot" of wearable accelerometry or actigraphy data demonstrates night-to-night patterns. The x-axis is the date, and the y-axis is time of day. Each day is repeated adjacent to and below the previous day. This aligns the nights of data and can be particularly useful in depicting circadian rhythm sleep disorders. (a) Actigraphy levels in a healthy control. (b) Actigraphy levels in a patient with borderline personality disorder.

Table 1

Aberrations in physiology and behavior associated with neuropsychiatric illness that are detectable by sensors in smartphones and wearables

Illness	Sensor type			
	Accelerometry	HR	GPS	Calls & SMS
Stress & depression	Disruptions in circadian rhythm and sleep	Emotion mediates vagal tone which manifests as altered HRV	Irregular travel routine	Decreased social interactions
Bipolar disorder	Disruptions in circadian rhythm and sleep, locomotor agitation during manic episode	ANS dysfunction via HRV measures	Irregular travel routine	Decreased or increased social interactions
Schizophrenia	Disruptions in circadian rhythm and sleep, locomotor agitation or catatonia, diminished overall activity	ANS dysfunction via HRV measures	Irregular travel routine	Decreased social interactions
PTSD	Inconclusive evidence	ANS dysfunction via HRV measures	Inconclusive evidence	Decreased social interactions
Dementia	Dementia Disruptions in circadian rhythm, diminished locomotor activity	Inconclusive evidence	Wandering away from home	Decreased social interaction
Parkinson's disease	Gait impairment, ataxia, dyskinesia	ANS dysfunction via HRV measures	Inconclusive evidence	Voice features can indicate vocal impairment