From smartphone data to clinically relevant predictions: A systematic review of digital phenotyping methods for depression prediction

Imogen E. Leaning[*†a,b], imogen.leaning@donders.ru.nl
Nessa Ikani[*c], n.ikani@tilburguniversity.edu
Hannah S. Savage[a,b], hannah.savage@donders.ru.nl
Alex Leow[d], AlexLeow@Alumni.ucla.edu
Christian Beckmann[a,b], christian.beckmann@donders.ru.nl
Henricus G. Ruhé[a,e], eric.ruhe@radboudumc.nl
Andre F. Marquand[a,b,f], andre.marquand@donders.ru.nl

*These authors contributed equally to this work.
[†]Corresponding author.

[a]Donders Institute for Brain, Cognition and Behaviour Radboud University Nijmegen, Nijmegen, the Netherlands
[b]Department for Cognitive Neuroscience, Radboud University Medical Center Nijmegen, Nijmegen, the Netherlands
[c]Department of Developmental Psychology, Tilburg School of Social and Behavioral Sciences, Tilburg University, Tilburg, the Netherlands
[d]Department of Psychiatry, Department of Biomedical Engineering and Department of Computer Science, University of Illinois Chicago, Chicago, United States of America
[e]Department of Psychiatry, Radboud University Medical Center Nijmegen, Nijmegen, the Netherlands
[f]Department of Neuroimaging, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

Declaration Of Interest:
Alex Leow is on the advisory board for Buoy Health and is a cofounder of KeyWise.
Christian Beckmann is a director of SBGNeuro.

Abstract

Smartphone-based digital phenotyping enables potentially clinically relevant information to be collected as individuals go about their day. This could improve monitoring and interventions for people with Major Depressive Disorder (MDD). The aim of this systematic review was to investigate current digital phenotyping methods used in MDD. We searched PubMed, PsycINFO, Embase (20/07/2022), Scopus (21/07/2022) and Web of Science (22/07/2022) for articles including: (1) MDD population, (2) smartphone-based features, (3) validated ratings. Risk of bias was assessed using criteria from several sources. Studies were compared within prediction goals. Fourteen studies (3249 participants) were included, covering different goals (e.g., predicting symptom severity). Studies achieved moderate model performance. Common themes included challenges from missing and complex data, and a lack of external testing sets. There is a risk of bias from missing data. Studies made progress towards relating digital phenotypes to clinical variables, often focusing on time-averaged features. Methods investigating temporal dynamics more directly may be beneficial for patient monitoring.

European Research Council consolidator grant: 101001118, Prospero registration: CRD42022346264, Open Science Framework: https://osf.io/s7ay4

Keywords: digital phenotyping, Major Depressive Disorder, smartphone

Introduction

Major depressive disorder (MDD) is one of the most common and debilitating mental disorders worldwide, associated with a high personal and societal burden (Lim, et al., 2012). Moreover, MDD is often linked to a high recurrence risk (Buckman, et al., 2018), with over half of people who experience one depressive episode going on to have a subsequent episode (Burcusa & Iacono, 2007). Importantly, early signs of the development of symptoms or recurrence of depression are often not identified, which impedes timely preventive strategies. The broad integration of smartphones into people's daily lives provides the unique opportunity to continuously and unobtrusively record behavioural dynamics in a naturalistic setting with high temporal resolution (Nelson & Allen, 2018). As such, it can offer insights into an individual's mental state and could be useful for symptom monitoring and just-in-time preventive efforts in both non-clinical and clinical contexts (e.g., predicting symptom onset, or future recurrent episodes in patients with Major Depressive Disorder (MDD)).

Developing tools that leverage smartphone data to its full potential may therefore enable earlier identification and intervention before worsening of symptoms or recurrence of depression, leading ultimately to better outcomes. Smartphones can collect a wide range of behavioural information, for example geolocation data derived from the Global Positioning System (GPS), an individual's use of social media or communication apps, general phone use/screen time, and typing-related data measuring psychomotor functioning or processing speed (Harari, Müller, Aung, & Rentfrow, 2017). Other technologies, such as wearable devices (e.g., wristbands), act as additional digital sources of behavioural or psychophysical measures. All these types of data can be used to create digital phenotypes, i.e., markers of behaviour or physiology calculated from digital measures, which could be indicative of clinically-relevant behaviours. In this review we will focus on digital phenotypes for MDD created using smartphones, as these devices are now an integral and ubiquitous part of our daily lives. By

installing monitoring tools on an individual's own device, greater ecological validity may therefore be achieved than by using other devices as the risk of the monitoring altering participants' behaviour may be lower than in studies where participants are required to adapt to wearing a device that they are not already accustomed to.

Digital phenotyping is a rapidly expanding technique, and a variety of different features have been explored in combination with various methods for classifying clinical labels or predicting clinically relevant information (e.g., depression scores). For example, Saeb, et al. (2015) carried out an initial exploration of possible depression-related GPS and phone usage features, which were then used to classify participants with and without depressive symptoms from the general population using a logistic regression classifier, and to predict Patient Health Questionnaire (PHQ) scores using a linear regression model. Relatedly, Farhan, et al. (2016) also predicted clinical diagnoses and questionnaire (PHQ-9) scores, using a support vector machine (SVM) for the former objective and support vector multivariate regression methods for the latter. Ware, et al. (2020) went beyond classifying clinical labels, and instead classified the presence or absence of specific symptoms in participants using SVM models. On a more general level, Müller, Chen, Peters, Chaintreau, and Matz (2021) classified participants as being either healthy or depressed using solely GPS data and penalized logistic regression, random forest and XGBoost models. These various studies have used a variety of methods (i.e., different modelling approaches, populations, features, study durations), to achieve a similar range of prediction goals.

In order to understand how digital phenotyping can be used to better understand behavioural dynamics underlying MDD and to advance precision medicine endeavours aimed at earlier identification and/or intervention of (recurrent) depressive symptoms, high model performance (e.g., low errors for regression models, and high sensitivity and specificity for classification models) is needed in addition to validation across multiple settings, including

various symptom severities and lifestyles (e.g., working vs non-working populations). The general aim of this review was, therefore, to investigate the current state of digital phenotyping research for populations with MDD, in particular to establish what current methods are able to achieve in terms of their predictive power, and where subsequent efforts need to be focused to advance digital phenotyping in depression. Specifically, this systematic review aims to answer the following questions:

1. Of the different features that have been constructed from smartphone data, which are the most predictive of clinically relevant variables in the context of MDD?
2. What are the different methods that have been used for various depression prediction tasks using smartphone data and to what extent have these methods been successful?

First, we provide an overview and general evaluation of smartphone features (constructed across the included studies) that have been correlated with clinically relevant variables for depression (e.g., self-reported symptom scores), as feature construction is important for successful prediction models. The correlation score was used as a tentative measure of clinical relevance. However, even a weakly correlated feature could be clinically relevant when used, for example, in conjunction with other features that provide more contextual information. Second, digital phenotyping studies in the field of MDD were compared that cover a variety of prediction tasks. In doing so, studies were grouped by prediction goal (i.e., studies that predict depression symptom severity from smartphone features (or vice versa); studies that predict clinical vs non-clinical labels or states (e.g., depressive state); studies that predict a state at future time points; and other prediction goals). Within each group of studies, the various methods used were compared, and assessed in terms of their benefits and limitations. Available metrics (e.g., classification accuracy, root mean squared error (RMSE)) were compared

between studies against the backdrop of factors such as study population and included features, where informative, to assist meaningful comparisons.

Method

Protocol and registration

This systematic review was guided by a protocol registered on Prospero (CRD42022346264) and the Open Science Framework (https://osf.io/s7ay4) and reported in line with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.

Information sources and search strategy

A comprehensive search was conducted on the following electronic databases: PubMed (July 20, 2022), PsycINFO (July 20, 2022), Embase (July 20, 2022), Scopus (July 21, 2022) and Web of Science (July 22, 2022). This search was restricted to studies published between January 2012 and July 2022 and included keywords related to (1) MDD, (2) digital phenotyping or monitoring, and (3) smartphones. For an overview of the exact keywords and search strings see S1 in supplementary materials.

Selection of studies and eligibility criteria

Two authors (IL and NI) independently screened all titles and abstracts, and selective full texts screenings, to identify eligible papers for inclusion. Full texts of the selected papers were then examined to determine the final selection. In case of disagreement on inclusion, a third author (AM) was consulted to resolve divergent assessments. Central issues were discussed with all authors.

Studies were considered eligible if (1) passively collected smartphone data was utilised (e.g., GPS, use of communication apps, nearby Bluetooth devices) obtained with Android or iPhone smartphones, (2) passive data was collected over the course of the participants' everyday lives (i.e., not collected during a laboratory session), and if (3) passively collected

smartphone data was related to measures assessing depressive symptoms and/or diagnostic MDD status (e.g., self-reports, Ecological Momentary Assessments (EMA), structured clinical interviews) for the purpose of external validation of the digital phenotypes. Studies combining passively collected smartphone data with other data types, such as data from other wearable devices, were also included. Studies were excluded if (1) data was collected solely through means other than smartphones (e.g., wearable devices such as smart wristbands); (2) studies did not include participants with MDD (e.g., studies that included participants without a formal clinical diagnosis); (3) digital phenotyping-related studies with objectives not listed above (e.g., data collection verification studies); (4) reviews, overview articles, commentaries, etc.

Data extraction, risk of bias assessment and quality assessment

Two authors (IL and NI) extracted data regarding study context, study sample, prediction goals, data acquisition, paradigms and analysis methods. For an overview of the exact data that was extracted see data extraction form S2 in supplementary materials.

Each included study was independently assessed by authors (IL and NI) for risk of bias using the criteria proposed by the Cochrane Collaboration Risk of Bias (RoB; Higgins, et al., 2016) and discrepancies were resolved with a third author (AM) (see Table S3 in supplementary materials). Five domains were rated as high risk, some concerns, low risk or unclear risk if there was risk of bias due to: (a) the used method for the randomisation sequence (selection bias); (b) allocation concealment (allocation bias); (c) blinding of participants and researchers (performance bias); (d) blinding of outcome assessment (detection bias); (e) incomplete outcome data (up to 10% drop out was rated as low risk) (attrition bias); (f) selective reporting (reporting bias). Items that were not relevant for a study were marked as ‘NA’.

Each included study was independently assessed by authors (IL and NI) for quality. Quality Assessment (QA) was assessed using items adapted from the guidelines created by Luo, et al. (2016), as well as from Benoit, Onyeaka, Keshavan, and Torous (2020) (see Table S4 in supplementary materials). The guidelines provided by Luo, et al. (2016) relate to machine learning methods, therefore, not all items were relevant/applicable for each included study. In these cases, relevant items were assessed and others listed as 'NA'.

Outcome measures

Outcomes included: (1) correlations between passively-collected smartphone data and clinical measures; (2) type of prediction strategy used to predict clinical labels or symptoms, and measures of performance of the prediction strategy. In addition, participant (sample type) and study information relevant for RoB assessment and QA was extracted.

Results

The search queries returned 14 eligible studies with several analysis and prediction goals (see Figure 1 for PRISMA flow chart. Examples of exclusion based on study design included studies that utilised solely digital phenotypes calculated from digital devices other than smartphones, or that used digital phenotypes calculated from actively collected smartphone data. Examples of exclusion based on publication type included published protocols, reviews and dissertations). Characteristics of included studies are provided in Table 1 and general methodological information for these studies is summarised in Table 2. Specific methodological information per analysis goal is summarised in Tables 3-6. Included studies used features calculated from a range of sensor streams, for example count-based and statistical features reflecting mobility sensors (e.g., GPS data) and communication sensors (e.g., app-based data). Studies were grouped by analysis goal to allow for comparison of methods with similar objectives. To this end, we first compared studies that correlated passive smartphone features with depression symptom severity (Cao, et al., 2020; Sverdlov, et al., 2021; Zhang, et al., 2022), and then investigated the methods used for predicting symptom severity (Braund, et al., 2022; Cao, et al., 2020; Faurholt-Jepsen, et al., 2022; Pedrelli, et al., 2020; Pellegrini, et al., 2022; Sverdlov, et al., 2021; Zhang, et al., 2021; Zhang, et al., 2022). Somewhat unexpectedly, two studies aimed to predict specific smartphone features from ratings of depression (Laiou, et al., 2022; Tønning, Faurholt-Jepsen, Frost, Bardram, & Kessing, 2021). We then shift our focus towards studies that aimed to classify participants into different diagnostic classes and mood states (Bai, et al., 2021; Cho, et al., 2019; Faurholt-Jepsen, et al., 2022; Sverdlov, et al., 2021). Some studies contained unique goals that were not shared with the other studies, and these goals are considered in a separate section. We also compared some key methodological choices, such as feature selection and processing, dimension reduction, and handling of missing data.

