Smartphone Applications for Monitoring Mood Disorders:

A Systematic Review

Flora Or, M.H.S.[1]

John Torous, M.D.[2]

Juliana Simms, B.S.[3]

Jukka-Pekka Onnela, Sc.D.[1]

Harvard T.H. Chan School of Public Health[1]

Harvard Medical School[2]

Harvard University[3]

Abstract

Background: Despite the great burden of mood disorders, much of the treatment gap remains due to under-recognition. The increasing prevalence of smartphone ownership creates an opportunity to better monitor behavioral phenotypes that are relevant to mood disorders. This review is aimed to evaluate the most up-to-date evidence on the feasibility of smartphone apps as a research tool for monitoring mood disorders.

Methods: We conducted the systematic search in nine bibliographic databases: Embase, PsycINFO, Web of Science, PubMed, CINAHL, EconLit, PAIS, ABI and INSPEC using search terms that capture the concepts of mood disorders and smartphone applications. We included 30 empirical studies that aimed at screening or monitoring mood disorders using passive smartphone data in human subjects.

Results: The included studies attempted to measure and quantify the following five phenotypic categories in the context of mood disorders: social activity, physical activity, sleep, voice, and mobility. Most included studies speculated that increased frequency in these phenotypes would indicate onset or relapse of a manic episode, and vice versa for major depressive episode, and tested the hypotheses with descriptive statistics. Common limitations that impact the robustness of statistical inferences include sparse or lack of clinical assessments as gold standards, small sample size, and data incompleteness.

Conclusion: There is high potential but limited evidence in using smartphone data to monitor mood disorders. Both internal and external validity of included studies were limited. Future studies should consider issues related to selection bias, information that can be learned from pattern of missing data, and within-person designs.

Mood disorders create great burden, yet much of the burden remains unaddressed globally (Kohn, Saxena, Levav, & Saraceno, 2004; Murray et al., 2013; Whiteford et al., 2013). The reasons for treatment gap include lack of awareness, limited access to care, denial of symptoms, stigma, and ineffective care (Corrigan, Larson, & Ruesch, 2009; Corrigan & Watson, 2002; Kohn et al., 2004). Traditionally, the diagnosis and monitoring of mood disorders rely on individuals' recall of symptoms elicited using clinical interviews or paper and pencil questionnaires. Recent studies have shown that recall bias in mood symptoms exist for recall periods as short as one day (Shiffman, Stone, & Hufford, 2008; Torous & Powell, 2015). Patients are more likely to consult their clinicians during a depressive episode rather than a manic episode, leading to misdiagnosis of bipolar disorder as unipolar depression(Bowden, 2001; Hirschfeld & Vornik, 2005). Most current clinical practices are targeted at "an average patient," which often leads to ineffective treatments. The increasing prevalence of smartphones and the advances in smartphone technology create opportunities for more personal, precise, scalable screening and monitoring of mood disorders. We define "screening" as early identification of risk for onset or relapse of mood disorders, and "diagnosis" as the determination of the state of health by a clinician(Težak, Kondratovich, & Mansfield, 2010). While no existing device has the capability to diagnose mood disorders, a typical smartphone has built-in accelerometer, touchscreen, GPS, camera, and Bluetooth and Wi-Fi that enable digital phenotyping which refers to "the moment-by-moment quantification of the individual-level human phenotype in situ using data from personal digital devices."(Onnela & Rauch, 2016) This area of research is in its infancy, and most investigations are pilot studies(Torous & Powell, 2015). Existing reviews of smartphone technologies for mood disorders are limited to those that primarily use this technology to field surveys, either using mobile web, phone calls or text messages (SMS) to

solicit responses. The reviews specific to smartphone applications (apps) designed to screen, monitor and manage symptoms of mood disorders appeared in the past three years (aan het Rot, Hogenelst, & Schoevers, 2012; Martínez-Pérez, De La Torre-Díez, & López-Coronado, 2013; Plaza, Demarzo, Herrera-Mercadal, & García-Campayo, 2013; Shen et al., 2015; Torous & Powell, 2015). To date, there is insufficient data to evaluate the effectiveness of smartphone apps to monitor mental health(Donker et al., 2013; Mohr, Cheung, Schueller, Brown, & Duan, 2013; Plaza et al., 2013; Seko, Kidd, Wiljer, & McKenzie, 2014). As this area of scientific enquiry has grown rapidly in the past two years, this review is aimed to evaluate the most up-to-date evidence on the feasibility of smartphone apps as a research tool for monitoring mood disorders.

Methods

Collaborating with a professional librarian, we conducted the systematic search in nine bibliographic databases: Embase, PsycINFO, Web of Science, PubMed, CINAHL, EconLit, PAIS, ABI, and INSPEC. We used controlled vocabularies including "mobile applications," "cell(ular) phones," "mobile device," "software applications," and text words including iPhone, iPad, Android, Android tablets, and Blackberries to identify studies that used smartphone apps. To identify literature on mood disorders, we searched with controlled vocabulary including "mood disorders," "affective disorders," "depression (emotion)," "dysthymic disorder," "endogenous depression," "postpartum depression," "recurrent depression," "treatment resistant depression," "bipolar disorders," "major depression," "mania," "seasonal affective disorders," along with synonyms of these conditions as text words appearing in abstracts or titles. We identified literature at the intersection of smartphone and mood disorders by combining the two searches with a "AND" boolean. We last updated this search on January 22, 2017. Two authors (FO and JS) screened the titles, abstracts, and texts based on the following inclusion and exclusion criteria.

We include articles that meet all the inclusion criteria: (1) empirical studies (2) written in English; (2) used passive smartphone data; (3) aimed at screening or monitoring mood disorders; (4) in humans. We exclude titles or abstracts written in languages besides English, non-empirical studies (i.e., comment, opinion or narratives), and empirical studies that involved non-human subjects or "app" that referred to concepts besides smartphone apps. In order not to duplicate previous reviews, we exclude studies that used ecological momentary assessment (EMA) because a comprehensive review on studies using electronic self-monitored mood has been recently published(Maria Faurholt-Jepsen, Munkholm, Frost, Bardram, & Kessing, 2016). To

ensure specificity and relevance of this rapidly progressing area of research inquiry, we exclude computer-based, web-based, or text-based studies, and effects of smartphone use. Due to functional differences between therapeutic and monitoring apps, we also exclude studies that used apps for treatment purposes.

Results

The specified searches identified 4960 articles. We removed 1694 duplicates based on authors and titles. We excluded 2809 out of the 3266 remaining publications by titles or abstracts using our exclusion criteria. Using our inclusion criteria, we assessed the abstracts of the remaining 457 records and included 30 studies in this review. These included studies recruited clinical patients in the hospital or in the community, and non-clinical participants for studies that lasted from one week to one year in the U.S. or in Europe. These included studies evaluated the feasibility of using smartphone sensors, including GPS, accelerometer, and Wi-Fi to monitor mood disorders. Many included studies aimed to inform the management of mood disorders but not to replace clinical diagnosis. The incentive structure varied in the types of payment (i.e., financial vs. goods), amount, and qualifying requirements in some studies, whereas other studies did not offer financial incentives for participation (M. Faurholt-Jepsen, Frost, et al., 2014; M. Faurholt-Jepsen et al., 2015)·(A. Grünerbl, Oleksy, P., Bahle, G., Haring, C., Weppner, J., & Lukowicz, P. , 2012; V. Osmani, Maxhuni, A., Grünerbl, A., Lukowicz, P., Haring, C., & Mayora, O. , 2013) (Table 1). The included papers evaluated 16 unique apps aimed to monitor either bipolar disorders or depression; see Table 2 for the functionality and characteristics of these apps.

