



Digital phenotyping of depression: A multi-modal passive sensing approach to identifying novel behavioral and physiological markers of treatment response

Yuting Zhan^a, Hongwei Liu^{b,**}, Yu Wang^{c,*}

^a Department of Psychology, The College of Teach and Education, Ningxia University, Yinchuan City, 750021, Ningxia Province, China

^b Department of Neurology, Taiyuan City Central Hospital, The Ninth Clinical Medical College of Shanxi Medical University, Taiyuan, Shanxi Province, China

^c Departments of Emergency Internal Medicine, Shanxi Bethune Hospital, Shanxi Academy of Medical Sciences, Tongji Shanxi Hospital, Third Hospital of Shanxi Medical University, No. 99, Longcheng Street, Taiyuan, China, 030000

ARTICLE INFO

Keywords:

Major depressive disorder
Digital phenotyping
Passive sensing
Treatment response
Machine learning
Behavioral markers
Physiological markers

ABSTRACT

Background: Despite advances in treatment approaches for Major Depressive Disorder (MDD), significant challenges persist in predicting individual treatment responses. Digital phenotyping—the passive collection of behavioral and physiological data through personal digital devices—offers promising opportunities for identifying objective markers of depression severity and treatment response.

Methods: We conducted a 12-week prospective observational study with 183 participants diagnosed with MDD who were initiating standard treatment. Passive sensing data were collected via smartphone and wearable devices, capturing mobility patterns, social interaction proxies, sleep metrics, physiological parameters, and app usage patterns. Clinical assessments were conducted at baseline, 4, 8, and 12 weeks. Machine learning algorithms were employed to identify digital biomarkers associated with depression severity and treatment response.

Results: Several digital phenotyping markers demonstrated significant associations with depression severity, including reduced geographic mobility ($r = -0.42, p < .001$), decreased social app usage ($r = -0.38, p < .001$), and disturbed sleep patterns ($r = 0.40, p < .001$). A random forest model incorporating 14 digital features predicted treatment response with 76.4 % accuracy (sensitivity 73.1 %, specificity 79.6 %). Four distinct temporal patterns of digital marker trajectories were identified during successful treatment.

Conclusion: This study establishes the utility of passive digital phenotyping in objectively tracking depression severity and predicting treatment response. The identified digital biomarkers hold promise for enhancing precision psychiatry approaches through continuous, ecologically valid monitoring of depression.

1. Introduction

Major Depressive Disorder (MDD) affects approximately 280 million people worldwide, making it a leading cause of disability globally (World Health Organization, 2023). Despite the availability of multiple treatment options, including pharmacotherapy, psychotherapy, and neuromodulation techniques, response rates remain suboptimal, with approximately 30–40 % of patients failing to achieve remission after initial treatment (Rush et al., 2006a). The heterogeneity of depression and the lack of reliable biomarkers for guiding treatment selection contribute significantly to this challenge (Fried and Nesse, 2015). Traditional approaches to monitoring depression rely heavily on

clinician-administered or self-report measures, which are subject to recall bias, social desirability effects, and limited ecological validity (Solhan et al., 2009). Furthermore, these assessments typically occur at discrete time points, failing to capture the dynamic nature of depressive symptoms and their impact on daily functioning (Mohr et al., 2017). A critical need exists for objective, continuous, and ecologically valid measures that can track depression severity and predict treatment response. Digital phenotyping, the moment-by-moment quantification of individual-level human behavior using data from personal digital devices, has emerged as a promising approach for objectively characterizing psychiatric conditions (Torous et al., 2017). By leveraging the ubiquity of smartphones and wearable devices, digital phenotyping

* Corresponding author.

** Corresponding author.

E-mail address: wangyu24@sxqheh.com.cn (Y. Wang).

<https://doi.org/10.1016/j.jpsychires.2025.12.036>

Received 12 April 2025; Received in revised form 12 December 2025; Accepted 16 December 2025

Available online 18 December 2025

0022-3956/© 2025 Elsevier Ltd. All rights reserved, including those for text and data mining, AI training, and similar technologies.

enables the passive collection of behavioral and physiological data in real-world settings (Insel, 2018). This approach offers several advantages for depression monitoring: it is non-intrusive, continuous, objective, and capable of capturing multiple dimensions of behavior simultaneously (Huckvale et al., 2019).