Correlation between passive smartphone features and depressive symptom severity

Three studies were identified that investigated correlations between features derived from passively-collected smartphone data (e.g., total amount of time spent at home, number of unique phone call partners/day, total number and duration of phone calls) and depressive symptom severity quantified using sum scores of self-reports, such as the PHQ, and observer-rated Montgomery–Åsberg Depression Rating Scale (MADRS). These studies are summarised in Table 3. Sverdlov, et al. (2021) and Cao, et al. (2020) investigated various features related to participants' communication behaviours. Sverdlov, et al. (2021) found that more severe depressive symptom scores tended to have lower entropy of usage time of communication apps, lower total count of communication apps usage, and lower WhatsApp usage. Cao, et al. (2020) found that a higher depression score is significantly correlated with lower social interaction levels (i.e., shorter phone call durations, fewer text messages sent), which seems consistent with Sverdlov, et al.'s (2021) findings from app-based investigations, despite the differences in data type. However, no significant correlation between symptom scores and ambient light intensity or smartphone screen usage was found (Cao, et al., 2020).

Various mobility features (e.g., location variance, number of unique places visited) have also been related to individuals' symptom scores. Sverdlov, et al. (2021) found that participants with higher symptom scores tended to maintain a lower average distance from home than participants with less severe symptom scores. Cao, et al. (2020) also found that individuals with higher symptom scores demonstrated lower mobility, as indicated by decreased step count, fewer places visited, and lower location variance, spending their time more uniformly across different places (indicated by higher normalized entropy). Zhang, et al. (2022) investigated correlations between mobility features and symptom scores provided by their vector autoregressive model. They found that within individuals, the proportion of time spent at their home location/s ('homestay') and short-term rhythm (i.e., behavioural rhythms

with frequency higher than one day, e.g., for many people, going to and from their home) were positively correlated with symptom scores. Other features, for example long-term rhythm (behavioural rhythms with frequency less than one day, e.g., a weekly grocery shop) and circadian rhythm, were negatively correlated with symptom scores. Between individuals, less features were found to be significantly correlated, specifically, only location variance and moving distance were negatively correlated with symptom scores. Overall studies generally identified that higher symptom scores were associated with lower mobility.

Regarding the significance of the results and correction for multiple testing, Zhang, et al. (2022) used the commonly chosen adjusted significance value of $p < .05$, whereas Cao, et al. (2020) listed various p-values ranging between $p < .01$ and $p < .1$. It is unclear whether the thresholds used by Cao, et al. (2020) were corrected for multiple testing. Overall, correlations were generally weak to moderate, with Cao, et al. (2020) reporting the strongest correlation magnitude of approximately 0.65 for daily step count.

Predicting depression symptom severity

Several studies investigated the possibility of using passively-collected smartphone data to predict depression symptom severity (displayed in Table 4). Linear regression and linear mixed-effect regression models were popular choices for this prediction goal. Sverdlov, et al. (2021) used the communication and mobility features discussed in the preceding section to predict depressive symptom sum scores using a multiple linear regression model. To select a subset of features for this model from their entire communication and mobility feature set, a stepwise variable selection method with a significance threshold of $p < 0.1$ was used. Sverdlov, et al. (2021) was the only study in this section of the review to mention their model selection method. The correlation between observed and predicted scores calculated in leave-one-out cross-validation was $r = 0.43$, showing a moderate correlation. Pellegrini, et al. (2022)

conducted a Principal Component Analysis on a set of passive smartphone features they created based on weekly summaries of GPS and accelerometer measures, and used the first principal component as a predictor in their linear mixed models. Pellegrini, et al. (2022) investigated various models with and without this passive smartphone feature and a baseline depressive symptom score, demonstrating that including a smartphone feature did not improve the prediction of depressive symptom scores, but instead was comparable to predictions by models using only questionnaire data. The model combining smartphone features and questionnaire data achieved a RMSE of 4.30, and the questionnaire-only model achieved an RMSE of 4.27 (with the possible MADRS score ranging from 0-60). It is unclear whether this difference is statistically significant.

Applying (penalised) linear regression models to passive smartphone data may help ensure that the models are less likely to overfit to the data, however, the relationship between depression symptom severity and smartphone features may be non-linear. Because of this, several papers chose to investigate predictions of symptom severity by non-linear regression models, often comparing these results to linear regression models. In a study investigating depression in an adolescent population, Cao, et al., (2020) used linear regression and support vector regression with a polynomial kernel to predict symptom severity from smartphone data and personal and parental ratings. The most accurate model was a support vector regression model combining all three of these data types (RMSE = 2.65, with the PHQ-9 range being 0-27. The range in RMSE values for the remaining support vector regression models was between 3.23 to 6.41). Interestingly, the most accurate models using only smartphone data were linear models (RMSE = 2.77. The range in RMSE values for the remaining linear regression models was between 3.23 to 10.78). Zhang, et al. (2021) used pairwise linear mixed-effect models to explore the relationship between passively-collected smartphone features and symptom severity in a cross-country study in populations with a recent history of depression. In doing

so, they focused on Bluetooth features calculated over a period of two weeks, including features related to second-order statistics (e.g., the average value of the daily maximum number of nearby Bluetooth device count (NBDC)), multiscale entropy and the frequency domain. These features were compared to the PHQ-8 score reflecting the same two-week interval. Ten of their second-order statistical features were negatively associated with depression symptom scores. Four features related to daily variance of NBDC were negatively associated with depression score. The multiscale entropy at five scales was associated with symptom severity (with multiscale entropy at short timescales indicating sequence complexity, and at longer timescales indicating periodicity), and five frequency domain features were also associated with severity. The models containing Bluetooth features provided better fits to the data than a model containing no Bluetooth features. In general, it was found that for increases in depression symptom severity score, the variance and periodicity of the smartphone features sequence decreased, and it became more irregular. Zhang, et al. (2021) also investigated hierarchical Bayesian linear regression, LASSO regression and XGBoost regression models to predict symptom severity. Their hierarchical Bayesian linear regression model achieved the best performance in terms of their selected metrics for the two different cross-validation scenarios used (RMSE = 3.89, 4.426, with the PHQ-8 range being 0-24), although performance was similar to the XGBoost model for one of the scenarios. Given the possible PHQ-8 range, depression severity predictions were good but not exceptionally accurate.

Zhang, et al. (2022) investigated relationships between smartphone features and depression score using vector autoregressive models. They considered (cross-)lagged effects between each time point and the subsequent time point occurring two weeks later. Residential location count was positively correlated with later depression scores ($\varphi = 0.05$), despite a negative correlation being found between this feature and depression score at the within-subjects level. Moreover, depressive symptom scores were shown to be negatively correlated

with later circadian rhythm ($\varphi = -0.07$), and preceding location entropy ($\varphi = -0.04$) and homestay ($\varphi = 0.09$). Importantly, Zhang, et al. (2022) also demonstrated individual differences in cross-lagged effects related to age and circadian rhythm.

Pedrelli, et al. (2020) aimed to predict depressive symptom severity using average ensemble of boosting and random forest (AdaBoost) models. These models included a combination of smartphone and wearable features, as well as weather-related features. Smartphone features covered measures of location and movement (using GPS, Wi-Fi and Cell tower signal), app/smartphone usage, and calls/SMS. Kernel Principal Component Analysis was used to reduce the dimensionality of this feature set from 877 to 25 features, and was also carried out separately for models with only smartphone or wearable features. This study achieved similar performance across models, with the lowest error in a time-split cross-validation scenario from the model including only mobile features (RMSE = 4.88, HDRS range of 0-52), and the lowest error in a user-split cross-validation scenario from the model including only wearable features (RMSE = 5.35). Comparing models in both validation scenarios, the correlation between observed and predicted depression symptom scores in the time-split model including smartphone features was the strongest. The machine learning models outperformed predictions made using group median baseline and individual screen baseline models, but not predictions using individual median HDRS scores. Moreover, feature ranking was applied to investigate which features were most useful in predicting depression severity, highlighting features related to phone engagement, activity level and wearable features as the most important.

Braund, et al. (2022) investigated participants with both bipolar disorder and MDD, using linear regression models to test the association between circadian rhythm and depressive symptoms, measured using PHQ-9, and mixed-effects linear models to investigate potential moderating effects of circadian rhythm on symptom prediction across six timepoints covering

a ten week period. The diagnostic groups were pooled for the linear regression models after results indicated that there were no significant interactions between circadian rhythm and diagnosis. For the mixed-effects regression models, groups were not pooled. Circadian rhythm was not found to be associated with depression severity and similarly, no interactions were found between time point, circadian rhythm or diagnosis, or time point and circadian rhythm for depression severity. However, there was quite low variability in depression symptom severity indicated graphically in the study (given that the PHQ-9 range is between 0 and 27), and so strong interactions may be difficult to detect.

Faurholt-Jepsen, et al. (2022) used two-level mixed effects regression models (first level: repeated measurements per patient; second level: between-subjects variation) to investigate differences in mobility patterns (quantified using GPS, Wi-Fi and cell tower signals) between participants with bipolar disorder and unipolar depression. During depressive states, participants with unipolar depression were found to cover a significantly larger area per day, and had a larger total distance and duration of moves per day compared to participants with bipolar disorder. Overall and during euthymic states, participants with unipolar depression were found to have greater location entropy during the daytime than participants with bipolar disorder.

Predicting passive smartphone features

Whilst many studies investigated whether depression symptom severity could be predicted by features derived from passively collected smartphone data, two studies were identified that sought to make the inverse prediction (i.e., predicting different smartphone features from measures of depressive symptoms). Choice of smartphone-based response variables were informed by previous research and clinical knowledge. These studies are displayed in Table 4, alongside the studies in the preceding section. Laiou, et al. (2022) used a linear regression

model to predict homestay based on depression symptom severity measured by PHQ-8 scores, also including age, gender, occupational status, median completeness and sampling constancy in their model. They found that high depression symptom severity was associated with longer home stay during the overall study period and for weekdays only, but not for weekends. Laiou, et al. (2022) investigated the impact of individual differences in age, gender and occupational status on symptom severity, finding that older individuals tended to spend more time at home. This was also the case for unemployed relative to employed individuals. Gender was not found to have a significant effect on homestay.

Tønning, Faurholt-Jepsen, Frost, Bardram, and Kessing (2021) investigated the prediction of several smartphone features, including daily averages of physical activity (number of steps, total distance moved), smartphone usage (total screen-on time, number of times screen was turned on) and social activity (number of incoming, outgoing, missed calls, duration of calls, number of incoming & outgoing text messages). In doing so, various predictors were investigated, including symptom severity scores from clinical ratings, and from patient-reported data (daily smartphone-based patient self-reports of symptoms). Using linear mixed-effects models to account for repeated measurements (random effects) within each participant, it was found that more severe depression scores were significantly associated with fewer screen turn ons, larger number of outgoing calls, and longer phone call durations. However, it was noted that this was a small number of significant results, especially given the high risk of chance findings, as multiple testing was not accounted for. Unlike Zhang, et al. (2022), Tønning, Faurholt-Jepsen, Frost, Bardram, and Kessing (2021) did not find a significant relationship between distance moved and symptom severity. Tønning, Faurholt-Jepsen, Frost, Bardram, and Kessing (2021) also investigated the relationship between the smartphone features and a mood score that patients provided via their smartphones. Lower smartphone-reported mood was associated with increased social activity and phone usage

indicated by screen time, call duration, number of incoming/missed calls and number of incoming/outgoing text messages. In this context, incoming communication was suggested to be increased due to concern from external sources.

Predicting diagnostic class

Studies with classification-related goals are displayed in Table 5. Sverdlov, et al. (2021) investigated two regression methods to classify participants as depressed or healthy. The first was a logistic regression method that utilised input variables (e.g., number of unique places visited, average distance from home, total number of WhatsApp calls, total usage count of apps) selected in a stepwise manner, and for the second method they applied a clinically-determined threshold to MADRS scores predicted by a multiple linear regression model (again using selected input variables) to split the participants into the two classes. The latter model achieved higher accuracy, sensitivity and area under the receiver operating characteristic curve (AUC) than the logistic regression model, and comparable specificity.

Faurholt-Jepsen, et al. (2022) classified participants with bipolar disorder and unipolar depression into overall diagnostic classes, as well as subclasses related to state (depressive state or euthymic state), using an ensemble of decision trees. With regards to overall diagnostic label classification, the models achieved a sensitivity of 0.70, specificity of 0.65, and AUC of 0.75 during cross-validation. In line with this, Faurholt-Jepsen, et al. (2022) aimed to classify solely the depressive periods of participants with bipolar disorder or unipolar depression, to investigate whether the depressive state of participants in the two diagnostic groups can be differentiated, achieving again a sensitivity of 0.70, a higher specificity of 0.77, and AUC of 0.79. Overall, it can be seen that these models could differentiate between classes with moderate success, bearing in mind that a random binary classifier would achieve about 50% accuracy.