Data completeness

Incompleteness in smartphone data is common in reviewed studies – planned and unplanned. Planned missingness occurs when the sensors (e.g., GPS) are programmatically triggered at a defined interval (Asselbergs et al., 2016; Canzian & Musolesi, 2015). Unplanned missingness occurs due to technical issues or non-adherence. Adherence is generally defined by the extent to which participants engage in the behavior of interest as recommended by the

researcher(s) (Dunbar, 1984). Reviewed studies measured adherence to enabling the smartphone sensors, or the proportion of completed self-assessment or clinical assessment. Most studies claimed smartphone apps to be excellent disease management tools because they require little input from the patients as individuals generally carry their phones on them (Prociow, Wac, & Crowe, 2012a). However, studies reported substantial data incompleteness from various smartphone sensors, typically due to participants turning off phone sensors or reduced frequency of using a study phone over time (Asselbergs et al., 2016; Dang, Mielke, Diehl, & Haux, 2016; A. Grunerbl et al., 2015)·(A. Grunerbl, Osmani, V., Bahle, G., Carrasco, J. C., Oehler, S., Mayora, O., ... & Lukowicz, P., 2014; A. Muaremi, Gravenhorst, Grunerbl, Arnrich, & Troster, 2014). Some included studies (66.7%) used study phones rather than personal phones. We speculate that the use of study phones might be a reason for missing data due to usage differences between personal and study phones(Belisario, 2015). Existing literature suggests that patients refuse to make calls using study phones and do not carry them at all times (D. Ben-Zeev, E. A. Scherer, R. Wang, H. Xie, & A. T. Campbell, 2015; M. Frost, Doryab, A., Faurholt-Jepsen, M., Kessing, L. V., & Bardram, J. E. , 2013; A. Grunerbl et al., 2015; A. Muaremi et al., 2014).

Adherence to completing self-assessments varied across incentive structures; see Table 1. The adherence rate of completing a one-time self-assessment at week-10 was 78% among students who were incentivized by raffling of technological products (D. Ben-Zeev, E. A. Scherer, et al., 2015). In a community sample that was not financially incentivized, the adherence rate for self-assessment and not deleting the smartphone app for four weeks or more was 22.2%. The adherence rate of bipolar patients' daily self-assessment who received clinician feedback ranged from 88 to 91% (M. Faurholt-Jepsen, Vinberg, et al., 2014; M. Frost, Doryab, A., Faurholt-Jepsen, M., Kessing, L. V., & Bardram, J. E. , 2013) as compared to 55.7% among

bipolar patients who did not receive clinician feedback(Beiwinkel et al., 2016). The shortage of smartphone data may be partially explained by participants switching off smartphone sensors and by authors excluding data that they deemed unsuitable for training and testing when ground truth of the patients' clinical states is missing or when there was a lack of change in mood states among patients(A. Grunerbl, Osmani, V., Bahle, G., Carrasco, J. C., Oehler, S., Mayora, O., ... & Lukowicz, P., 2014; A. Muaremi et al., 2014). In a 12-week observational study where each bipolar patient was expected to generate 84 datasets from GPS, accelerometer, call and text logs and microphone, only 19 to 71 smartphone datasets per patient were used in the data analysis (A. Grunerbl et al., 2015). When smartphone audio, call and text logs were not considered, 35 to 71 smartphone datasets were included in the analysis (A. Grunerbl, Osmani, V., Bahle, G., Carrasco, J. C., Oehler, S., Mayora, O., ... & Lukowicz, P., 2014). In a 12-month study in bipolar inpatients, physical activity data from GPS, accelerometer, and sensors for cell towers were only complete 78.2% of the time, and social activity data from communication logs of the study phones were available on 56.1% of days(Beiwinkel et al., 2016). In community samples, data from 30-50% of the participants were excluded from the analysis because there were insufficient self-assessment data (Asselbergs et al., 2016; Hung, Yang, Chang, Chiang, & Chen, 2016; S. Saeb et al., 2015).

Phenotypes and Mood Disorders

The manifestation of mood disorders involves changes in behavioral phenotypes that can potentially be assessed by smartphone data, and those that cannot be assessed by smartphone data; see Table 3. Most included studies speculated that increased frequency in behavioral phenotypes would indicate onset or relapse of a manic episode, and vice versa for a major depressive episode. We organized this review on the feasibility of monitoring mood disorders

using smartphone data by the five phenotypes that included studies aimed to evaluate: social activity, physical activity, sleep, voice, and mobility. Table 2 maps these five phenotypes onto the types of smartphone data and the Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM 5) mood disorders symptoms. Some studies used one sensor to quantify several phenotypes, while other studies used multiple sensors to ascertain a given phenotype.

Sleep

Although sleep plays a significant role in the development and progression of mood disorders (Association, 2013; Lopresti, 2013), only two studies monitored sleep using smartphone data (D. Ben-Zeev, E. A. Scherer, et al., 2015; Farhan, Lu, et al., 2016). Ben-Zeev et al. (2015) used touch screen "lock" duration, stationary time per accelerometer reading, ambient silence per microphone, and ambient darkness detected by light sensor to approximate daily sleep duration, while Farhan (2016b) used light sensor to infer sleep duration. While the obtained estimates were not verified by actigraphy, Ben-Zeev et al. (2015) adopted an approach that yielded similar results to self-reports in another study(D. Ben-Zeev, E. A. Scherer, et al., 2015). The reported associations between sleep inferred by smartphone data and PHQ9 score were inconsistent across time among university students in Ben-Zeev et al. (2015), whereas Fanhan (2016b) reported more normal sleep patterns among university students with lower PHQ9 scores.

Social Activity

The manifestation of mood disorders in social activity includes social isolation in depression, and increased goal-directed activity or talkativeness in mania,(Beiwinkel et al., 2016; Farhan, Lu, et al., 2016; A. Grünerbl, Oleksy, P., Bahle, G., Haring, C., Weppner, J., & Lukowicz, P. , 2012)[43]. The operational definitions of social activity varied across included studies. Some studies used smartphone communication logs, Bluetooth, and microphones to

capture some aspects of social activity from in-person or phone conversations, and time spent in crowded places or in quiet places (Beiwinkel et al., 2016; D. Ben-Zeev, E. A. Scherer, et al., 2015; Farhan, Yue, et al., 2016; M. Faurholt-Jepsen, Frost, et al., 2014; M. Faurholt-Jepsen et al., 2015; M. Frost, Doryab, A., Faurholt-Jepsen, M., Kessing, L. V., & Bardram, J. E. , 2013; A. Grunerbl et al., 2015; A. Grünerbl, Oleksy, P., Bahle, G., Haring, C., Weppner, J., & Lukowicz, P. , 2012; Prociow et al., 2012a). Other studies used the detection of human speech in the ambient noise by the smartphone microphone as a proxy for social interaction in community samples.

For bipolar patients, the studies mostly ascertained social activities using information such as daily call times, call durations, and number of unique contacts from call and text logs (Beiwinkel et al., 2016; M. Faurholt-Jepsen, Frost, et al., 2014; M. Faurholt-Jepsen et al., 2015; M. Frost, Doryab, A., Faurholt-Jepsen, M., Kessing, L. V., & Bardram, J. E. , 2013; A. Grunerbl et al., 2015; A. Muaremi et al., 2014; V. Osmani, 2015). The strength of call and text logs is data completeness independent of sensors which the subjects may switch off. However, 16.7-40% of the patients did not make calls with the assigned study phones (A. Grünerbl et al., 2015; Amir Muaremi, Gravenhorst, Grünerbl, Arnrich, & Tröster, 2014). While some evidence suggested that reduced text messages frequency was associated with increased depressive symptoms (Beiwinkel et al., 2016), one study found that mildly depressed patients had longer and a greater number of calls relative to those who were severely depressed and those who were healthy (A. Grünerbl, Oleksy, P., Bahle, G., Haring, C., Weppner, J., & Lukowicz, P. , 2012). The relationship between calling and texting behavior and severity of illness may be nonlinear.