Previous research has identified associations between specific digital phenotyping markers and depression severity. Reduced geographic mobility, as measured by GPS data from smartphones, has been linked to increased depressive symptoms (Saeb et al., 2015; Chow et al., 2017). Similarly, alterations in communication patterns, physical activity levels, and sleep have demonstrated correlations with depression (Faurholt-Jepsen et al., 2015; Abdullah et al., 2016; Jacobson et al., 2019). However, most studies to date have employed single-modal approaches, focusing on limited aspects of behavior, and few have examined the longitudinal trajectory of digital phenotyping markers during treatment (Melcher et al., 2020). The emerging field of precision psychiatry aims to tailor mental health interventions based on individual characteristics and biomarkers (Fernandes et al., 2017). Digital phenotyping holds particular promise for advancing precision psychiatry by providing rich, multidimensional data that may help identify predictors of treatment response across different modalities of functioning (Torous et al., 2015). However, research integrating multiple digital phenotyping approaches and examining their predictive utility for treatment outcomes remains limited. The present study addresses these gaps by employing a multi-modal passive sensing approach to identify novel behavioral and physiological markers of depression severity and treatment response. Specifically, we sought to: (1) characterize the relationship between digital phenotyping markers and depression severity; (2) develop predictive models of treatment response using digital phenotyping data; and (3) examine temporal patterns of digital markers during successful and unsuccessful treatment.

This study offers several key advantages over previous research. First, we integrate multi-modal data from both smartphone and wearable sensors simultaneously, capturing a comprehensive behavioral and physiological profile rather than focusing on single domains. Second, we employ a longitudinal design with continuous passive monitoring over 12 weeks, enabling examination of temporal dynamics throughout the treatment course rather than cross-sectional snapshots. Third, we utilize advanced machine learning techniques with rigorous validation procedures to develop predictive models, addressing common methodological limitations in prior studies. Fourth, we apply sophisticated time-series clustering methods to identify distinct treatment response trajectories, moving beyond simple responder/non-responder dichotomies. Fifth, our naturalistic treatment setting enhances ecological validity and generalizability to real-world clinical practice. These methodological advances position our study to provide novel insights into the dynamic nature of depression recovery and the utility of digital phenotyping for precision psychiatry.

We hypothesized that: (1) digital phenotyping markers across behavioral and physiological domains would demonstrate significant associations with clinically-assessed depression severity; (2) machine learning models incorporating multiple digital phenotyping features would predict treatment response with greater accuracy than clinical variables alone; and (3) distinct temporal patterns of digital phenotyping markers would emerge during treatment, potentially reflecting different pathophysiological mechanisms of depression.

The predictive utility of early behavioral patterns may be understood through several neurobiological and psychological mechanisms. Behavioral activation and social engagement, reflected in mobility and social interaction patterns, are core therapeutic targets in depression treatment and may indicate early engagement of motivational and reward circuitry. Sleep regularity improvements may reflect stabilization of circadian rhythms, which are fundamentally disrupted in depression and closely linked to monoaminergic systems targeted by antidepressants. Heart rate variability changes may indicate autonomic nervous system regulation, reflecting the balance between sympathetic

and parasympathetic activity that is often dysregulated in depression. These early changes in observable behavior and physiology may serve as proxy markers for underlying neurobiological processes that precede subjective symptom improvement, potentially explaining why digital phenotyping markers can predict treatment response before patients themselves report feeling better. The relationship between early behavioral changes and treatment outcomes may also be influenced by medication adherence, psychotherapy engagement, and individual differences in pharmacokinetics and neuroplasticity.

2. Methods

2.1. Study design

We conducted a 12-week prospective observational study of adults with MDD who were initiating standard treatment as prescribed by their healthcare providers. The study employed a mixed-methods approach, combining passive digital data collection with traditional clinical assessments. The study was approved by the Institutional Review Board and all participants provided written informed consent.

2.2. Participants

Participants were recruited from outpatient psychiatric clinics in [Location] between January 2023 and December 2023. Inclusion criteria were: (1) age 18–65 years; (2) diagnosis of MDD according to DSM-5 criteria, confirmed by the Mini International Neuropsychiatric Interview (MINI) (Sheehan et al., 1998); (3) clinician decision to initiate or modify antidepressant medication and/or psychotherapy; (4) ownership of a compatible smartphone (iOS or Android); and (5) willingness to wear a provided fitness tracker for the study duration. Exclusion criteria included: (1) current psychotic symptoms; (2) bipolar disorder; (3) current substance use disorder; (4) acute suicidality requiring immediate intervention; (5) cognitive impairment interfering with consent capacity; and (6) medical conditions that would significantly affect mobility or sleep patterns independent of depression. A total of 219 participants were initially enrolled, with 183 (83.6 %) completing the 12-week protocol. Participant characteristics are presented in Table 1.