Predicting mood state/episode label

Rather than focusing on diagnostic labels, many studies chose to focus on mood states (these studies are also included in Table 5). This approach may be a useful step towards predicting clinically relevant changes in state for those who already have a diagnosis, or to predict relapse for those in remission. Bai, et al. (2021) aimed to classify participants with MDD into two groups, steady state and mood swing state, as well as four subgroups (steady state: in remission, currently depressed; mood swing: drastic (i.e., difference between maximum and minimum PHQ-9 scores is greater than or equal to ten), moderate (i.e., difference in scores is greater than or equal to five)), using a variety of machine learning methods (i.e., support vector machines, K-nearest neighbours, decision trees, naïve Bayes, random forest and logistic regression). Statistical smartphone features were calculated for the different types of phone calls and times of call, call duration, number of people involved in the calls and the entropy of callers. Some features from wearable devices (e.g., step count, heart rate) were also used. The success rate of classification was found to vary depending on which classes and features were included in each binary classifier, with classification between "steady-remission" and "swing-moderate" achieving the highest accuracy (0.8092) in a model using only the smartphone features. With this same feature set, classification between "steady-depressed" and "swing-drastic" achieved the lowest accuracy (0.6618). Amongst their models with various feature subsets, classification accuracies between "steady-remission" and "mood swing" (drastic and moderate) were higher than accuracies between "steady-depressed" and "mood swing" (drastic and moderate). The authors speculated that patients with continuous depressive symptoms may exhibit similar behaviours to those experiencing mood swings. The set of features that generally gave models with the highest accuracies contained features related to call logs, sleep data, step count data, and heart rate data. The models using features related to smartphone data generally achieved the lowest accuracies. Although Bai et al. (2021) chose to focus on binary classifiers, a multi-

class classifier may have more clinical utility as in the real world the class label would not yet be known, therefore it would be uncertain which binary classifier would be appropriate.

Random forest models were used by Cho, et al. (2019) to classify mood state as biased or neutral, and mood episode as depressive, manic, hypomanic, or no episode. Behavioural patterns from participants with bipolar disorder were also investigated. Cho, et al. (2019) used statistical features related to light exposure during bedtime and daytime, and many Fitbit features. The mood state prediction and mood episode prediction models achieved higher accuracy than a random classifier. For the mood state prediction of patients with MDD, and cut-off values of 10, 30 and 50% of the absolute mood score, the accuracy ranged from 0.61-0.67, sensitivity from 0.39-0.61, specificity from 0.42-0.74, and AUC from 0.56-0.69, respectively. There was no cut-off value that clearly achieved the best performance. For the mood episode prediction of patients with MDD, the accuracy was 0.751 and 0.712 for "No Episode" and "Depressive Episode" respectively. Moreover, sensitivity was 0.935 and 0.409, specificity was 0.395 and 0.878, and the AUC value was 0.781 and 0.798, demonstrating good but slightly inconsistent performance between the depressive episode and no episode classes. Cho, et al. (2019) also investigated individual models of mood state and episode classification, finding that the personalised mood state model outperformed the general model in all cases. For mood episode prediction, the personalised model achieved better performance in almost all cases.

Future time point prediction

One of the models reported in Cho, et al. (2019) involved making predictions related to future time points. That is, Cho, et al. (2019) used a classifier to predict mood states three days following the window covered by the data collection of passive smartphone features, using a mixture of smartphone-derived light exposure features and wearable features. The number of

days used to test the model (i.e., the three days) and the number of days used to train the model (18 days) were selected during parameter tuning, with longer periods (up to 300 days for training days and 30 days for testing) also being investigated. However, the shorter period of three days was found to be a more reasonable window for mood state prediction than longer periods during their parameter selection process, suggesting that later mood state prediction was difficult. This method differed from other studies that aimed to make predictions for mood at the end of smartphone data collection (e.g., Pellegrini, et al., 2022), and gives an example of how digital phenotyping research can shift towards predicting upcoming depression states.

Other goals

A few studies investigated digital phenotyping in participants with MDD, but could not be categorised into one of the above groups. For completeness, these studies are summarised in Table 6. Emden, Goltermann, Dannlowski, Hahn, and Opel (2021) investigated differences in study participation or retention between various diagnostic groups (affective, anxiety, and psychotic disorder groups and healthy controls), which yielded no significant differences between groups. Focusing on participants with MDD, Matcham, et al. (2022) investigated whether depressed mood was associated with data availability, not identifying a difference in data availability between those with no or mild depressive symptoms and more severe symptoms. Braund, et al. (2022) investigated differences in circadian rhythm between participants with MDD and participants with bipolar disorder, also not identifying a difference between groups. Moreover, Zhang, et al. (2022) investigated the relationship between mobility and depression severity using a dynamic structural equation model, however, the model failed to converge.

Comparison of Study Methods

Feature construction

Due to the vast range of sensors on smartphones, there are many different options of feature sets that are available or chosen for digital phenotyping. Sensors used by the studies identified in this review included, for example, GPS, light, steps, app data, smartphone on-off status, Wi-Fi and Bluetooth. Due to our focus on passive smartphone data, it was unlikely that any microphone sensors would be included here as these usually involve active participation of the participant to make recordings. The sensor data can be processed in many ways to create features, for example count features, statistical features and frequency-based measures. The number of studies using the various feature types are displayed in Figure 2.

Several different processing steps were used in the calculation of features. Clustering was sometimes used to group GPS samples into separate locations, for example by using the Density-based spatial clustering of applications with noise (DBSCAN) algorithm (Faurholt-Jepsen, et al., 2022; Zhang, et al., 2022) or K-means clustering (Cao, et al., 2020). Various thresholds were also used, for example to set minimum requirements for clusters. Various requirements were used in the identification of a home location. Several studies chose to use dimension reduction or feature selection methods to reduce the number of features in their feature set/s, including Principal Component Analysis (Pellegrini, et al., 2022; Pedrelli, et al., 2020), L1-Based Feature Selection and Tree-Based Feature Selection (Bai, et al., 2021).

Handling of missing data

A common issue amongst the studies in this review was the prevalence of missing smartphone data. Various tactics were used to handle this issue. Some studies did not mention any strategy for handling missing data, whilst others did not explicitly indicate the strategy used to deal with this, but acknowledge some kind of criteria. For example, by mentioning the use of “rich sensor data” (Cao, et al., 2020), or stating that analyses were only performed when sufficient data was

deemed to be available (Braund, et al., 2022). Other studies explicitly stated thresholds for data inclusion (Bai, et al., 2021; Faurholt-Jepsen, et al., 2022; Laiou, et al., 2022; Zhang, et al., 2021; Zhang, et al., 2022); if too high a percentage of data for a sample was missing then the sample was excluded. There was no consistent selection of thresholds for data inclusion across studies.

In other studies, researchers chose to exclude days that had any variable missing (Pellegrini, et al., 2022; Pedrelli, et al., 2020; Cho, et al., 2019). However, Cho, et al. (2019) indicated that not being able to incorporate missing data in models can lead to loss of large volumes of data, as illustrated by the exclusion of approximately 89% of their samples due to the inability of the selected model to account for missing data. Occasionally imputation was carried out on missing data (Pedrelli, et al., 2020; Pellegrini, et al., 2022; Zhang, et al. 2021), usually in combination with a minimum criterion of available data.

Discussion

MDD is a very common and debilitating mental disorder, associated with a high recurrence risk. The recent decade witnessed an upsurge of digital phenotyping studies to better understand dynamics of MDD and to advance precision methods efforts for (relapse) prevention. This systematic review investigated the predictive power of methods used in smartphone-based digital phenotyping research for MDD, identifying a total of 14 studies. These studies overall supported the use of digital phenotyping in MDD, but conjunctively showed that the field is still in relatively early stages, with much room for improvement in predictive performance and the understanding of individual differences in digital phenotypes still to be more rigorously developed.

Three of the studies investigated correlations between smartphone-derived features and depressive symptoms, and eight studies predicted depressive symptom scores from passive smartphone data (or in two studies, the inverse prediction). Four studies also sought to predict classes of data, including diagnosis, mood state or mood episode, with one study predicting mood state 3 days in advance (Cho, et al., 2019). Overall, studies tended to achieve moderate model performance, with no studies achieving notably high performance (e.g., the highest out-of-sample classification accuracies achieved during cross-validation tended to be in the range of 80-84%), and frequently reported lower accuracies, indicating relatively varied predictive value of current smartphone-based digital phenotyping methods. Moreover, few studies aimed to predict responses across time points. Finally, a small number of studies investigated individual differences, demonstrating differences between participants of different ages and employment statuses (e.g., in time spent at home (Zhang, et al., 2022; Laiou, et al., 2022)), as well as some gender differences (e.g., in location entropy and residential location count (Zhang, et al., 2022)). This is consistent with the common intuition that individuals have different phone

use and behavioural habits, and suggests that factors such as these could be included in prediction models to improve personalised predictions.

In the following sections, we describe key methodological themes identified in this review, and discuss the importance of feature construction in digital phenotyping. We then discuss important study differences that affect the comparison between studies, and identify limitations of this review, before discussing recommendations for the field of digital phenotyping.

Key methodological themes

The richness and complexity of digital phenotyping data brings about several challenges that need to be overcome through making careful methodological decisions. Due to the high temporal resolution of the data and many available feature options that contribute to high heterogeneity of the data, many studies chose to include time-averaged features to summarise their data. This is a practical way to reduce the large volumes of data to a smaller, more manageable, number of features. However, depending on the chosen level of granularity, this approach can greatly reduce information relating to the temporal dynamics of individual time series. Inclusion of measures reflecting variations in rhythm may have been done to overcome this disadvantage. For some features, the duration of data collection may affect their utility. For example, features such as circadian rhythm may be more reliably calculated when data is collected over a longer period of time, as this may provide indications of the relevance of variations given a person's "usual" behaviours. As many studies focused on shorter periods (e.g., two week windows), future studies may seek to collect passive smartphone data over longer periods of time, such as in the range of months rather than weeks. Studies also generally validated time-averaged digital phenotyping measures against symptom measures that are also time-averaged, for example by comparing averages from two weeks of digital phenotyping data to a PHQ score reflecting depressive symptoms over two weeks. Whilst this is a necessary

step during the early stages of digital phenotyping, it may detract from one of the original motivations behind digital phenotyping; i.e., to develop tools that can be used for real-time patient monitoring.

There was an absence of external validation datasets used in the identified studies, with performance results either reflecting the overall dataset, or an internal validation dataset (i.e., a subset of the overall dataset). Methods such as k-fold cross-validation help to give an indication of performance on withheld data, but do increase the risk of bias. External validation/testing sets may be difficult to obtain due to limited data availability, which could be aided in the future through more cross-collaboration and sharing of datasets.

High prevalence of missing data is a serious concern in smartphone-based digital phenotyping research, as the various different functionalities of smartphones may be unable to function under certain conditions (e.g., low battery, poor Wi-Fi connection, or intended smartphone switch-offs by users). Strategies for handling missing data therefore need to be applied. Studies identified by this review used minimum data inclusion criteria, exclusion of samples with missing values and/or imputation of missing values. The latter two options allow flexibility in the chosen models as they do not need to be able to account for missing data. However, as data may be missing due to the behaviours of the participants (e.g., missingness as a result of an ongoing depressive episode), it is possible that imputing missing data may disguise important behavioural changes. For example, Cohen, et al. (2023) found that their "data quality" feature, calculated based on the data that was successfully collected relative to the amount of data that was expected to be collected, was useful in predicting relapse in schizophrenia. This may also be the case for predicting clinically significant changes in depression.

Feature construction

As a large number of available sensors and options exist that researchers can select from before data collection begins, appropriate feature construction is key in digital phenotyping. This feature construction seems particularly guided by the technical availability of specific sensors and options, as different smartphone operating systems (e.g., Android, IoS) and applications have various restrictions. For example, Apple smartphones tend to have more restrictions on what data can be collected, and in applications such as WhatsApp, information about calls and messages cannot be accessed. Besides "technically-driven feature construction", concerns about ethics and/or privacy (e.g., Maher, et al., 2019) can also drive the choice of features. For example, no study analysed the predictive value of content of general phone text messages, or voice recordings. Passive voice collection is likely less popular due to ethical issues and privacy concerns, although active voice recordings have shown promise in differentiating between people with depression and healthy controls (Silva, Lopes, Galdino, & Almeida, 2021). Of the excluded studies, some did record voices in defined settings or through using specific exercises or assessments (e.g., Abbas, et al., 2021).