Voice

Variations in speech patterns might be indicative of persistent elevated, expansive, or

irritable mood; increased talkativeness; racing thoughts in mania, and low mood in depression (Association, 2013). Speech recordings from therapy session and vocal exercise have shown promising clinical utility for mental illness in the past (Cummins, 2015), which leads to the question whether smartphone audio data could assist monitoring of mood disorders. Reviewed studies used audio features, such as number of conversations, speaking length, pitch, and volume, to ascertain clinically relevant outcomes including mood and social rhythm. However, the completeness of smartphone audio data was strongly affected by patients' adherence to carrying and using the study phones as prescribed. Despite this limitation, included studies reported that audio features from daily calls classified patients into one of the seven possible mood states, ranging from severe depression to severe mania, with an accuracy of 70-80% (A. Grunerbl et al., 2015; A. Maxhuni, Muñoz-Meléndez, A., Osmani, V., Perez, H., Mayora, O., & Morales, E. F. , 2016; A. Muaremi et al., 2014; V. Osmani, 2015). Audio features resulted in classification accuracy similar to that of call and text metadata, and a fusion of both data streams did not improve the classification accuracy (69-83%) (A. Grunerbl et al., 2015; A. Muaremi et al., 2014; V. Osmani, 2015). However, the cost and quality of data collection differ between these two data streams. Smartphone audio data are more voluminous to collect, often involve legal and ethical considerations, while call and text metadata tend to be more complete and readily available in Android devices. Smartphone audio data from daily calls outside of clinical assessment seemed to have difficulties differentiating depression from euthymia(Gideon, Provost, & McInnis, 2016; Karam et al., 2014). Contrary to the expectation that increased speech frequency was associated only with (hypo)mania, increased speech frequency in audio was also observed during the transition from hypomania to depression (Guidi et al., 2015) and the correlations between audio features and mood were relatively weak(S. Abdullah et al., 2016;

Guidi et al., 2015). The findings in the reviewed papers suggest that GPS and accelerometer data potentially provide superior performance in classifying or predicting clinically relevant outcomes without the burden of collecting audio data from patients (A. Grunerbl et al., 2015; A. Maxhuni, Muñoz-Meléndez, A., Osmani, V., Perez, H., Mayora, O., & Morales, E. F. , 2016; A. Muaremi et al., 2014; V. Osmani, 2015); see Table 4.

Mobility

The reviewed literature speculated linkage between mood disorders and mobility. Depressive symptoms, such as loss of motivation and social withdrawal were expected to be negatively associated with distance traveled, regularity in location patterns, and time spent indoors; and vice versa for a manic episode. Some studies used GPS with or without other sensors (e.g., accelerometer, microphone) to ascertain changes in social rhythms (Saeed Abdullah, 2016) or mood in bipolar patients (A. Grunerbl, Osmani, V., Bahle, G., Carrasco, J. C., Oehler, S., Mayora, O., ... & Lukowicz, P., 2014). Other studies investigated the relationship between daily stress, depression severity and physical movement in community samples recruited from craigslist (S. Saeb et al., 2015) or from universities(D. Ben-Zeev, E. A. Scherer, et al., 2015; Farhan, Lu, et al., 2016; Farhan, Yue, et al., 2016). Location data from GPS were often summarized into distance travelled, number of location clusters (Saeed Abdullah et al., 2016; Dror Ben-Zeev, Emily A Scherer, Rui Wang, Haiyi Xie, & Andrew T Campbell, 2015; Gruenerbl et al., 2014), ratio of time spent outdoor to time spent indoor (A. Grünerbl et al., 2012), and the regularity of travel patterns (S. Saeb et al., 2015) on a 24-hour cycle. When GPS data were unavailable, the studies used number of cell towers detected, wearable light sensors that distinguish between outdoor natural and indoor artificial light sources, and detection of Bluetooth devices tied to known locations to monitor mobility.

Analyses of mobility data were exploratory. Common approaches to investigate the association between statistical summaries of location data and clinical outcomes (e.g., PHQ-9 score, YMRS, HAMD), were standard statistical methods, such as linear regression (Dror Ben-Zeev et al., 2015; A. Grünerbl et al., 2012; Sohrab Saeb, Mi Zhang, Christopher J Karr, et al., 2015). Exploratory analysis using t-tests suggested that circadian movement, normalized entropy, location variance, home stay, phone usage duration, and phone usage frequency were different between individuals with PHQ9 scores ≥ 5 and those whose with PHQ9 score ≤ 5 in a community sample recruited from craigslist (Sohrab Saeb, Mi Zhang, Christopher J Karr, et al., 2015). Correlational analyses between mobility data, including circadian movement, location variance, normalized entropy and home stay, and depression score suggested that inactivity and reduced regularity in daily routines were associated with higher depression score (Sohrab Saeb, Mi Zhang, Mary Kwasny, et al., 2015).

Studies that used machine learning to predict clinical outcomes (i.e., social rhythm, mood, PHQ9 score) were able to infer social rhythm stability in bipolar patients using number of location clusters, distance traveled, frequency of conversation inferred from audio data, and duration of non-sedentary activity on a daily basis. Abdullah et al. (2016) performed feature ranking, where least-contributing features to the model were discarded until the most important features remained, and they found location cluster and total distance traveled over a day to be the most important features (Saeed Abdullah et al., 2016). Another study that used machine learning and summary statistics such as time spent outdoors, distance traveled, entropy, percentage of time spent at home was able to predict mood state with 80%-87% accuracy (Farhan, Lu, et al., 2016; A. Grünerbl et al., 2012). Addition of accelerometer did not improve the prediction accuracy of mood states (A. Grünerbl et al., 2012). Notwithstanding the potential of using

smartphone location data to monitor mood disorders, there were several limitations, such as sparse smartphone and ground truth data.

Physical Activity

Traditionally, subjective reports of changes in physical activity have been used to ascertain depressive symptoms, such as loss of energy, psychomotor retardation, and symptoms of mania, including irritability, excessive energy, reduced need for sleep, and psychomotor agitation(A. Maxhuni, Muñoz-Meléndez, A., Osmani, V., Perez, H., Mayora, O., & Morales, E. F. , 2016). The advancement of mobile technology provides an opportunity to overcome limitations of patient self-reports on physical activity (e.g., recall bias) by using smartphone accelerometer data to monitor bipolar disorders in clinical samples over 2-24 weeks (Saeed Abdullah et al., 2016; M. Frost, Doryab, Faurholt-Jepsen, Kessing, & Bardram, 2013; A. Grünerbl et al., 2015; A. Grünerbl et al., 2012; A. Maxhuni et al., 2016; Venet Osmani et al., 2013; Prociow, Wac, & Crowe, 2012b) and to screen for depression in community samples over 2-10 weeks (D. Ben-Zeev, E. A. Scherer, et al., 2015; Sohrab Saeb, Mi Zhang, Mary Kwasny, et al., 2015). Smartphone accelerometer data was often summarized into ratio of stationary to sedentary duration (Saeed Abdullah, 2016) or the ratio of time spent moving to the time with little or no movement (D. Ben-Zeev, E. A. Scherer, et al., 2015; Farhan, Yue, et al., 2016; A. Grünerbl et al., 2012). Two challenges in processing accelerometer data were the physical orientation of the phone and to determine whether the person was being stationary or not carrying the phone in times of movement. To address issues of phone orientation, which was largely unknown to the researchers, rotationally invariant statistical summaries were used in the analysis (Dror Ben-Zeev et al., 2015; A. Grünerbl et al., 2015; Venet Osmani et al., 2013). To define stationary states, studies used pre-determined or experimentally determined threshold to

distinguish lack of movement from missing data. For example, complete lack of movement (A. Grunerbl, Osmani, V., Bahle, G., Carrasco, J. C., Oehler, S., Mayora, O., ... & Lukowicz, P., 2014) or movement slower than 1km/h were interpreted as the phone not being on the patient (Farhan, Yue, et al., 2016; Sohrab Saeb, Mi Zhang, Mary Kwasny, et al., 2015). One study derived this threshold experimentally(D. Ben-Zeev, S. M. Schueller, et al., 2015). Findings regarding the clinical validity of using smartphone accelerometer data to monitor mood disorders were mixed. Among bipolar patients, daily sedentary time was weakly correlated with mood and moderately correlated with self-assessed energy score, but was the third most important feature in predicting social rhythm stability using support vector machine (Saeed Abdullah et al., 2016). Based on Pearson correlation between physical activity levels (i.e., low, moderate, high) and mood states, the authors concluded that physical activity in the morning was associated with mood scores much more strongly than daily activity levels (Venet Osmani et al., 2013). Rather than evaluating activity level on scheduled cycle, one study used smartphone acceleration data during phone conversation and semi-supervised learning, which resulted in over 80% prediction accuracy in mood states (A. Maxhuni et al., 2016); see Table 5.

Conclusions

The area of inquiry regarding the use of smartphone to monitor mood disorders is in its infancy. Comparison across studies included is difficult due to differences in the sample characteristics, incentives, measures, and methods. Common limitations include sparse or lack of clinical assessments as gold standards and data incompleteness, which led to the challenges in data management and analytic approaches. There is considerable selection bias of subjects due to unwillingness to use an Android study phones or to participate in research (Dror Ben-Zeev et al., 2015; Faurholt‐Jepsen et al., 2015; M. Frost et al., 2013; A. Grünerbl et al., 2015; Amir Muaremi et al., 2014). Most existing studies provided participants with study phones, in part because some participants did not own smartphones compatible with the apps(Saeb, Lattie, Schueller, Kording, & Mohr, 2016). We suspect that participants use study phones and personal phone differently. Smartphone apps that support both Android and iOS phones would facilitate the use of personal rather than study phones(Farhan, Yue, et al., 2016), thus reducing data incompleteness and erratic usage.