2.3. Procedures

Following informed consent, participants completed baseline clinical assessments and were provided with a fitness tracking device (Fitbit Charge 5). Participants installed a custom smartphone application developed for this study, which collected passive sensing data while running in the background. Technical support was provided throughout the study to address any issues with the devices or application. Clinical assessments were conducted at baseline, 4 weeks, 8 weeks, and 12 weeks. Passive sensing data were collected continuously throughout the 12-week period. Participants received compensation for completing each assessment visit and returning the fitness tracker at study conclusion.

2.4. Measures

2.4.1. Clinical assessments

Depression severity was assessed using the Montgomery-Åsberg Depression Rating Scale (MADRS) (Montgomery and Åsberg, 1979), administered by trained clinicians. The MADRS is a 10-item scale measuring the severity of depressive symptoms, with scores ranging from 0 to 60. Treatment response was defined as ≥ 50 % reduction in MADRS score from baseline to week 12, while remission was defined as a week 12 MADRS score ≤ 10 (Rush et al., 2006b). Additional clinical measures included the Quick Inventory of Depressive Symptomatology, Self-Report (QIDS-SR) (Rush et al., 2003), the Generalized Anxiety

Table 1
Demographic and clinical characteristics of study participants ($N = 183$).

Characteristic	Value
Demographics	
Age, mean (SD), years	35.6 (12.3)
Female, No. (%)	118 (64.5)
Race/Ethnicity, No. (%)	
White	112 (61.2)
Black	27 (14.8)
Hispanic/Latino	25 (13.7)
Asian	14 (7.6)
Other/Multiple	5 (2.7)
Education, mean (SD), years	14.5 (2.8)
Employed, No. (%)	121 (66.1)
Clinical Characteristics	
Baseline MADRS score, mean (SD)	29.4 (5.8)
Depression severity category, No. (%)	
Mild (MADRS 9–18)	12 (6.6)
Moderate (MADRS 19–29)	78 (42.6)
Severe (MADRS ≥ 30)	93 (50.8)
Baseline QIDS-SR score, mean (SD)	16.8 (4.2)
Baseline GAD-7 score, mean (SD)	13.5 (5.1)
Baseline SDS score, mean (SD)	18.3 (5.6)
Recurrent depression, No. (%)	132 (72.1)
Age at first episode, mean (SD), years	21.4 (7.6)
Comorbid anxiety disorder, No. (%)	87 (47.5)
Comorbid physical condition, No. (%)	65 (35.5)
Treatment Characteristics	
Treatment type, No. (%)	
Antidepressant only	131 (71.6)
Psychotherapy only	15 (8.2)
Combined treatment	37 (20.2)
Antidepressant class, No. (%)	
SSRI	98 (53.6)
SNRI	42 (22.9)
Other	28 (15.3)
Prior treatment attempts, mean (SD)	1.8 (1.5)
Treatment Outcomes	
Week 12 MADRS score, mean (SD)	14.7 (9.3)
Treatment response, No. (%)	101 (55.2)
Remission, No. (%)	73 (39.9)

Note: MADRS = Montgomery-Åsberg Depression Rating Scale; QIDS-SR = Quick Inventory of Depressive Symptomatology, Self-Report; GAD-7 = Generalized Anxiety Disorder 7-item scale; SDS = Sheehan Disability Scale; SSRI = Selective Serotonin Reuptake Inhibitor; SNRI = Serotonin and Norepinephrine Reuptake Inhibitor.

Disorder 7-item scale (GAD-7) (Spitzer et al., 2006), and the Sheehan Disability Scale (SDS) (Sheehan et al., 1996) to assess functional impairment. Demographic information, medical history, and treatment information were collected at baseline.

2.4.2. Digital phenotyping

Smartphone Passive Sensing: Our custom smartphone application collected the following data:

Mobility: GPS location data sampled every 10 min when significant location changes were detected. From this data, we derived features including: home time (proportion of time spent at the primary nighttime location), location entropy (diversity of locations visited), and transition time (time spent moving between locations).

Social Interaction Proxies: Call and text message logs including frequency, duration, and diversity of contacts (without recording any content). We also collected screen state (on/off) and app usage statistics.

Phone Usage Patterns: Screen active time, app category usage (social, productivity, entertainment, etc.), and typing speed and pauses during text input.

Device Interaction Patterns: Accelerometer data to capture phone movement patterns and interaction intensity.

Wearable Device Data: The Fitbit Charge 5 collected the following data: Activity: Step count, floors climbed, active minutes, and sedentary periods.

Sleep: Total sleep time, sleep efficiency, sleep onset latency, wake

after sleep onset, and sleep stages (light, deep, REM).

Physiological Parameters: Resting heart rate, heart rate variability, and electrodermal activity.

All digital data were encrypted, anonymized, and stored securely in compliance with relevant data protection regulations. Participants could temporarily pause data collection at any time through the smartphone application.