Moreover, selected feature sets tended to be influenced by domain knowledge of depression symptoms ("theory-driven feature construction"). For example, measures of home stay were common, and many studies included measures of behavioural rhythm, such as circadian rhythm (see Figure 2). One study (Cao, et al., 2020) attempted to include a proxy measure of sleep, although with known limitations as it is not expected that participants use their smartphones immediately before and after sleep. Relatedly, more "data-driven" feature constructions were also seen to be used, in which a broad range of features are calculated to investigate which are most useful (e.g., statistical features calculated for a measure of movement). There are also other features that seem to fall between theory- and data-driven; for example, screen time is a commonly used feature yet despite its common perception, it may not have a notable negative impact on mental health (Aschbrenner, et al., 2019). As we do not

yet fully understand the impact of this behaviour on mental health, its use as a digital marker cannot be considered to be completely theory-driven. Studies also did not record reasons a person may have for greater screen time (e.g., family or work commitments), so these features are lacking context that could contribute to our understanding of their potential relationship with depression.

Interestingly, several studies investigated individual differences in age, gender and occupational status, to inform predictions for individuals. Understanding the impact these differences have on smartphone-related behaviours may enable the development of more personalised digital phenotyping or prevention tools, for example through stratifying individuals into informative groups or detecting recurrence in depression. It is likely that the smartphone-derived features themselves may vary in usefulness between individuals (e.g., some individuals may never use the basic call function, whereas others may regularly make calls using this function). As such, feature selection could perhaps be carried out for each individual for use in individual models, or for subsets of individuals, although it would need to be investigated whether this extra computational step would lead to increases in model performance. Studies were yet to consider other factors that could affect smartphone measures that may impact individual predictions, such as family- or work-related smartphone usage or locations. For example, in the case of an individual who works from home, minimal mobility away from the home location and large call volume are likely unrelated, or rather inseparable, from their mood status. Measures related to expected smartphone usage and lifestyle may, therefore, help to inform smartphone-derived predictions on the individual level and interactions with changing contexts, and could be collected in future studies.

With all of the smartphone-derived features used, it should be noted that digital phenotyping is limited to the assessment of data on behaviours, activities or physical responses that can be passively registered by a smartphone. That is, it cannot assess underlying motives

or experiences behind these behaviours, activities or physical responses (unless EMA or self-report are additionally administered). For example, for an individual the qualitative inference can be made that using their smartphone at night may lead to depression, but the more direct relationship could be that the smartphone use during the night might indicate insomnia associated with depression. Without assessments of underlying motives or experiences, digital phenotyping results must be interpreted with care.

Important study differences

The aim of this review was not to compare smartphone-derived digital phenotypes to other phenotypes, such as wearable-derived digital phenotypes, however as some studies combined smartphone features with other features, we can make some preliminary comments on the effect of combining different types of phenotypes. For example, Pedrelli, et al. (2020) included wearable features, and found that it was inconclusive which modality performed better to predict residual depressive symptom scores. In their models classifying mood states, Bai, et al. (2021) found that their best performing model combined a smartphone feature (call logs) with wearable features (sleep, step count, heart rate), therefore outperforming models using smartphone data alone. Future reviews could seek to compare studies focusing on wearables more generally in predictions for MDD.

Aside from using passive features from other devices, one study chose to use previous depressive symptom scores to predict the following score, finding that this improved prediction (Pellegrini, et al., 2022). This same study found that including smartphone data in their various models did not always improve predictions, but noted that due to the convenience of smartphone data, smartphone-based models may still be worthwhile (also noted by Cao, et al. (2020)). Another study included historical weather data in their model (Pedrelli, et al., 2020), showing that it is possible to include broader contextual information that may affect an individual's behaviour in models. Interestingly, Pedrelli, et al. (2020) also found that individual

median HDRS scores provided better predictions than their machine learning models using passive data. Digital phenotyping researchers could seek to incorporate more contextual information in prediction models, including seasonal/time-related information.

Regarding studies that focused on predicting depressive symptom severity, several studies (Braund, et al., 2022; Cao, et al., 2020; Zhang, et al., 2021; Zhang, et al., 2022) used participant-rated depressive symptoms in the form of PHQ scores. Others focused on clinician-rated depressive symptoms, such as the MADRS (Sverdlov, et al., 2021; Pellegrini, et al., 2022) and the HDRS (Pedrelli, et al., 2020). Using participant-rated symptom severity scores may allow for more frequent symptom assessment and therefore a higher temporal resolution of symptom course to be predicted, whereas clinician-rated symptom severity may allow for more consistent measurements between participants, but is less convenient for frequent assessments.

Varying quantities of data (i.e., units of analysis) were defined as a sample by each study, with some studies treating each day as a separate sample. The PHQ was a commonly used assessment tool for measuring depressive symptoms, and as this tool assesses symptoms over a period of 2 weeks, most studies using this tool selected passive smartphone features calculated from the two weeks preceding the administration of the questionnaire. Other studies chose to focus on a participant's entire data, which allows for longer periods to be analysed. Thus, the unit of analysis that is selected highly depends on research aims or the intended clinical application.

As is the case for other health applications, an open question in digital phenotyping research is whether to develop individual or group models, or perhaps combinations of the two. This review identified studies that predominantly applied group models, with individual factors commonly addressed in models as covariates (Faurholt-Jepsen, et al., 2022; Zhang, et al., 2021; Zhang, et al., 2022), or as predictors of interest themselves (Pellegrini, et al., 2022; Laiou, et al., 2022). Two studies compared group models to individual models (Pedrelli, et al., 2020;

Cho, et al., 2019). In these studies, it was found that individual models often outperformed group models. In terms of developing models that are useful in practice, it may be worthwhile to investigate how other patients' data may be useful for individual predictions, as varying amounts of data can be expected per patient; for example, some patients may have more than a year's worth of data available, whereas new patients may only have a few weeks of data available but could still be experiencing clinically significant changes. Models with some shared group parameters may therefore be able to contribute to clinical predictions, which could also be informed by appropriate clustering of patients using factors such as employment status and other important individual differences.

Limitations

This systematic review identified several studies that have made important progress in linking behavioural phenotypes to clinically-relevant variables such as symptom severity and mood state, despite challenges arising from the nature of digital phenotyping data. These included frequent issues with missing data, and the need to combine various high temporal resolution channels in a meaningful way. This often led to the exclusion of large volumes of data and the common use of time-averaged features, risking the loss of useful information relating to temporal dynamics. The digital phenotyping field still needs to achieve higher model performance before the models can be clinically useful without adding additional burden to clinicians in the form of difficult-to-interpret-models or models with low predictive power and consequently high rates of false positive and false negative predictions.

Our review has some limitations, especially with regards to the effect of our search criteria. Studies were restricted to those with MDD populations to avoid too much heterogeneity between different psychiatric populations and/or too general populations with relatively low symptom scores. This may have introduced a selection bias in favour of studies from research groups in WEIRD ('Western, Educated, Industrialized, Rich, and Democratic')

countries or regions, which could have more resources contributing to their mental health care systems and easier access to a population with diagnosed major depression. This criterion also led to online studies being excluded, as possible MDD diagnosis of participants could not be confirmed using clinical tools. By restricting to studies including an MDD population, this consequently limited the number of studies that could be compared within each analysis goal. In addition, differences in the methodologies used by each of the studies make it more complicated to determine what the overall most predictive variables are for the different goals, limiting the possibilities to make direct comparisons. Future reviews could seek to focus more broadly on studies of depressive symptoms and/or other psychiatric populations within single prediction goals, for a more in-depth comparison of the methods. Our criteria also required passive data to be collected in a naturalistic setting, which could exclude some feature types. For example, it is difficult to collect speech in a naturalistic setting, hence we did not identify any studies with speech features that met our criteria.

Recommendations

The popularity of digital phenotyping continues to grow as the smartphone maintains its place in today's world. In order to fully take advantage of digital phenotyping's clinical potential in MDD, attention should be paid to careful model development.

Firstly, given the changing nature of human behavior, it is important to acknowledge the temporal dynamics of clinically-relevant changes in the formulation of prediction goals and selection of prediction methods. That is, future approaches may seek to investigate temporal dynamics more directly through choosing models which can handle time series data. For example, a recent paper on predicting schizophrenia relapse using smartphone data applied an anomaly detection approach to investigate whether daily features are anomalous relative to nearby days (Cohen, et al., 2023). To gain further insight into temporal dynamics of individuals' experiences associated with change, digital phenotyping approaches can be

combined with EMA data. A shift towards investigations of temporal dynamics may provide more timely predictions of clinically-relevant changes.

Secondly, before models can reasonably be expected to be used by clinicians in practice, their performance should be improved. Greater collaboration between research groups could allow for larger datasets to be used in model development, and more investigations of generalisability. Replicating results in external validation or testing datasets aids in generalising results to broader settings. As such, increasing efforts to externally validate model performance can help strengthen arguments that digital phenotyping tools can be useful in clinical practice.

Thirdly, models should account for missing data to avoid excessive sample exclusions. To summarise the steps that can be taken to handle the challenge of missing data, efforts can be made to minimise missing data during data collection. Clear instructions should be given to participants so they do not accidentally switch off app functionalities required for collection. However, even if users do not accidentally cause data collection to be impacted, large volumes of missing data can still occur. Incoming data should be regularly inspected to ensure prolonged periods of missing data are not occurring, and researchers can then take action to restore app functionality if this is indeed the case. Data should be inspected across the various sensors that are investigated in case the issue is not affecting all sensors. It could be useful to develop automatic data-checking tools to identify periods of missing data, especially before incorporating prediction models in clinical practice. Even once all has been done to minimise missing data occurring during data collection, there will inevitably still be some instances of missing data that need to be handled. Ideally, minimal participants/samples need to be discarded, although minimum data availability requirements may be needed to filter out participants/samples that are missing large volumes of data. Thresholds for missing value requirements do not necessarily need to be consistent across studies, but could be investigated

during model selection/training. For the remaining participants/samples, an appropriate imputation method could be considered (and models with and without imputation compared). Models could also be chosen based on their ability to manage missing data, for example, Hidden Markov Models can accommodate for missing timepoints. Whilst there is not necessarily a one-size-fits-all approach to handling missing data, overall, it seems that to model digital phenotyping data, minimum data availability requirements are needed, and ideally models allowing for data to be missing should be used. The potential bias arising due to the non-randomness of missing data and the strategy used to handle missing data remain unexplored.

Conclusion

As the field of digital phenotyping develops, we get closer towards the goal of making insightful clinical predictions that can help people with depression, through earlier identification of changes in symptom course and possible onset of future episodes. The studies identified in this review demonstrated moderate success across various prediction goals, including predicting symptom severity and mood state, despite challenges from complex, high-dimensional time series and a high propensity for missing data. Once models with current prediction goals can achieve higher performance across different settings and MDD populations, digital phenotyping research could start to shift towards investigating how to implement these models in practice, for example whether rolling windows should be used to analyse the incoming temporal data. With careful model decisions and implementations, including clinically- and technically-informed feature construction and appropriate validations, digital phenotyping methods for MDD could be generalised to other disorders, with the eventual goal to one day be able to make online predictions of mental disorders that can be directly used by clinicians for improved individualised interventions and patient outcomes.