In addition to selection bias, the uncertainty of mood states associated with missing observations brings into the question the internal validity of the collected data. It was challenging to assess the robustness of statistical conclusions of the included studies due to missing data. For instance, authors of the studies excluded data from their analyses when data were deemed unsuitable for training models (e.g., unavailable ground truth data, lack of change in mood states) (A. Grunerbl, Osmani, V., Bahle, G., Carrasco, J. C., Oehler, S., Mayora, O., ... & Lukowicz, P., 2014; A. Muaremi et al., 2014). A study using a community sample excluded, in the analytic stage, half of the participants whose data were unavailable more than half of the time (S. Saeb et al., 2015). Most of the included studies did not consider the validity of statistical inferences in the presence of missing data. We conjecture that data missingness in these included studies is

likely associated with the state or condition of the subject rather than complete randomness (Wahle, Kowatsch, Fleisch, Rufer, & Weidt, 2016), which requires further enquiry in the future.

As a further attempt to address some of the existing limitations, our team has developed a research platform for high throughput smartphone based digital phenotyping, designed for collecting and storing system using AWS cloud computer infrastructure, and the iOS front end or android back end systems. The platform will contain protocols that can be imported or exported, customized to fit the needs of different studies, and will be made openly available to investigators. The goal is to facilitate transparent and meticulous reporting of data management and analysis to ensure replicabilty which is particularly important in this field where only 6% of medical studies are completely reproducible according to one estimate (Prinz, Schlange, & Asadullah, 2011).

While most included studies expected the intensity of phenotypes to increase from severe depression to normal state, and from normal state to severe mania, many included studies presented counterintuitive findings. Contrary to the expected increase in physical and social activity associated with worsening of mania or lessening of depression, some studies found that increased activities were associated with the alleviation of both depressive and manic symptoms. Future research should avoid overly simplistic model specifications in modeling the complex behavioral variations of mood disorders.

Within-subject and between-subject design could yield dramatically different conclusions(Beiwinkel et al., 2016; A. Grünerbl et al., 2012), where individual difference across mood states ranged from 35-2700% in one study (A. Grünerbl et al., 2012). Potential reasons for these observed variations might be differences in the participants' lifestyle and phone usage. Within-person variations in social activity may also be more relevant than between-person variation. For instance, text and call logs would be a less valid indicator for social activity for

people who use social media as primary forms of communications, although within-subject comparisons over time would remain valid among these individuals. We believe that it would be useful to survey participants about their phone use habits prior to the start of the study, and draw comparisons longitudinally using within-person analyses. Given the variation in expected mobility patterns throughout the week, it would be meaningful to stratify mobility data into week days and weekends (Saeb et al., 2016). While it is feasible to use smartphone data to monitor mood disorders, it is important to consider issues surrounding selection bias, data management, and analytic approaches. While most included studies focused on the association between self-reports and passively collected smartphone data, the use of passively collected smartphone data to predict clinical events or crises, such as hospitalization, could be potentially fruitful. Greater reliance on passive data would better facilitate large sample sizes over long study periods in the future, thus alleviate concerns that understandably exist, given the novelty of the field.

The 30 included studies suggested that there is high potential but limited evidence in using smartphone data to measure and quantify behavioral phenotypes in the context of mood disorders. While most studies speculated associations between the symptom severity and the phenotypic categories, the evidence is inconclusive due to issues such as selection bias and missing data, we think that pattern of missing data, and within-person designs are potentially fruitful areas to explore in future studies.

Acknowledgement

JT is supported by the Natalia Mental Health Foundation and a Dupont Warren Fellowship from the Harvard Medical School Department of Psychiatry. JPO is supported by NIH/NIMH 1DP2MH103909-01 (PI: Onnela) and the Harvard McLennan Dean's Challenge Program (PI: Onnela).

References

aan het Rot, M., Hogenelst, K., & Schoevers, R. A. (2012). Mood disorders in everyday life: A systematic review of experience sampling and ecological momentary assessment studies. *Clinical psychology review, 32*(6), 510-523.

Abdullah, S., Matthews, M., Frank, E., Doherty, G., Gay, G., & Choudhury, T. (2016). Automatic detection of social rhythms in bipolar disorder. *Journal of the American Medical Informatics Association, 23*(3), 538-543.

Abdullah, S., Matthews, M., Frank, E., Doherty, G., Gay, G., & Choudhury, T. (2016). Automatic detection of social rhythms in bipolar disorder. *J Am Med Inform Assoc, 23*(3), 538-543. doi:10.1093/jamia/ocv200

Asselbergs, J., Ruwaard, J., Ejdys, M., Schrader, N., Sijbrandij, M., & Riper, H. (2016). Mobile phone-based unobtrusive ecological momentary assessment of day-to-day mood: an explorative study. *Journal of medical Internet research, 18*(3).

Association, A. P. (Ed.) (2013). *Diagnostic and statistical manual of mental disorders* (5 ed.). Arlington, VA: American Psychiatric Publishing.

Beiwinkel, T., Kindermann, S., Maier, A., Kerl, C., Moock, J., Barbian, G., & Rössler, W. (2016). Using smartphones to monitor bipolar disorder symptoms: a pilot study. *JMIR mental health, 3*(1).

Belisario, J. S. M. J. J., Huckvale K, O'Donoghue J, Morrison CP, Car J. (2015). Comparison of self-administered survey questionnaire responses collected using mobile apps versus other methods. *Cochrane Database Syst Rev, 27*(7).

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

Ben-Zeev, D., Scherer, E. A., Wang, R., Xie, H., & Campbell, A. T. (2015). Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health. *Psychiatr Rehabil J, 38*(3), 218-226. doi:10.1037/prj0000130

Ben-Zeev, D., Schueller, S. M., Begale, M., Duffecy, J., Kane, J. M., & Mohr, D. C. (2015). Strategies for mHealth Research: Lessons from 3 Mobile Intervention Studies. *Administration and Policy in Mental Health and Mental Health Services Research, 42*(2), 157-167. doi:10.1007/s10488-014-0556-2

Bowden, C. L. (2001). Strategies to reduce misdiagnosis of bipolar depression. *Psychiatric Services, 52*(1), 51-55.

Canzian, L., & Musolesi, M. (2015). *Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis.* Paper presented at the Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing.

Corrigan, P. W., Larson, J. E., & Ruesch, N. (2009). Self‐stigma and the "why try" effect: impact on life goals and evidence‐based practices. *World Psychiatry, 8*(2), 75-81.

Corrigan, P. W., & Watson, A. C. (2002). Understanding the impact of stigma on people with mental illness. *World psychiatry, 1*(1), 16-20.

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T. F. . (2015). A review of depression and suicide risk assessment using speech analysis. *Speech Communication, 71*, 10-49.

Dang, M., Mielke, C., Diehl, A., & Haux, R. (2016). Accompanying Depression with FINE-A Smartphone-Based Approach. *Studies in health technology and informatics, 228*, 195.

Donker, T., Petrie, K., Proudfoot, J., Clarke, J., Birch, M.-R., & Christensen, H. (2013). Smartphones for smarter delivery of mental health programs: a systematic review. *Journal of medical Internet research, 15*(11), e247.

Dunbar, J. (1984). Adherence measures and their utility. *Controlled Clinical Trials, 5*(4), 515-521.

Farhan, A. A., Lu, J., Bi, J., Russell, A., Wang, B., & Bamis, A. (2016). *Multi-view Bi-Clustering to Identify Smartphone Sensing Features Indicative of Depression.* Paper presented at the Connected Health: Applications, Systems and Engineering Technologies (CHASE), 2016 IEEE First International Conference on.

Farhan, A. A., Yue, C., Morillo, R., Ware, S., Lu, J., Bi, J., . . . Wang, B. (2016). *Behavior vs. Introspection: Refining prediction of clinical depression via smartphone sensing data.* Paper presented at the 7th Conference on Wireless Health, WH.

Faurholt-Jepsen, M., Frost, M., Vinberg, M., Christensen, E. M., Bardram, J. E., & Kessing, L. V. (2014). Smartphone data as objective measures of bipolar disorder symptoms. *Psychiatry Res, 217*(1-2), 124-127. doi:10.1016/j.psychres.2014.03.009

Faurholt-Jepsen, M., Munkholm, K., Frost, M., Bardram, J. E., & Kessing, L. V. (2016). Electronic self-monitoring of mood using IT platforms in adult patients with bipolar disorder: A systematic review of the validity and evidence. *BMC psychiatry, 16*(1), 7.

Faurholt-Jepsen, M., Ritz, C., Frost, M., Mikkelsen, R. L., Christensen, E. M., Bardram, J., . . . Kessing, L. V. (2015). Mood instability in bipolar disorder type I versus type II-

continuous daily electronic self-monitoring of illness activity using smartphones. *Journal of Affective Disorders, 186*, 342-349. doi:10.1016/j.jad.2015.06.026

Faurholt-Jepsen, M., Vinberg, M., Frost, M., Christensen, E. M., Bardram, J., & Kessing, L. V. (2014). Daily electronic monitoring of subjective and objective measures of illness activity in bipolar disorder using smartphones--the MONARCA II trial protocol: a randomized controlled single-blind parallel-group trial. *BMC Psychiatry, 14*, 309. doi:10.1186/s12888-014-0309-5

Faurholt‐Jepsen, M., Vinberg, M., Frost, M., Christensen, E. M., Bardram, J. E., & Kessing, L. V. (2015). Smartphone data as an electronic biomarker of illness activity in bipolar disorder. *Bipolar disorders, 17*(7), 715-728.

Frost, M., Doryab, A., Faurholt-Jepsen, M., Kessing, L. V., & Bardram, J. E. (2013). *Supporting disease insight through data analysis: refinements of the monarca self-assessment system.* Paper presented at the Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing.

Frost, M., Doryab, A., Faurholt-Jepsen, M., Kessing, L. V., & Bardram, J. E. . (2013). *Supporting disease insight through data analysis: refinements of the monarca self-assessment system.* Paper presented at the Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing.

Gideon, J., Provost, E. M., & McInnis, M. (2016). *Mood state prediction from speech of varying acoustic quality for individuals with bipolar disorder.* Paper presented at the Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on.

Gruenerbl, A., Osmani, V., Bahle, G., Carrasco, J. C., Oehler, S., Mayora, O., . . . Lukowicz, P. (2014). *Using smart phone mobility traces for the diagnosis of depressive and manic episodes in bipolar patients.* Paper presented at the Proceedings of the 5th Augmented Human International Conference.

Grünerbl, A., Muaremi, A., Osmani, V., Bahle, G., Oehler, S., Tröster, G., . . . Lukowicz, P. (2015). Smartphone-based recognition of states and state changes in bipolar disorder patients. *IEEE Journal of Biomedical and Health Informatics, 19*(1), 140-148.

Grunerbl, A., Muaremi, A., Osmani, V., Bahle, G., Ohler, S., Troster, G., . . . Lukowicz, P. (2015). Smartphone-based recognition of states and state changes in bipolar disorder patients. *IEEE J Biomed Health Inform, 19*(1), 140-148. doi:10.1109/jbhi.2014.2343154

Grünerbl, A., Oleksy, P., Bahle, G., Haring, C., Weppner, J., & Lukowicz, P. (2012). *Towards smart phone based monitoring of bipolar disorder.* Paper presented at the Proceedings of the Second ACM Workshop on Mobile Systems, Applications, and Services for HealthCare.

Grünerbl, A., Oleksy, P., Bahle, G., Haring, C., Weppner, J., & Lukowicz, P. . (2012). *Towards smart phone based monitoring of bipolar disorder.* Paper presented at the In Proceedings of the Second ACM Workshop on Mobile Systems, Applications, and Services for HealthCare.

Grunerbl, A., Osmani, V., Bahle, G., Carrasco, J. C., Oehler, S., Mayora, O., ... & Lukowicz, P. (2014). *Using smart phone mobility traces for the diagnosis of depressive and manic episodes in bipolar patients.* Paper presented at the In Proceedings of the 5th Augmented Human International Conference.

Guidi, A., Salvi, S., Ottaviano, M., Gentili, C., Bertschy, G., de Rossi, D., . . . Vanello, N. (2015). Smartphone Application for the Analysis of Prosodic Features in Running Speech with a Focus on Bipolar Disorders: System Performance Evaluation and Case Study. *Sensors (Basel), 15*(11), 28070-28087. doi:10.3390/s151128070

Hirschfeld, R., & Vornik, L. A. (2005). Bipolar disorder—costs and comorbidity. *Am J Manag Care, 11*(3 Suppl), S85-S90.

Hung, G. C.-L., Yang, P.-C., Chang, C.-C., Chiang, J.-H., & Chen, Y.-Y. (2016). Predicting Negative Emotions Based on Mobile Phone Usage Patterns: An Exploratory Study. *JMIR Research Protocols, 5*(3).

Karam, Z. N., Provost, E. M., Singh, S., Montgomery, J., Archer, C., Harrington, G., . . . Ieee. (2014). ECOLOGICALLY VALID LONG-TERM MOOD MONITORING OF INDIVIDUALS WITH BIPOLAR DISORDER USING SPEECH *2014 Ieee International Conference on Acoustics, Speech and Signal Processing*. New York: Ieee.

Kohn, R., Saxena, S., Levav, I., & Saraceno, B. (2004). The treatment gap in mental health care. *Bulletin of the World health Organization, 82*(11), 858-866.

Lopresti, A. L., Hood, S. D., & Drummond, P. D. (2013). A review of lifestyle factors that contribute to important pathways associated with major depression: diet, sleep and exercise. . *J Affect Disord, 148*(1), 12-27.

Martínez-Pérez, B., De La Torre-Díez, I., & López-Coronado, M. (2013). Mobile health applications for the most prevalent conditions by the World Health Organization: review and analysis. *Journal of medical Internet research, 15*(6), e120.

Maxhuni, A., Muñoz-Meléndez, A., Osmani, V., Perez, H., Mayora, O., & Morales, E. F. (2016). Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients. *Pervasive and Mobile Computing, 31*, 50-66.

Maxhuni, A., Muñoz-Meléndez, A., Osmani, V., Perez, H., Mayora, O., & Morales, E. F. . (2016). Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients. *Pervasive and Mobile Computing*.

Mohr, D. C., Cheung, K., Schueller, S. M., Brown, C. H., & Duan, N. (2013). Continuous evaluation of evolving behavioral intervention technologies. *American journal of preventive medicine, 45*(4), 517-523.

Muaremi, A., Gravenhorst, F., Grunerbl, A., Arnrich, B., & Troster, G. (2014). Assessing Bipolar Episodes Using Speech Cues Derived from Phone Calls. In P. Cipresso, A. Matic, & G. Lopez (Eds.), *Pervasive Computing Paradigms for Mental Health* (Vol. 100, pp. 103-114). Berlin: Springer-Verlag Berlin.

Muaremi, A., Gravenhorst, F., Grünerbl, A., Arnrich, B., & Tröster, G. (2014). *Assessing bipolar episodes using speech cues derived from phone calls.* Paper presented at the International Symposium on Pervasive Computing Paradigms for Mental Health.

Murray, C. J., Vos, T., Lozano, R., Naghavi, M., Flaxman, A. D., Michaud, C., . . . Abdalla, S. (2013). Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The lancet, 380*(9859), 2197-2223.

Onnela, J.-P., & Rauch, S. L. (2016). Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *Neuropsychopharmacology*.

Osmani, V. (2015). Smartphones in Mental Health: Detecting Depressive and Manic Episodes. *Ieee Pervasive Computing, 14*(3), 10-13.

Osmani, V., Maxhuni, A., Grünerbl, A., Lukowicz, P., Haring, C., & Mayora, O. (2013). *Monitoring activity of patients with bipolar disorder using smart phones.* Paper presented at the Proceedings of International Conference on Advances in Mobile Computing & Multimedia.

Osmani, V., Maxhuni, A., Grünerbl, A., Lukowicz, P., Haring, C., & Mayora, O. . (2013). *Monitoring activity of patients with bipolar disorder using smart phones.* Paper presented at the In Proceedings of International Conference on Advances in Mobile Computing & Multimedia.