References

Abbas, A., Sauder, C., Yadav, V., Koesmahargyo, V., Aghjayan, A., Marecki, S., . . . Galatzer-Levy, I. R. (2021). Remote digital measurement of facial and vocal markers of major depressive disorder severity and treatment response: a pilot study. *Frontiers in digital health, 3*, 610006. doi:10.3389/fdgth.2021.610006

Aschbrenner, K. A., Naslund, J. A., Tomlinson, E. F., Kinney, A., Pratt, S. I., & Brunette, M. F. (2019). Adolescents' use of digital technologies and preferences for mobile health coaching in public mental health settings. *Frontiers in public health, 7*, 178. doi:10.3389/fpubh.2019.00178

Bai, R., Xiao, L., Guo, Y., Zhu, X., Li, N., Wang, Y., . . . Wang, G. (2021). Tracking and monitoring mood stability of patients with major depressive disorder by machine learning models using passive digital data: Prospective naturalistic multicenter study. *JMIR mHealth and uHealth, 9*(3). doi:10.2196/24365

Benoit, J., Onyeaka, H., Keshavan, M., & Torous, J. (2020). Systematic review of digital phenotyping and machine learning in psychosis spectrum illnesses. *Harvard Review of Psychiatry, 28*, 296–304. doi:10.1097/HRP.0000000000000268

Braund, T. A., Zin, M. T., Boonstra, T. W., Wong, Q. J., Larsen, M. E., Christensen, H., . . . O'Dea, B. (2022). Smartphone Sensor Data for Identifying and Monitoring Symptoms of Mood Disorders: A Longitudinal Observational Study. *JMIR Mental Health, 9*(5). doi:10.2196/35549

Buckman, J. E., Underwood, A., Clarke, K., Saunders, R., Hollon, S. D., Fearon, P., & Pilling, S. (2018). Risk factors for relapse and recurrence of depression in adults and how they operate: A four-phase systematic review and meta-synthesis. *Clinical psychology review, 64*, 13–38. doi:10.1016/j.cpr.2018.07.005

Burcusa, S. L., & Iacono, W. G. (2007). Risk for recurrence in depression. *Clinical psychology review, 27*, 959–985. doi:https://doi.org/10.1016/j.cpr.2007.02.005

Cao, J., Truong, A. L., Banu, S., Shah, A. A., Sabharwal, A., & Moukaddam, N. (2020). Tracking and predicting depressive symptoms of adolescents using smartphone-based self-reports, parental evaluations, and passive phone sensor data: Development and usability study. *JMIR Mental Health, 7*(1). doi:10.2196/14045

Cho, C. H., Lee, T., Kim, M. G., In, H. P., Kim, L., & Lee, H. J. (2019). Mood prediction of patients with mood disorders by machine learning using passive digital phenotypes based on the circadian rhythm: Prospective observational cohort study. *Journal of Medical Internet Research, 21*(4). doi:10.2196/11029

Cohen, A., Naslund, J. A., Chang, S., Nagendra, S., Bhan, A., Rozatkar, A., . . . others. (2023). Relapse prediction in schizophrenia with smartphone digital phenotyping during COVID-19: a prospective, three-site, two-country, longitudinal study. *Schizophrenia, 9*, 6. doi:10.1038/s41537-023-00332-5

Emden, D., Goltermann, J., Dannlowski, U., Hahn, T., & Opel, N. (2021). Technical feasibility and adherence of the Remote Monitoring Application in Psychiatry (ReMAP) for the assessment of affective symptoms. *Journal of Affective Disorders, 294*, 652-660. doi:10.1016/j.jad.2021.07.030

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. *2016 IEEE wireless health (WH)*, (pp. 1–8). doi:https://doi.org/10.1109/WH.2016.7764553

Faurholt-Jepsen, M., Busk, J., Rohani, D. A., Frost, M., Tønning, M. L., Bardram, J. E., & Kessing, L. V. (2022). Differences in mobility patterns according to machine learning

models in patients with bipolar disorder and patients with unipolar disorder. *Journal of Affective Disorders, 306*, 246-253. doi:10.1016/j.jad.2022.03.054

Harari, G. M., Müller, S. R., Aung, M. S., & Rentfrow, P. J. (2017). Smartphone sensing methods for studying behavior in everyday life. *Current opinion in behavioral sciences, 18*, 83–90. doi:10.1016/j.cobeha.2017.07.018

Higgins J.P., Sterne J.A., Savović J., Page M.J., Hróbjartsson A., Boutron I., Reeves B., Eldridge S. (2016). A revised tool for assessing risk of bias in randomized trials In: Chandler J., McKenzie J., Boutron I., Welch V. (editors). Cochrane Methods. *Cochrane Database of Systematic Reviews 2016, 10* (Suppl 1). doi:dx.doi.org/10.1002/14651858.CD201601

Laiou, P., Kaliukhovich, D. A., Folarin, A. A., Ranjan, Y., Rashid, Z., Conde, P., . . . Hotopf, M. (2022). The Association between Home Stay and Symptom Severity in Major Depressive Disorder: Preliminary Findings from a Multicenter Observational Study Using Geolocation Data from Smartphones. *JMIR mHealth and uHealth, 10*(1). doi:10.2196/28095

Lim, S. S., Vos, T., Flaxman, A. D., Danaei, G., Shibuya, K., Adair-Rohani, H., . . . others. (2012). A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The lancet, 380*, 2224–2260. doi:10.1016/S0140-6736(12)61766-8

Luo, W., Phung, D., Tran, T., Gupta, S., Rana, S., Karmakar, C., . . . others. (2016). Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *Journal of medical Internet research, 18*, e323. doi:10.2196/jmir.5870

Maher, N. A., Senders, J. T., Hulsbergen, A. F., Lamba, N., Parker, M., Onnela, J.-P., . . . Broekman, M. L. (2019). Passive data collection and use in healthcare: A systematic review of ethical issues. *International journal of medical informatics, 129*, 242–247. doi:https://doi.org/10.1016/j.ijmedinf.2019.06.015

Matcham, F., Leightley, D., Siddi, S., Lamers, F., White, K. M., Annas, P., . . . Hotopf, M. (2022). Remote Assessment of Disease and Relapse in Major Depressive Disorder (RADAR-MDD): recruitment, retention, and data availability in a longitudinal remote measurement study. *BMC Psychiatry, 22*(1). doi:10.1186/s12888-022-03753-1

Müller, S. R., Chen, X., Peters, H., Chaintreau, A., & Matz, S. C. (2021). Depression predictions from GPS-based mobility do not generalize well to large demographically heterogeneous samples. *Scientific Reports, 11*, 14007. doi:10.1038/s41598-021-93087-x

Nelson, B. W., & Allen, N. B. (2018). Extending the passive-sensing toolbox: using smart-home technology in psychological science. *Perspectives on psychological science, 13*, 718–733. doi:10.1177/1745691618776008

Pedrelli, P., Fedor, S., Ghandeharioun, A., Howe, E., Ionescu, D. F., Bhathena, D., . . . Picard, R. W. (2020). Monitoring Changes in Depression Severity Using Wearable and Mobile Sensors. *Frontiers in Psychiatry, 11*. doi:10.3389/fpsyt.2020.584711

Pellegrini, A. M., Huang, E. J., Staples, P. C., Hart, K. L., Lorme, J. M., Brown, H. E., . . . Onnela, J. P. (2022). Estimating longitudinal depressive symptoms from smartphone data in a transdiagnostic cohort. *Brain and Behavior, 12*(2). doi:10.1002/brb3.2077

Saeb, S., Zhang, M., Karr, C. J., Schueller, S. M., Corden, M. E., Kording, K. P., . . . others. (2015). Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *Journal of medical Internet research, 17*, e4273. doi:10.2196/jmir.4273

Silva, W. J., Lopes, L., Galdino, M. K., & Almeida, A. A. (2021). Voice acoustic parameters as predictors of depression. *Journal of Voice*. doi:https://doi.org/10.1016/j.jvoice.2021.06.018

Sverdlov, O., Curcic, J., Hannesdottir, K., Gou, L., Luca, V. D., Ambrosetti, F., . . . Jacobs, G. E. (2021). A Study of Novel Exploratory Tools, Digital Technologies, and Central Nervous System Biomarkers to Characterize Unipolar Depression. *Frontiers in Psychiatry, 12*. doi:10.3389/fpsyt.2021.640741

Tønning, M. L., Faurholt-Jepsen, M., Frost, M., Bardram, J. E., & Kessing, L. V. (2021). Mood and Activity Measured Using Smartphones in Unipolar Depressive Disorder. *Frontiers in Psychiatry, 12*. doi:10.3389/fpsyt.2021.701360

Ware, S., Yue, C., Morillo, R., Lu, J., Shang, C., Bi, J., . . . Wang, B. (2020). Predicting depressive symptoms using smartphone data. *Smart Health, 15*, 100093. doi:10.1016/j.smhl.2019.100093

Zhang, Y., Folarin, A. A., Sun, S., Cummins, N., Ranjan, Y., Rashid, Z., . . . Dobson, R. J. (2021). Predicting depressive symptom severity through individuals' nearby bluetooth device count data collected by mobile phones: Preliminary longitudinal study. *JMIR mHealth and uHealth, 9*(7). doi:10.2196/29840

Zhang, Y., Folarin, A. A., Sun, S., Cummins, N., Vairavan, S., Bendayan, R., . . . Dobson, R. J. (2022). Longitudinal Relationships Between Depressive Symptom Severity and Phone-Measured Mobility: Dynamic Structural Equation Modeling Study. *JMIR Mental Health, 9*(3). doi:10.2196/34898

Figures

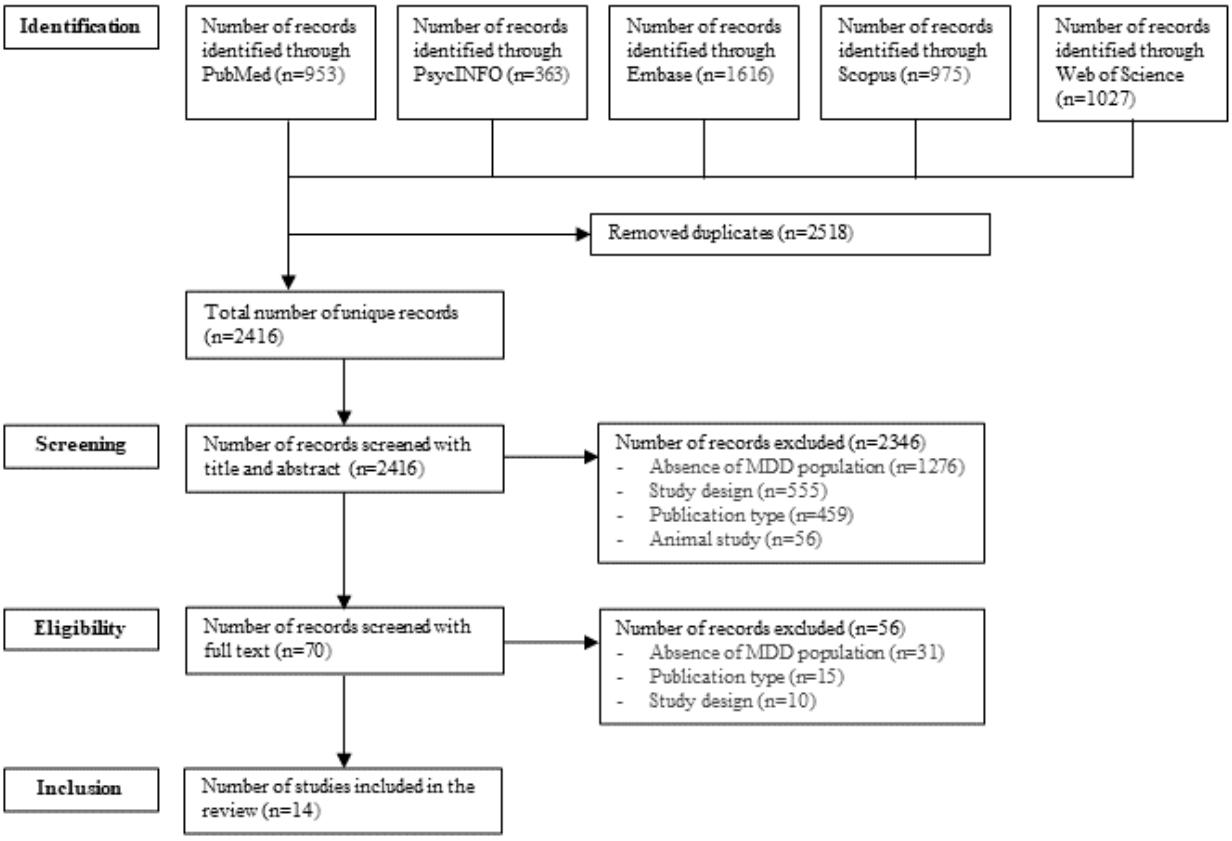

Figure 1. Flowchart of selection and inclusion process following the PRISMA Statement.