Plaza, I., Demarzo, M. M. P., Herrera-Mercadal, P., & García-Campayo, J. (2013). Mindfulness-based mobile applications: literature review and analysis of current features. *JMIR mHealth and uHealth, 1*(2), e24.

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets? *Nat Rev Drug Discov, 10*(9), 712. doi:10.1038/nrd3439-c1

Prociow, P., Wac, K., & Crowe, J. (2012a). Mobile psychiatry: Towards improving the care for bipolar disorder. *International Journal of Mental Health Systems, 6*. doi:10.1186/1752-4458-6-5

Prociow, P., Wac, K., & Crowe, J. (2012b). Mobile psychiatry: towards improving the care for bipolar disorder. *International journal of mental health systems, 6*(1), 5.

Saeb, S., Lattie, E. G., Schueller, S. M., Kording, K. P., & Mohr, D. C. (2016). The relationship between mobile phone location sensor data and depressive symptom severity. *PeerJ, 4*, e2537.

Saeb, S., Zhang, M., Karr, C. J., Schueller, S. M., Corden, M. E., Kording, K. P., & Mohr, D. C. (2015). Mobile Phone Sensor Correlates of Depressive Symptom Severity in Daily-Life Behavior: An Exploratory Study. *J Med Internet Res, 17*(7), e175. doi:10.2196/jmir.4273

Saeb, S., Zhang, M., Karr, C. J., Schueller, S. M., Corden, M. E., Kording, K. P., & Mohr, D. C. (2015). Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *Journal of medical Internet research, 17*(7), e175.

Saeb, S., Zhang, M., Kwasny, M., Karr, C. J., Kording, K., & Mohr, D. C. (2015). *The relationship between clinical, momentary, and sensor-based assessment of depression.* Paper presented at the Proceedings of the 9th International Conference on Pervasive Computing Technologies for Healthcare.

Saeed Abdullah, M. M., Ellen Frank, Gavin Doherty, Geri Gay, Tanzeem Choudhury. (2016). Automatic Detection of Social Rhythms in Bipolar Disorder. *J Am Med Inform Assoc, 23*(3), 538-543.

Seko, Y., Kidd, S., Wiljer, D., & McKenzie, K. (2014). Youth mental health interventions via mobile phones: a scoping review. *Cyberpsychology, Behavior, and Social Networking, 17*(9), 591-602.

Shen, N., Levitan, M.-J., Johnson, A., Bender, J. L., Hamilton-Page, M., Jadad, A. A. R., & Wiljer, D. (2015). Finding a depression app: a review and content analysis of the depression app marketplace. *JMIR mHealth and uHealth, 3*(1), e16.

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annu. Rev. Clin. Psychol., 4*, 1-32.

Težak, Ž., Kondratovich, M. V., & Mansfield, E. (2010). US FDA and personalized medicine: in vitro diagnostic regulatory perspective. *Personalized Medicine, 7*(5), 517-530.

Torous, J., & Powell, A. C. (2015). Current research and trends in the use of smartphone applications for mood disorders. *Internet Interventions, 2*(2), 169-173.

Wahle, F., Kowatsch, T., Fleisch, E., Rufer, M., & Weidt, S. (2016). Mobile sensing and support for people with depression: a pilot trial in the wild. *JMIR mHealth and uHealth, 4*(3).

Whiteford, H. A., Degenhardt, L., Rehm, J., Baxter, A. J., Ferrari, A. J., Erskine, H. E., . . . Johns, N. (2013). Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *The Lancet, 382*(9904), 1575-1586.

PRISMA Diagram

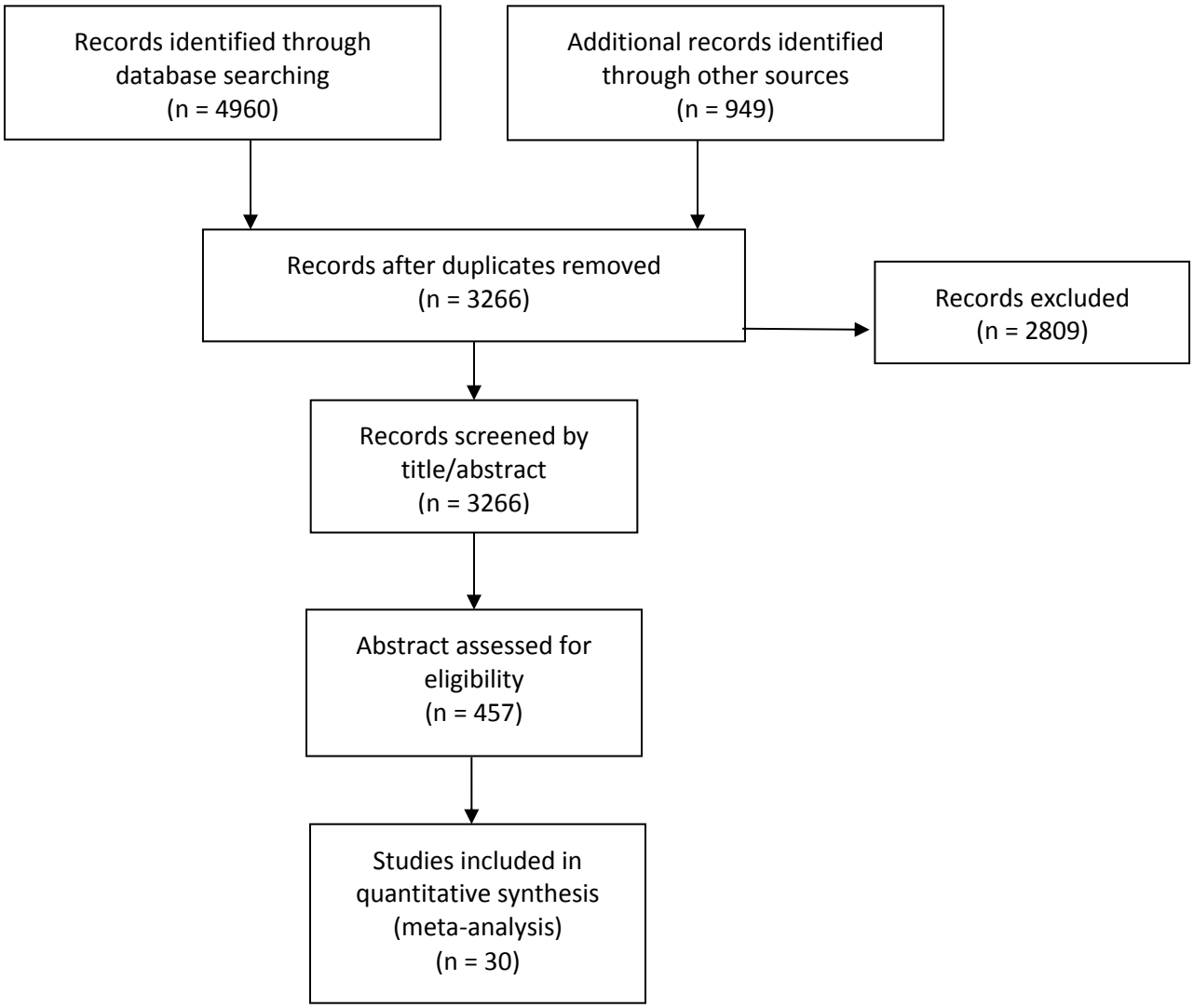

Table 1. Overview of Included Studies

Study, country, n	Adherence	Duration	App	Condition (diagnostic criteria)	Inpatient/outpatient/community	Incentive
Abdullah et al., 2016, U.S. (n=7)	NA	4 weeks	Mood Rhythm	Bipolar disorders (clinically diagnosed, criteria not specified)	Outpatient	$50 for each week of participation, $25 for each completed questionnaire and $50 for the final interview. Patient compensation was not contingent on adherence to the daily protocol of use.
Giudi et al., 2015, France (n=1)	NA	14 weeks	PSYCHE	Bipolar II (DSM-IV-TR)	Outpatient	NA
Grunerbl et al., 2015, Austria (n=10)	19 - 71 out of 84 datasets per patient per day	12 weeks	MONARCA	Bipolar (ICD 10)	Inpatient/ outpatient	NA
Karam et al., 2014, U.S.(n=6)	NA	6 months	PRIORI	Bipolar disorders (clinically diagnosed, criteria not specified)	Outpatient	NA
Maxhuni et al., 2016, Austria (n=5)	NA	12 weeks	MONARCA	Bipolar (ICD 10)	Inpatient/ outpatient	NA
Muaremi et al., 2014, Austria (n=12)	NA	12 weeks	MONARCA	Bipolar disorders (clinically diagnosed, critiera not specified)	Inpatient/ outpatient	NA
Osmani et al., 2015, Austria (n=12)	NA	12 weeks	MONARCA	Bipolar disorders (clinically diagnosed, critiera not specified)	Inpatient/outpatient	NA
Faurholt-Jepsen, et al., 2016a, Denmark (n=29)	NA	12 weeks	MONARCA	Bipolar (ICD 10)	Outpatient	No compensation
Dang et al., 2016, Germany (n=4)	NA	1 week	Fine	MDD (clinically diagnosed, criteria not specified)	Outpatient	NA
Beiwinkel et al., 2016, Germany (n=13)	55.7% (no clinician feedback); activity: 78.2%; social data	12 months	SIMBA	Bipolar (DSM-IV)	Outpatient	NA

	56.1%					
Ben-Zeev et al., 2015, US (n=47)	One-time self-assessment at week-10 was 78%	10 week	FOCUS	Healthy/MDD (NA)	Community	T-shirts and raffle of multiple Jawbone UP wristbands, and Google Nexus smartphone at various weeks
Faurholt-Jepsen, et al., 2014, Denmark (n=17)	Daily self-assessment (10-week) 87 (with clinician feedback)	3 months	MONARCA	Bipolar (ICD 10)	Outpatient	No compensation
Faurholt-Jepsen, et al., 2015, Denmark (n=61)	NA	6 months	MONARCA	Bipolar (ICD 10)	Outpatient	No compensation
Frost et al., 2013, Denmark, (n=6)	Daily self assessment 91% (with clinician feedback)	6 months	MONARCA 2.0	Bipolar disorders (clinically diagnosed, criteria not specified)	Outpatient	No compensation
Grunerbl et al., 2012, Austria (n=10)	NA	8weeks	MONARCA	Bipolar disorders (clinically diagnosed, criteria not specified)	Inpatient/ outpatient	Study phone
Grunerbl et al., 2014, Austria (n=12)	NA	12 weeks	MONARCA	Bipolar (ICD 10)	Inpatient/outpatient	NA
Osmani et al., 2013, Austria (n=9)	NA	3 months	MONARCA	Bipolar (ICD 10)	Inpatient/outpatient	NA
Prociow et al., 2012, UK (n=4 healthy, 1 bipolar)	NA	2 weeks	NA	Healthy/bipolar (self-identified)	Outpatient	NA
Saeb et al., 2015a, US (n=28)	NA	2 weeks	Purple Robot	Healthy/MDD (self-identified)	Community	NA
Asselbergs et al. 2016, Netherlands (n=27)	daily self-assessment 88.8%	6 weeks	iYouVU	Healthy/MDD (self-identified)	Community	EMA response rates ≥50%: €20; rates ≥75%: €35; rates ≥95%: €47.50

Braun et al., 2016, Germany and Spain (n=36)	NA	2 weeks	VoiceApp	Healthy/MDD (self-identified)	Community	NA
Canzian et al., 2016, UK, (n=28)	Mood Traces	2 weeks	MoodTraces	Healthy/MDD (self-identified)	Community	One winner of a Nexus 5 mobile phone and five winners that have received a 10 pounds Amazon voucher each among all the participants that have completed the daily questionnaire at least 50 times in a two-month span.
Farhan et al., 2016a, US (n=79)	NA	Undetermined	LifeRhythm	Healthy/MDD (self-identified/DSM-5)	Community	$15 Amazon gift card for every two weeks of active participation
Farhan et al., 2016b, US (n=60)	NA	10 weeks	StudentLife	Healthy/MDD (self-identified)	Community	NA
Faurholt-Jepsen, et al., 2016b, Denmark (n=28)	NA	12 weeks	MONARCA	Bipolar (ICD 10)	Outpatient	NA
Gideon, et al., 2016, US (n=37)	NA	6-12 months	PRIORI	Bipolar disorders (clinically diagnosed, criteria not specified)	Outpatient	NA
Hung et al., 2016, Taipei (n=28)	NA	2 weeks	iHOPE	Healthy (self-identified)	Community	NA
Saeb et al., 2016, US (n=48)	NA	10 weeks	StudentLife	Healthy/MDD (self-identified)	Community	NA
Saeb et al., 2015b, US (n=18)	NA	2 weeks	Purple Robot	Healthy/MDD (self-identified)	Community	$35 per week
Wahle et al., 2016, Switzerland and Germany (n=126)	22.2% provided 2 self assessment for 4 weeks or more	8 weeks	MOSS	Healthy/MDD (self-identified)	Community	NA

Table 2. Overview of Smartphone Applications

App	Platform	Passive Data Streams	Bipolar or MDD	Commercial	Personal vs. Study Phone
MONARCA	Android	Audio, communication, GPS and accelerometer	Bipolar	No	Personal or study phone
MONARCA 2.0	Android	Audio, communication, GPS, accelerometer, phone usage	Bipolar	No	NA
Mood Rhythm	iOS/Android	Audio, communication, GPS, accelerometer, Wi-fi, network	Bipolar	No	Study phone
PRIORI	Android	Audio	Bipolar	No	Study phone
PSYCHE	Android	Audio	Bipolar	No	NA
SIMBA	Android	GPS, communication, network	Bipolar	NA	Study phone
Fine	Android	Communication, GPS and accelerometer, phone usage	MDD		Study phone
FOCUS	Android	Audio, GPS, accelerometer, Wi-fi, light sensor	MDD	No	Study phone
iHOPE	Android	Communication and phone usage	MDD	No	Personal phone
iYouVU	Android	Communication, accelerometer, phone usage, phone camera log	MDD	No	Personal phone
LifeRhythm	iOS/Android	GPS, accelerometer, network	MDD	No	Personal phone
MoodTraces	Android	GPS	MDD	Yes	Personal phone
MOSS	Android	Communication, GPS, accelerometer, phone usage, schedule	MDD	No	NA
Purple Robot	Android	GPS, phone usage	MDD	Yes	Study phone
StudentLife	iOS/Android	Audio, communication, GPS and accelerometer, phone usage, light sensor	MDD	No	Personal or study phone
VoiceApp	Android	Audio	MDD	No	NA

Table 3. Mapping of Phenotypes on Smartphone Data Sources and DSM 5 criteria

Phenotypes	**Smartphone Data Sources**	**Symptoms of Major Depressive Episode**	**Symptoms of Manic Episode**
Social Activity	Microphone, GPS, Wi-Fi, communication log	Diminished interest in nearly all activities most of the day	Increased goal-directed activity; increased talkativeness
Mobility	GPS, accelerometer, microphone, Bluetooth, sensor for cell towers	Fatigue or decreased energy	Increased goal-directed activity; distractibility
Physical Activity	Accelerometer	Psychomotor agitation or retardation; fatigue or decreased energy	Increased psychomotor agitation
Sleep	Screen time, GPS, Wi-Fi, communication log, light sensors	Insomnia or hypersomnia	Decreased need for sleep
Voice	Microphone	Depressed mood	Persistent elevated, expansive, or irritable mood; increased talkativeness; flight of ideas or racing thoughts;
Other symptoms	NA	Significant change in weight or appetite; inappropriate guilt or feelings of worthlessness; difficulty concentrating or making decisions; recurrent thoughts of death, suicidal thoughts, plans, or attempts	Increased in risky behaviors; inflated self-esteem or grandiosity