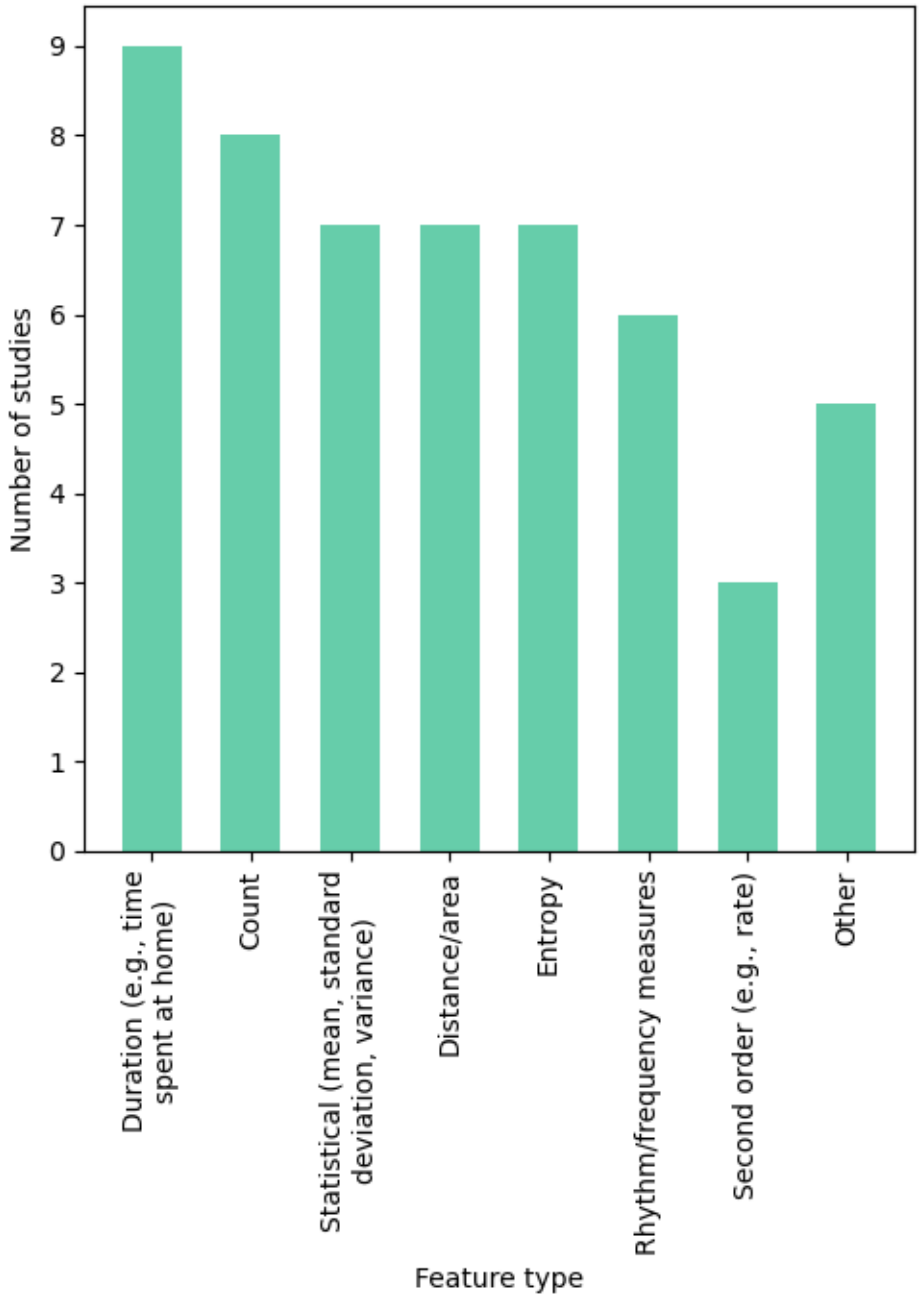

Figure 2. Number of studies (out of the 14 selected studies) using each broad category of feature type (N.B., studies generally used more than one feature type).

Tables

Table 1. Characteristics of Included Studies

Author (year of publication)	Country	Population	% Female	Age (*M*)	Ethnicity	Sample size
Bai et al. (2021)	China	MDD	NP	NP	NP	334
Braund et al. (2022)	Australia	MDD (n = 79); BD (n= 42)	65.3	41.4	NP	121
Cao et al. (2020)	USA	MDD	84.6	14.93	NP	13
Cho et al. (2019)	Korea	MDD (n =18); BD I (n =18); BD II (n = 19)	49.1	25.92	NP	55
Emden et al. (2020)	Germany	MDD (n = 409); BD (n = 48); AD (n = 58); PD (n = 21); HC (n = 458)	67.3	35.99	NP	997
Faurholt-Jepsen et al. (2022)	Denmark	MDD (n = 75); BD (n = 65)	56.6	44.2	NP	140
Laiou et al. (2021)	UK, Netherlands, Spain	MDD	75.0	48 (median)	NP	164
Matcham et al. (2022)	UK, Netherlands, Spain	MDD (n = 378); HC (n = 245)	75.6	46.4	NP	623

Pedrelli et al. (2020)	USA	MDD	74.0	33.7	White = 71%, Hispanic/Latino = 23%, Asian = 16%, Haitian/Black/African-American = 12%, American Indian/Alaskan = 3%, mixed-race = 6%, and other = 3%	41
Pellegrini et al. (2021)	USA	MDD (n = 10); BD (n = 10); schizophrenia or schizoaffective disorder (n = 10); HC (n = 11)	63.0	43.0	White = 71%, African-American = 20%, Asian = 7%, Other = 2%	41
Sverdlov et al. (2021)	The Netherlands	MDD (n = 20); HC (n = 20)	33.0	31.2	White = 87.5%, Mixed = 7.5%, Asian = 2.5%, Black or African-American = 2.5%	40
Tønning et al. (2021)	Denmark	MDD	52.7	44.4	NP	74
Zhang et al. (2021)	Spain, The Netherlands, UK	MDD	74.1	51.0 (median)	NP	316
Zhang et al. (2022)	Spain, The Netherlands, UK	MDD	74.1	50.0 (median)	NP	290

AD = Anxiety Disorder; BP (I or II) = Bipolar Disorder (Type 1 or Type 2); GPS = Global Positioning System; HC = Healthy Control; M = Mean; MDD = Major Depressive Disorder; NP = Not Provided; PD = Psychotic Disorder

Table 2. General Methodological Information Of All Included Studies

Author (year of publication)	What did a single measurement refer to?	Inclusion and exclusion criteria for data	Outlier removal	Handling of missing values	Variable/feature processing, selection, generation, or reduction	Potential bias or factors affecting generalisability
Bai et al. (2021)	3 consecutive PHQ-9 results & smartphone data collected between 1st & last PHQ-9	-	-	Sample excluded if lasts <1 week, or if contained <3 days of effective data	- Phone usage features calculated using screen on/off status - Periods used: midnight to 3 AM, 3 AM to 6 AM, 6 AM to 9 AM, 9 AM to noon, noon to 3 PM, 3 PM to 6PM, 6 PM to 9 PM, 9 PM to midnight - Two feature selection models used to find feature subset: L1-Based Feature Selection, Tree-Based Feature Selection	Imbalance of data led most of the models to have higher recall than accuracy
Braund et al. (2022)	One participant	-	Not mentioned	Circadian rhythm only calculated when sufficient data available.	Circadian rhythm: extent to which individual's sequence of locations followed 24-hour rhythm. Determined at baseline based on changes in GPS location during first 2 weeks. Least squares spectral analysis performed to estimate amount of energy that fell into 24-hour frequency bins. Circadian rhythm calculated as logarithm of sum of	- Low symptom levels - No healthy controls limits generalizability to nonclinical

					energy for longitude and latitude.	
Cao et al. (2020)	2 weeks of data	-	Not mentioned	Not mentioned, though says fitted regressors for "rich sensor data"	- Estimated moving speed at each location, labelled as stationary or transition state - Points in stationary states: used K-means clustering - Points in transitional states: used speed to categorise as automobile, walking, unknown	- Small sample size - But no evidence that parents' evaluations biased by their own mental health status
Cho et al. (2019)	Each day	-	-	Removed days if any variable missing	- Daily mood scores (−3-+3) converted to absolute mood score (AMS; 0-3) - Higher AMS: mood worse & unstable - Neutral mood: if average AMS for following 3 days within bottom 50% of all AMS (vice versa for biased mood) - Average light exposure calculated during defined bedtime & daytime - Extended daily features to include previous 3, 6, 12 days. SD & gradient coefficient computed	Application of HAMS & LAMS may not accurately reflect mood state
Emden et al. (2020)	One participant	-	-	-	Each GPS location event consists of longitude, latitude, timestamp.	No identified differences between individuals willing &

					Acceleration provided as (x, y, z) vector. - Steps & distance walked recorded from Apple HealthKit/Google Fit	not willing to download app
Faurholt-Jepsen et al. (2022)	One Participant (during affective state)	-	Outliers (points with unrealistic acceleration) and duplicate samples removed.	>=50 location samples/day required	- Places: specific locations - Stops: visits to any place - Moves: sequences of locations between stops - Stops: locations sequentially grouped using maximum distance threshold. New group created If sample >threshold from current group median. Minimum distance 50m, minimum duration 20min. Stops <=5min & 5m apart merged - Places: identified using DBSCAN clustering - Moves: location sequences in-between stops; minimum duration 5 min, minimum distance 50 m - Distance between 2 days: counted corresponding hrs with same location, computed mean	- Did not account for multiple testing - Imbalanced classes but method accounted for this - Differences in psychopharmacological treatment between patients with BD & UD - Mobility patterns can be very individual, may not generalize across groups - Low symptom severity

					- Routine index: mean of distances to each day of preceding 28 days - Radius of gyration: deviation from centroids of daily stops. Distances weighted by time spent at location - Depressive state: smartphone-based mood ≤-1 - Euthymic state: -1< mood <1	
Laiou et al. (2021)	PHQ-8 record combined with GPS data from preceding 2 weeks	- Segment needed: 14 days of GPS recordings, daily median GPS sampling period ≤11 minutes, daily number of acquired GPS data points ≥48 - Data before January 1, 2020 (excluded COVID-19 period)	GPS data points with accuracy >20 meters removed.	Only biweekly segments that met the criteria analyzed: 14 days of GPS recordings available, daily median sampling period of GPS signal =11 minutes, daily number of acquired GPS data points =48.	- For each day: computed number of GPS points collected for each hour & over whole day - Daily completeness: ratio between actual number of GPS points collected each day & expected number - Sampling constancy: divided actual number of GPS points collected per hour by expected number, computed SD - Home location: calculated median longitude & latitude of all points in segment between 12 AM-6 AM. All points in this time >60m from initial home location	- Nonhomogeneous population from three European countries, but most from one site - Antidepressant side effects or illness may make people more likely to spend time at home - GPS criteria decreased number of available segments - Poor signals, e.g. in rural areas, led to more missing data

					removed. Home then set as median longitude & latitude of remaining points. - Home stay: ratio between number of points ≤60m from home & total number - Each variable transformed using Yeo–Johnson then zero-mean, unit-variance normalization	
Matcham et al. (2022)	One participant	Data considered "available" if: complete/ usable; incomplete/potentially usable; not usable	-	-	- People who provided any passive data divided into: 0–25%, 26–50%, 51–75%, >75% of expected data - Participants grouped as "symptomatic", "some/no/mild symptoms"	- Higher % of white & female participants than in general or depressed populations-Mean age & gender distribution comparable to other MDD samples - Depressive symptoms & pandemic may affect dropout rate & adherence - Difficult to compare ethnic groups between the two countries who had this data
Pedrelli et al. (2020)	Features from same day (midnight-to-midnight) of HDRS	-	Used Theil-Sen estimator, random sample consensus, huber algorithms: have built-in sampling	- Excluded days with data missing due to technical problems - Extrapolated missing latitude & longitude values	- Calculated features for 6-h intervals & per day. - Down-sampled location data - Location used to retrieve weather data from DarkSky API	- Small sample size - Low variability of depressive symptoms - Unclear whether test of significance is most appropriate metric to compare models

			procedure allowing fraction of data points to be outliers.		- Used kernel PCA to get 25 transformed features - HDRS residual scores: subtracted screening score from scores from later visits	
Pellegrini et al. (2021)	Multiple MADRS assessments for each subject	-	-	Imputed missing GPS trajectories using resampling method. Excluded data points with missing values for 1+ of the predictors.	Computed 32 candidate summary statistics from GPS and accelerometer data. PCA on the 32 statistics and used the first principal component as a predictor.	- Predictors may be inaccurate for low proportions of collected data - may underestimate more severe depression due to omission of PHQ suicide item - modest sample size
Sverdlov et al. (2021)	One participant	-	Low data quality observations removed	No missing data imputation	Took averages of valid longitudinal measurements within subject.	- No age or gender matching done - More females than males - Various antidepressants used for varying durations - Small sample size - Short study duration - Expect heterogeneity of depressed sample reflects non-clinical trial depressed populations
Tønning et al. (2021)	- Aim 2: 3 days preceding HDRS - Aim 3: each day	-	-	Missing items from ratings & questionnaires not included in the	Aims 1 & 2: - Averages of daily smartphone data (passive & patient-reported) used day of	- Severely ill patients - Data collected in period following discharge - Study participation &/or intervention

				summed scores, no imputations or assumptions on missing items made.	rating & 3 days before HDRS-17 - If data not available, used following 3 days Aim 3: - Used all smartphone-based patient-reported data with same-day passive data, w/out averages or sums	might have affected patients' behavior - Longer call durations could be due to more incoming calls from concerned friends/family
Zhang et al. (2021)	14 days preceding PHQ-8	Data before February 2020 (excluded COVID-19 period)	-	- Usable days contained >=12 hours of data - Included PHQ-8 intervals with >=10 usable days - Imputed missing hours using linear interpolation - Prediction task subset: >=3 valid intervals per participant	- Calculated max, min, mean, SD of daily NBDC data from all valid days - For each feature: calculated 2nd-order features (max, min, mean, SD) - Complexity & periodicity on different timescales (1-24h): calculated MSE features for each scale - Used Fast Fourier transformation. Three frequency intervals: low (0-0.75 cycles/day), middle (0.75-1.25), high (>1.25) - Calculated sums of spectrum power & spectral entropy in each interval, %s of spectrum powers in each interval to total spectrum power	- Large, multicenter dataset - Long follow-up - Lower data resolution - Cannot differentiate between phones & other devices, strangers' & acquaintances' devices - Manually set spectrum frequency intervals