Table 4. Summary of Smartphone Audio Data in Monitoring Mood Disorders

Study, country (n)	Duration	App	Data Source	Clinically Relevant Outcome	Methods	Audio Findings	Additional Findings
Braun et al., 2016, Germany and Spain (n=36)	2 weeks	VoiceApp	Speech recorded of text reading	HAMD-17	Unspecified correlation between HAMD score and audio data	r>0.80 between audio features and HAMD 65-75% of all cases	NA
Faurholt-Jepsen, et al., 2016b, Denmark (n=28)	12 weeks	MONARCA	Phone calls	YMRS and HAMD from biweekly clinical assessment	what methods were used to classify patients into one of the how many possible mood states ranging from X to Y	Classification accuracy of depressed or euthymic state was 0.70 (s.d. 0.13) with a sensitivity of 0.64 (s.d. 0.25), and the classification accuracy for a manic or mixed state versus a euthymic state was 0.61 (s.d. 0.04) with a sensitivity of 0.71 (s.d. 0.09).	Addition of other objective data did not improvement prediction accuracy of audio data alone
Gideon, et al., 2016, US (n=37)	6-12months	PRIORI	Phone calls with clinician and others	YMRS and HAMD from weekly clinical assessment by phone	Support vector machine with linear and radial-basis-function kernel for binary mood state classification	AUC of 0.72±0.20 for mania and AUC of 0.75±0.14 for depression	
Giudi et al., 2015, France (n=1)	14 weeks	PSYCHE	Picture commenting task	Voice segments recorded by a computer	Spearman correlation between mood and audio	Spearman correlation between mood state changes and median absolute deviation of the distribution (0.54, p-value = 0.0392); No statistically significant correlations between audio features and QID or YMRS scales	N/A

Grunerbl et al., 2015, Austria (n=10)	12 weeks	MONARCA	Phone calls	Psychiatric assessment and psychological state examination were performed every 3 weeks over a period of 12 weeks using HAMD and YMRS	Naive Bayes classifier and other classifiers, such as k-nearest neighbors, were used to classify patients into one of the 7 classes ranging from mania to depression.	Classification accuracy 70%	Classification accuracy using call data 66%
Karam et al., 2014, U.S.(n=6)	6 months	PRIORI	Phone calls with clinicians and others	Weekly clinical assessment using YMRS and HAMD by phone with trained clinicians of mood state over the past week	Support vector machine with linear and radial-basis-function kernel for binary mood state classification	AUC for audio-based classification of hypomania (depression): clinical evaluation calls trained with clinical evaluation calls 0.81 ± 0.17 (0.67 ± 0.18); unstructured calls trained with audio data from clinical evaluation calls on the same day 0.61 ± 0.09 (0.49 ± 0.08); unstructured calls trained with clinical evaluation calls the day before or after 0.47±0.05 (0.52 ± 0.09).	N/A
Maxhuni et al., 2016, Austria (n=5)	12 weeks	MONARCA	Phone calls	Psychiatric assessment and psychological state examination were performed every 3 weeks over a period of 12 weeks using HAMD and YMRS	Implementation of several classifiers (e.g., random forest, support vector machine, k-nearest neighbors) was used to classify patients into one of the 7 classes ranging from mania to depression.	Classification accuracy using spectral characteristics 82% or emotional characteristics 82%	Classification accuracy using accelerometer data 81% – 85%; accelerometer and audio data combined: 79%-86%

Muaremi et al., 2014, Austria (n=12)	12 weeks	MONARCA	Phone calls	Psychiatric assessment and psychological state examination were performed every 3 weeks over a period of 12 weeks using HAMD and YMRS	Support vector machine, logistic regression, random forest and neural networks were used to classify patients into one of the 7 classes ranging from mania to depression.	Classification accuracy using conversation characteristics 78% or patient voice characteristics 80%	Classification accuracy using call data 77%; combination of all data streams 83%
Osmani et al., 2015, Austria (n=12)	12 weeks	MONARCA	Phone calls	Psychiatric assessment and psychological state examination were performed every 3 weeks over a period of 12 weeks using HAMD and YMRS	Naïve Bayes classifier, k-nearest neighbors, search tree, and a conjunctive rule learner were used to detect change at the individual level.	Classification accuracy 70%	Classification accuracy using accelerometer data 72%; GPS 81%; combination of GPS and accelerometer 76%; call data 66%; combination of call and audio 69%

Table 5. Summary of Studies Using Smartphone Accelerometer Data

Study, country, n	App	Duration	Clinically Relevant Outcome	Methods	Physical Activity Findings
Abdullah et al., 2016, U.S. (n=7)	Mood Rhythm	4 weeks	Self-report on Social Rhythm Metric (5 items)	Support vector machine with linear kernel using recursive feature elimination	Non-sedentary duration ranked 3rd in predicting stable vs. unstable status
Grunerbl et al., 2015, Austria (n=10)	MONARCA	12 weeks	Psychiatric assessment and psychological state examination were performed every 3 weeks over a period of 12 weeks using HAMD and YMRS	Naive Bayes classifier and other classifiers, such as k-nearest neighbors, were used to classify patients into one of the 7 classes ranging from mania to depression.	Classification accuracy 70%
Maxhuni et al., 2016, Austria (n=5)	MONARCA	12 weeks	Psychiatric assessment and psychological state examination were performed every 3 weeks over a period of 12 weeks using HAMD and YMRS	Implementation of several classifiers (e.g., random forest, support vector machine, k-nearest neighbors) was used to classify patients into one of the 7 classes ranging from mania to depression.	Classification accuracy using time domain 80.78% or frequency domain 84.54%
Osmani et al., 2015, Austria (n=12)	MONARCA	12 weeks	Psychiatric assessment and psychological state examination were performed every 3 weeks over a period of 12 weeks using HAMD and YMRS	Naïve Bayes classifier, k-nearest neighbors, search tree, and a conjunctive rule learner were used to detect change at the individual level.	Classification accuracy 72%
Ben-Zeev et al., 2015, US (n=47)	FOCUS	10 week	Self-report on PHQ9 at the end of study period	Penalized functional regression of active period on depression	The association was small and non-significant (b= 0.00031; p=0.61)

Frost et al., 2013, Denmark, (n=6)	MONARCA 2.0	6 months	Self-report on mood from highly depressed (-3) to highly manic (+3) daily	chi-square correlations between objective data (phone usage, social activity, physical activity, and mobility) self-reported mood score	Physical activity was not highly correlated with mood scores
Grunerbl et al., 2012, Austria (n=10)	MONARCA	8weeks	Daily self-report on activity of daily life and psychiatric assessment every 3 weeks using HAMD, ADS, and MSS	Descriptive statistics of motion ratio in manic, euthymic, depressed states; linear regression of motion ratio on mood states	Mean increase of 21.3% motion ratio in the transition from depression to euthymic state and a reduction of 33.7% motion ratio in the transition from mania to euthymic state; motion ratio correlated with self-assessment within 90% confidence interval in bipolar patients in a linear regression.
Grunerbl et al., 2014, Austria (n=12)	MONARCA	12 weeks	Psychiatric assessment every 3 weeks over a period of 12 weeks using HAMD, ADS, and MSS	Implementation of several classifiers (e.g., Na¨ive Bayes, k-nearest neighbor, j48 search tree, conjunctive rule learner) were used to classify patients into one of the 7 classes ranging from mania to depression.	Classification accuracy 72%
Osmani et al., 2013, Austria (n=9)	MONARCA	3 months	Psychiatric assessment and psychological state examination were performed every 3 weeks over a period of 3 months using HAMD and YMRS	Pearson correlation coefficient between physical activity levels (none, moderate, high) during each daily interval (overalll, morning, afternoon, evening, respectively) and psychiatric evaluation scores	Much stronger correlation between the individual daily intervals than there is for the overall activity levels

Asselbergs et al. 2016, Netherlands (n=27)	iYouVU	6 weeks	Self-report on mood from low (-2) to high (+2) 5 times a day	Personalized mood prediction models were trained using forward stepwise regression (FSR) of phone usage, communication log and percentage of duration performing "high activity"	correct cross-validated predictions of the personalized models 55% to 76%.
Farhan et al., 2016a, US (n=79)	LifeRhythm	undetermined	DSM5 based clinical assessment at the initial screening and self-report on PHQ9	Pearson's correlation between percentage of time being active or percentage of time being inactive and PHQ9 score	Non-significant correlation (correlation and p-value were not reported)
Farhan et al., 2016b, US (n=60)	StudentLife	10 weeks	self-report on PHQ9	Support vector machine to classify people into clusters based on daily averages and trends of physical activity, light information, conversation, and location data, as well as variability in location	People in the low PHQ9 group were more active as compared to people in the high PHQ9 group
Wahle et al., 2016, Switzerland and Germany (n=126)	MOSS	8 weeks	Binary outcome based on biweekly self-report on PHQ9 using a cut-off of 11	Random Forest and Support Vector Machine leave-one-out cross validation using Wi-fi, accelerometer, GPS and phone usage data	Classification accuracy: 59.1% - 60.1%