					- Prediction task subset: difference w/in each participant's scores >= 5 for inclusion	
Zhang et al. (2022)	14 days preceding each PHQ-8 record	- Data before February 2020 (excluded COVID-19 period) - location records with error >165 meters removed - amount of missing location data in interval <=50%	Location records with error > 165 meters removed	Missing location data in a PHQ-8 interval limited to 50%	- Calculated distances between consecutive locations & instantaneous speeds at all locations - Location clustering: used "density-based spatial clustering of applications with noise" method - Moving distance: divided total distance by available time in interval - Home: selected all location clusters visited at night in interval. Cluster with most records set as home - Homestay: % time spent at all home locations. Frequency-domain: used linear interpolation & fast Fourier transformation to get spectrums of longitude & latitude (scaled in cycles/day). Grouped bands as low, middle, high frequency	Unexpectedly, gender not found to be significantly correlated with PHQ-8

		- Summed power in each band of longitude & latitude, divided by sum of total spectral power of longitude & latitude

Table 3. Methodological Details Of Studies Correlating Passive Smartphone Features With Depression Symptom Severity

Author (year of publication)	Predictor variables	Response variable/s	Modelling techniques	Quality metrics used to assess model performance	Results	Interpretation of final model
Cao et al. (2020)	- Total number of steps during nighttime/daytime per day - GPS: total distance, transition time, location variance, number of frequently visited places/day, normalized entropy - Calls: number of unique phone call partners/day, total number & duration of phone calls, average call duration with frequent contacts/day, text message number/length, average text message length with frequent contacts/day, total number of hours with above-threshold average light intensity during nighttime per day (~sleep duration), above-threshold % of hours with screen on during nighttime/daytime per day	Biweekly psychometric scores (HAM-D, PHQ-9)	Pearson correlation	Significance levels marked as P<.1, P<.05, P<.01	- Participants with higher depression scores had lower mobility (fewer steps), visited fewer places, had lower location variance, but spent time more uniformly across different places - Higher depression score correlated with shorter phone call durations, fewer text messages - No significant correlation between ambient light intensity or smartphone screen usage & scores	- Overall results indicate depression reduces mobility - But authors uncertain of interpreting positive association between symptoms & location entropy - Sleep duration proxy not significantly correlated with depressive symptoms
Sverdlov et al. (2021)	Possible variables: total number of staypoints, number of places visited once, number of unique places visited, total amount of time spent at	Total MADRS score	Pairwise correlations	-	Participants with higher MADRS had lower average distance from home (r=−0.25), lower entropy of communication apps usage time (r=−0.31),	- Evidence of moderate correlations between some BeHapp features & MADRS

	home, average distance from home, average repetition per contact for outgoing, incoming, missed calls, total number of WhatsApp calls, entropy and mean of usage time of apps, total usage count of apps				lower total count of communication apps usage (r=−0.42), lower number of WhatsApp usage (r=−0.43)	- Expect higher symptom severity associated with lower social activity
Zhang et al. (2022)	- Each PHQ-8 score & mobility feature: location variance, moving time/distance, number of clusters, location/normalized entropy, homestay, residential location count, long-term/circadian/ short-term rhythm - Covariates: age, gender, work status	-	Correlations at within- & between-individual levels estimated using vector autoregressive model	Significant at adjusted P <.05	- Within-individual level: except for moving time, all mobility features significantly correlated with PHQ-8; homestay, short-term rhythm positively, other features negatively - Between-individuals: location variance, moving distance negatively correlated with PHQ-8. Individual Differences: - Older & employed participants had significantly lower intercepts of PHQ-8 - Age significantly, negatively correlated with number of clusters, location entropy, residential location count - Work status significantly correlated with most mobility features (except moving time, residential location count)	- Within-individual level: higher severity of depressive symptoms, lower mobility - Between-individual level: many correlations not significant, perhaps due to individual differences. Location variance, moving distance still significant so may be useful for predicting symptom severity

Table 4. Methodological Details Of Studies Predicting Depression Symptom Severity Or Predicting Passive Smartphone Features From Depression Symptom Severity

Author (year of publication)	Goal of analysis	Predictor variables	Response variable/s	Modelling techniques	Validation/assessment strategy	Validation/assessment metrics	Results from validation/assessment	Clinical implications
Braund et al. (2022)	To determine the relationship between circadian rhythm in GPS data and symptoms of mental health measures across 6 time points among a clinical sample of adults diagnosed with MDD or BD.	- Circadian rhythm - Random effect: participant; fixed effect: time point	PHQ-9	- Linear regression models (interaction terms between circadian rhythm & diagnostic group first included) - Mixed linear models (interaction terms between fixed effects of time point & diagnostic group first included)	-	P values corrected using Benjamini-Hochberg procedure	- No interaction found between time points & circadian rhythm for people with MDD - No significant interactions found among time point, circadian rhythm, diagnosis category or time point & circadian rhythm - Non-significant moderating effect of baseline circadian rhythm & change in depression severity across time points.	Specific implications for MDD not mentioned
Cao et al. (2020)	To investigate whether smartphone apps are useful in evaluating, monitoring depression symptoms in clinically depressed adolescents	- Total number of steps during nighttime/daytime per day - GPS: total distance, transition time, location variance, number of frequently visited places/day,	PHQ-9	Linear and nonlinear regressors (support vector regressor with a polynomial kernel)	- Divided dataset into training & test sets (4:1 split) - Tried different feature subsets to find which gave highest accuracy	RMSE of predicted PHQ-9 on test set	- Teens' or parents' inputs only: similar accuracies (RMSE 3.38, 3.47) (using linear regressor) - Teens' inputs & smartphone data: predicted PHQ-9 with RMSE 3.23 (support vector regressor)	- App could reduce user effort by just using passive data, as monitoring effect is comparable to non-passive model

	compared with PHQ-9, HAM-D, HAM-A.	normalized entropy - Calls: number of unique phone call partners/day, total number & duration of phone calls, average call duration with frequent contacts/day, text message number/length, average text message length with frequent contacts/day, total number of hours with above-threshold average light intensity during nighttime per day (~sleep duration), above-threshold % of hours with screen on during nighttime/daytime per day					- Including parents' evaluations: reduced RMSE to 2.65 (lowest error achieved) (support vector regressor), - Only smartphone data: RMSE 2.77 (linear regressor) - Lower mobility levels & fewer social interactions predictive of more severe symptoms	
Faurholt-Jepsen et al. (2022)	To investigate differences in passively collected smartphone-based location data between BD and UD	Daily features: number of stops, total duration of stops, number of places, number of moves, total distance moved, total duration of moves, routine	Number of stops, places & moves, duration of stops, routine index	Two-leveled mixed effects regression model (first level: repeated measurements per patient; second level:	-	-	- Compared with patients with BD during depressive state, patients with UD during depressive state covered significantly larger area/day, had	- Location entropy is possible 'digital activity marker' discriminating between

	overall, during depressive states, during euthymic states.	index, radius of gyration, location entropy. Covariates for all analyses: age, gender, work status		between-individuals variation).			borderline significantly larger total distance of moves/day and significantly larger total duration of moves/day. - Compared with patients with BD overall & during euthymic state, patients with UD overall & during euthymic state had higher location entropy - Patients with UD regardless of state & during euthymia had more location entropy during the day.	patients with BD & UD - May help with earlier correct diagnosis & treatment
Laiou et al. (2021)	To examine the association between the overall severity of MDD symptoms and a measure of daily mobility patterns	- PHQ-8 score - age, gender, occupational status, median completeness, sampling constancy	Homestay	Linear regression model	Randomly selected biweekly segment per participant, pooled data across participants & fitted model. Repeated 100 times to get CI for each regression coefficient.	Variable significantly related to home stay if 95% 2-sided CI did not include 0.	- Greater overall severity of MDD symptoms associated with prolonged home stay (also observed for weekday data, but not on weekends). Individual differences: - Spent more time at home with increasing age. Strength of relationship between home stay	Expect stronger relationship between homestay and symptom severity for younger people

							& MDD severity moderated by age - Employed spent less time at home compared with unemployed participants - Gender, median completeness, sampling constancy had no significant impact on home stay	
Pedrelli et al. (2020)	- To evaluate the feasibility and performance of an ML model combining physiological & smartphone features to assess depressive symptoms severity, as well as separate models for wearable & smartphone features. - To investigate which features are most	-877 features (473 from mobile sensors) - Latitude, longitude, % of time at home, total distance traveled, transition time, stationary time; use of various apps; incoming/outgoing calls/SMS; display on duration; - Also had wearable features, weather-related features - Reduced dimensionality to get 25 predictors - Two other models: only wearable features; only smartphone features	Residual HDRS scores	- Average ensemble of boosting - Random forest - Boruta algorithm to rank features	- 10-fold-cross-validation: ran 5 times for every model & split scenario (user-split or time-split), & averaged performance.	- MAE & RMSE of ML models compared to MAE & RMSE of estimates from: group median HDRS values, individual HDRS at screening, individual median HDRS from following visits. - Pearson correlation coefficient between true & estimated HDRS scores on test sets - Features ranked by importance	- All ML models performed similarly: MAE between 3.88-4.74, correlations between 0.46-0.7 - Time-split scenario: lowest MAE from model that included only mobile features - User-split scenario: all performed similarly, lowest MAE obtained by model using only wearable features - Correlation between observed, estimated depression in time-split model including features from mobile phone	- Uncertain if one modality outperforme d others - Activity level among most important features for estimating symptom severity

	important for estimating depressive symptom severity						(r = 0.7) was the strongest - Time-split: estimates from ML models more accurate than from group median & individual screening models, but not better than individual median model - Feature ranking: 39 features identified as important (54% from phone) - Related to: mobile phone engagement, activity level from mobile sensors, wearables - Top 5 features (ranked): average time phone screen on over 24 h, wearable feature, average SD of latitude & longitude from 12 p.m.-6 p.m, average duration phone screen on from 8 a.m.-6 p.m., average latitude over 24 h	
Pellegrini et al. (2021)	To predict a future MADRS score from	- GPS: number of significant locations visited, time spent at	MADRS score	Linear mixed models	Leave-one-subject-out cross-validation	Averaged RMSE across participants	- 1st PC explained 46% of variance (highest weights from GPS features)	- More work required to replace clinical raters

	baseline assessments of MADRS, PHQ on the phone, passively collected smartphone data, or combination of these.	home, distance traveled, maximum diameter, maximum home distance, radius of gyration, average & SD of flight length & duration, probability of pause, significant location entropy, circadian routine, weekend–weekday routine, missing data - Accelerometer: activity level, missing data - Android: number of outgoing calls, unique numbers dialed - PCA on all features, 1st PC used as predictor - Models with & w/out PHQ-8 surveys - All models: included baseline MADRS & demographics, then omitted baseline MADRS		- Including passive variables did not improve average RMSE, maybe due to varying data quality (NB only see separate results for MDD in graphs, not numerically stated results)	for MADRS assessment

Sverdlov et al. (2021)	To explain between-subject variation in MADRS, predict individual MADRS scores, using digital	biomarker data.	- Possible variables: total number of staypoints, number of places visited once, number of unique places visited, total amount of time spent at home, average distance from home, average repetition per contact for outgoing, incoming, missed calls, total number of WhatsApp calls, entropy and mean of usage time of apps, total usage count of apps - predictors used were most significant features found in variable selection step	Total MADRS score	Multiple linear regression model	Model fit assessed using plots of observed vs. predicted MADRS scores & using model residual plots	Proportion of variance in MADRS scores explained by model.	- Correlation r=0.43 between observed & predicted MADRS total scores	- Depression may be associated with decreased communication, less movement, more home stay - Models may be able to relate expected change in MADRS to change in digital phenotypes
Tønning et al. (2021)	To investigate associations between smartphone measures of social & physical activity,	- Clinical ratings: HDRS-17 - Daily smartphone-based patient-reportings of mood	Daily averages: Physical activity - number of steps, total distance moved,	Linear mixed-effects models accounting for repeated measurements within each participant.	-	-		- Higher HDRS-17 associated with fewer screen turn on's, positively associated with number of outgoing calls, duration of calls	Increase of total screen time, number of missed & incoming calls may show increase in

	phone usage, with HDRS-17 & with same-day smartphone-based patient-reportings		Smartphone usage - total screen-on time, number of times screen was turned on; social activity - number of incoming, outgoing, missed calls, duration of calls, number of incoming & outgoing text messages				- Patient-reported mood negatively associated with social activity & phone usage (screen time, call duration, number of incoming/missed calls, number of incoming/outgoing text messages) - Association between automatically generated smartphone measures & clinical ratings: 4/30 statistically significant results, high risk of chance findings	sedentary behavior, social withdrawal, concern from others
Zhang et al. (2021)	- To explore associations between statistical Bluetooth features, nonlinear features (of complexity, regularity & periodicity) & PHQ-8 - To predict severity of depressive symptoms using these	- 49 Bluetooth features - 3 categories: second-order statistics, multiscale entropy, frequency domain - Last PHQ-8 score - Covariates: age, gender, education years	Depressive symptom severity measured using PHQ-8.	- Pairwise linear mixed-effect models (random participant intercepts) - Hierarchical Bayesian linear regression - LASSO regression - XGBoost regression	Leave-all-out & leave-one-out time-series cross-validation	- Linear mixed-effect models: z-test to evaluate significance. Benjamini-Hochberg method to correct for multiple comparisons, adjusted P value threshold set to .05 - Prediction models: RMSE & predicted coefficient of determination	Linear mixed-effect models: - 10 second-order statistical features negatively associated with PHQ-8 - Minimum value of daily maximum NBDC in past 14 days had strongest association, four features related to daily variance of NBDC all negatively associated	- Decreased mobility & engagement in activities may be why participants with higher PHQ-8 have lower variance-related features - Could predict depression for participants

	features & ML models		- MSE at scales 1, 2, 3 positively associated with PHQ-8; at scales 16, 22 negatively associated - Five FD features associated with PHQ-8 - Spectrum power of three frequency intervals negatively associated with PHQ-8 - MF_sum had strongest association - HF spectral entropy positively associated with PHQ-8 - Second-order statistical Bluetooth features model & all Bluetooth features model fitted data significantly better than model without Bluetooth features. Prediction models: - LAO: Hierarchical Bayesian linear regression model using all Bluetooth features had better performance (R=0.526, RMSE=3.891) than	with few observations, but more accurate when more training data available

							LASSO & XGBoost models - LOO: Hierarchical Bayesian linear model with all Bluetooth features had best performance (R=0.387, RMSE=4.426), but XGBoost model had similar performance	
Zhang et al. (2022)	To explore the relationships and the direction of the relationships between depressive symptom severity and phone-measured mobility over time; to investigate effects of individual differences on models.	- Covariates: age, gender, work status	- Each PHQ-8 score & mobility feature: Location Variance, Moving Time/Distance, Number of Clusters, Location/ Normalized Entropy, Homestay, Residential Location Count, Long-term/ Circadian/ Short-term Rhythm	- Vector autoregressive models	-	Adjusted significance threshold P<.05	- Lagged effects: positive lagged effects in PHQ-8 & mobility features - Cross-lagged effects: PHQ-8 negatively correlated with following circadian rhythm of mobility, and preceding location entropy, homestay. Residential location count positively correlated with following PHQ-8 (despite negative correlation at within-individual level). Individual differences: - Older participants had lower lagged effects for moving distance, homestay	- Increase in PHQ-8 probably preceded by less time spent at different places, more time at home, or more traveling - Traveling may reduce depressive symptoms, but worsen depressive feelings

			- Female participants had lower lagged effects on location entropy, residential location count - Employed participants had lower lagged effects on PHQ-8, higher lagged effects on normalized entropy - Cross-lagged effects: age negatively correlated with a circadian rhythm coefficient

Table 5. Methodological Details Of Studies Predicting Diagnostic Class Or Mood State/Episode Label

Author (year of publication)	Goal of analysis	Predictor variables	Response variable/s	Modelling techniques	Model selection strategy	Performance metrics used to select model	Results from model selection	Internal validation	Internal validation metrics used	Results	Interpretation of the final model	Clinical implications
Bai et al. (2021)	Examine feasibility of monitoring mood status and stability of patients with MDD using ML models and passively collected data.	- For each type of call (incoming, outgoing, rejected) and for all phone calls during each period: time of call being made, duration, number of different people the call was made to/from, entropy of callers - Average & median of phone usage duration & number of times used (also ratio for different	2 groups and 4 subgroups: - Steady (Remission, Depressed) - Swing (Drastic, Moderate)	Support vector machines, K-nearest neighbors, decision trees, naïve Bayes, random forest, logistic regression	Parameter tuning not presented	-	-	10-fold cross-validation	Average accuracy rate and recall rate of all folds	- Selected phone features (including app usage & call logs): classification between Steady-remission & Swing-moderate had highest accuracy (80.92%) & recall (95.50%). Classification between Steady-depressed & Swing-drastic had lowest accuracy (66.18%) & recall (65.71%). Classification between all Steady samples & all Swing samples	- Prediction accuracies between Steady-remission & Mood Swing better than those between Steady-depressed & Mood Swing (drastic and moderate) - Patients with ongoing depressed symptoms may have similar behaviour patterns to those with mood swings - Differences in	Features chosen by model reflect some of the MDD symptoms: low sleep quality, reduced social interaction, reduced physical activity

	times of day) - Average, SD & entropy of app usage duration & number of times used for 8 app categories (also ratio for different times of day) - 8 categories: Instant messaging, social networking, shopping, entertainment, music, food delivery, others, all apps - Difference, mean value, SD of each feature calculated for each sample		had 66.76% accuracy & 80.93% recall - Selected features (including from wearables): classification between Steady-remission & Swing-moderate had highest accuracy (80.56%) & recall (97.08%). Classification between Steady-remission & Swing-drastic had lowest accuracy (74.29%) & recall (84.31%). Classification between all Steady samples & all Swing samples had 76.67% accuracy & 90.44% recall - Selected features of call	behaviour between patients in remission & those with mood swings may be more significant - Best combination used call logs, sleep, step count, heart rate - Models using features from all collected data had lower accuracies than those using features from all collected data except for app usage - Models using phone data (including app usage & calls) had lowest

		- Wearable device features used in model combining smartphone and wearable data								logs, sleep, step count & heart rate: classification between Steady-depressed & Swing-drastic had highest accuracy (84.27%) & recall (85.33%). Classification between all Steady samples & all Swing samples had lowest accuracy (75.64%) & recall (89.93%)	overall accuracies - NB low prevalence of Steady-remission class.	
Cho et al. (2019)	Determine whether mood states/episodes can be predicted using only automatically recorded data, w/o any mood information, by ML.	- Light exposure during bedtime, light exposure during daytime - Mean, SD, gradient calculated - Also used Fitbit features for steps,	- Classes for mood state prediction within next 3 days: biased, neutral - Classes for mood episode prediction:	Random forest	- Tuned values for time period of days for model training (p) and for model testing (q) - Repeated experim	AUC	p=18, q=3	- Repeated training/testing evaluations by moving timepoint split from start to end of timeline - Evaluation of Mood State Prediction Model (Neutral or	Average sensitivity, specificity, accuracy, AUC	- Mood State Prediction (Neutral or Biased): Mood state labeling with 10, 30 and 50% cut-off for patients with MDD: accuracy 61, 67, 65%; sensitivity 61, 39, 57%; specificity 42, 74, 68%; AUC 56, 67, 69%	- In model selection step found that future mood prediction more successful for shorter period - Light exposure showed clear differences according	- For stable mood state, may be beneficial for patients to coordinate light exposure with their circadian rhythm

	sleep, heart rate	depressive, manic, hypomanic, none	ents for p ranging from 3-300 days and q from 3-30 days		Biased): used 10%, 30%, 50% cut-off thresholds between mood states. Tested performance for each threshold - Compared accuracy between general & individual models for: mood state for next 3 days, depressive episode, manic episode, hypomanic episode	- Mood Episode Prediction Model: for patients with MDD (No Episode/Depressive Episode), accuracy 75.1%/71.2%, sensitivity 93.5%/40.9%, specificity 39.5%/87.8%, AUC 78.1/79.8% - Future mood state: every personalised model outperformed general model - Episode prediction: personalised models almost always outperformed general model (NB no separate results for MDD) - Models with missing features removed (NB numerical	to mood state	- Mood prediction algorithm may help manage mood and avoid more severe episodes

										results not separately reported for MDD): AUC 0.684 without steps features, 0.687 without sleep features, 0.683 without heart rate features, 0.683 without light-related features		
Faurholt-Jepsen et al. (2022)	Investigate use of passively collected location data in classifying BD and UD.	- Daily features: number of stops, total duration of stops, number of places, number of moves, total distance moved, total duration of moves, routine index, radius of gyration, location entropy	- Classes: BD or UD across depressive or euthymic states	Balanced bagging classifier (ensemble of decision trees) used to account for imbalanced group labels	Parameter tuning not presented	-	-	10-fold stratified cross-validation	- Sensitivity, specificity, positive predictive value and negative predictive value, ROC AUC	- Classifying patients with UD during depressive state vs patients with BD during depressive state (sensitivity, specificity, PPV, NPV, AUC): 0.70, 0.77, 0.68, 0.78, 0.79 - Classifying patients with UD overall versus patients with BD overall (sensitivity, specificity, PPV, NPV, AUC): 0.70,	- Patients with BD during both depressive & euthymic states classified with higher AUC than patients with UD during depressive & euthymic states	- Location entropy may be useful ‘digital activity marker’ to discriminate between BD & UD - May help in achieving earlier, correct diagnosis - But not yet sure which features

										0.65, 0.72, 0.64, 0.75		most useful ‘digital markers’ of BD or UD
Sverdlov et al. (2021)	Build classifiers of depressed and healthy subjects using models with select digital biomarkers	- Possible variables: total number of staypoints, number of places visited once, number of unique places visited, total amount of time spent at home, average distance from home, average repetition per contact for outgoing, incoming, missed calls, total number of WhatsApp calls, entropy and mean	- Classes: depressed or healthy	- Logistic regression - Threshold approach based on predicted MADRS score using multiple linear regression model with selected digital biomarkers	Used stepwise variable selection method	Threshold to select significant predictors: $p<0.1$	- Logistic regression method: Total count of communication apps usage - Linear regression method: number of places visited once, number of unique places visited, average distance from home, average repetition per contact for outgoing, incoming and missed calls, total count of communica	Leave-one-out cross validation	- Sensitivity, specificity, overall classification accuracy, ROC AUC	- Logistic model (Sensitivity, Specificity, Accuracy, AUC): 0.64, 0.64, 0.64, 0.72 - Linear regression model (Sensitivity, Specificity, Accuracy, AUC): 0.86, 0.64, 0.75 , 0.83	- Linear model-based classifier generally achieved better performance than logistic classifier	-

	of usage time of apps, total usage count of apps - Logistic regression method: used most significant predictors from variable selection - Linear regression method: used MADRS score predicted from multiple linear regression model using most significant predictors	tion apps usage	

Table 6. Methodological Details Of Studies With Other Analysis Goals

Author (year of publication)	Goal of analysis	Predictor variables	Response variable/s	Modelling techniques	Quality metrics used to assess performance	Results
Braund et al. (2022)	Investigate differences in circadian rhythm between participants with MDD & BD	Diagnostic group	Circadian rhythm	2-tailed t-tests	Significance threshold p<.05	No difference found
Emden et al. (2020)	To evaluate feasibility & adherence of ReMAP in transdiagnostic sample.	Diagnosis, age, sex, depression severity, global level of functioning, previous psychiatric hospitalizations, verbal IQ, education	Participation in study (yes or no), two-week retention, one-year retention, participation duration, rate of days with passive events	- Kaplan-Meier survival analysis, multiple standard regression analyses, Cox proportional hazard regression analysis - Random forest (used inner 5-fold Cross Validation for parameter optimisation, outer 5-fold Cross Validation for evaluation)	- Significance threshold p<.05 - Mean balanced accuracy	- HC, MDD, BD, AD, PD app users did not differ wrt total participation duration or rate of passive data transfer - Random forest: no significant prediction for participation or retention

Matcham et al. (2022)	To examine whether depressed mood is associated with availability of data	People who provided passive data throughout follow-up divided into: those who provided 0–25%, 26–50%, 51–75, >75% of expected data	Participants who are symptomatic at baseline & those who have no/mild symptoms	Chi-squared tests	P-value significance threshold not stated	- Most data available for GPS location & battery level; least data available for phone usage - No evidence of difference in data availability between those with and w/o depression at baseline
Zhang et al. (2022)	To explore the relationships and the direction of the relationships between depressive symptom severity and phone-measured mobility over time; to investigate effects of individual differences on models.	Investigated direction of relationships between mobility features (location variance & entropy, normalized entropy, moving time & distance, number of clusters, homestay, residential location count, long- & short-term rhythm, circadian Rhythm) & PHQ-8	-	Dynamic structural equation modeling: considered lagged & cross-lagged effects between time point & following time point	Adjusted P-value threshold <.05	DSE model (with all mobility features) failed to converge