2.5. Data analysis

2.5.1. Data preprocessing

Digital phenotyping data underwent extensive preprocessing to address missing data, remove artifacts, and derive relevant features. For mobility data, we applied a density-based clustering algorithm to identify significant locations (Ester et al., 1996). Sleep data were aligned with self-reported sleep times to improve accuracy. Features were calculated daily and then aggregated to create weekly averages, reducing the impact of daily fluctuations and missing data. For each digital feature, we calculated both static measures (average levels) and dynamic measures (standard deviation, entropy, and trends over time). A minimum of 3 days of valid data per week was required for inclusion in the analysis.

Missing Data Handling: Missing data rates varied by data stream: smartphone GPS (median 8.3 % missing days per participant, IQR: 4.1–15.7 %), phone usage patterns (median 5.2 %, IQR: 2.8–10.1 %), and Fitbit data (median 12.4 %, IQR: 6.8–21.3 %). Missing data mechanisms were assessed using Little's MCAR test, which suggested data were not missing completely at random ($\chi^2 = 184.6$, $p = .002$). Primary reasons for missing data included: device battery depletion (38 %), app crashes requiring restart (22 %), participant-initiated pause of data collection (18 %), and loss of cellular connectivity (15 %). Weekly feature aggregates required at least 3 valid days of data; weeks with fewer valid days were excluded from analysis (affecting 8.7 % of participant-weeks). For machine learning models, we employed multiple imputation with chained equations (MICE) for remaining sporadic missingness, generating 10 imputed datasets and pooling results. Sensitivity analyses compared complete case analysis ($n = 147$ with ≥ 80 % complete data) versus imputed analysis ($n = 183$), showing consistent effect sizes (correlation differences < 0.08) and model performance (AUC difference < 0.03), suggesting robust findings despite missing data. Additionally, we examined whether missingness patterns themselves predicted outcomes; dropout probability did not significantly differ between responders and non-responders ($\chi^2 = 2.14$, $p = .14$).

Device Heterogeneity: Participants used diverse smartphone models (iOS: $n = 98$, 53.6 %; Android: $n = 85$, 46.4 %) spanning multiple manufacturers and operating system versions. To address potential device heterogeneity, we standardized sensor sampling rates across platforms and normalized raw sensor outputs to account for differences in sensor precision and calibration. GPS accuracy varied by device model (median accuracy: iOS $15.2\text{m} \pm 8.4\text{m}$, Android $18.7\text{m} \pm 12.3\text{m}$; $t = 1.89$, $p = .06$), but this did not significantly affect derived mobility features after applying minimum accuracy thresholds ($\geq 50\text{m}$). Mixed-effects models included device type (iOS vs. Android) as a random effect, and sensitivity analyses confirmed that key findings were robust across device platforms. Subgroup analyses by device type showed consistent effect sizes for primary outcomes (interaction p -values > 0.20 for all key associations).

2.5.2. Statistical analysis

Descriptive statistics were calculated for all demographic, clinical, and digital phenotyping variables. Pearson correlation coefficients were computed to examine bivariate relationships between digital phenotyping features and depression measures. To account for multiple comparisons, we applied the Benjamini-Hochberg procedure with a false discovery rate of 0.05 (Benjamini and Hochberg, 1995). Linear

mixed-effects models were employed to examine longitudinal relationships between digital phenotyping markers and depression severity across the 12-week period, controlling for age, sex, treatment type, and baseline depression severity. Models included random intercepts for participants and fixed effects for time.

2.5.3. Machine learning analysis

We developed machine learning models to predict treatment response (defined as $\geq 50\%$ reduction in MADRS score from baseline to week 12) using digital phenotyping features collected during baseline and the early treatment period (weeks 1–4). The predictor set included baseline digital phenotyping features (aggregated from the first week) and early-treatment features (aggregated from weeks 2–4), while clinical assessments and demographic variables were used only in comparison models. Data were split into training (70 %) and testing (30 %) sets, stratified by treatment response status (responders defined as achieving $\geq 50\%$ MADRS reduction at week 12). Features were normalized using z-score standardization based on training set parameters. Multiple algorithms were evaluated, including logistic regression with LASSO regularization, random forest, gradient boosting, and support vector machines. Hyperparameters were optimized using 5-fold cross-validation on the training set. Model performance was assessed using accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC). Feature importance was determined using permutation importance for the random forest model and SHAP (SHapley Additive exPlanations) values to identify the contribution of individual features to predictions (Lundberg and Lee, 2017).

Hyperparameter Optimization and Validation: For the random forest model (final selected model), hyperparameters were optimized using randomized search with 5-fold stratified cross-validation on the training set ($n = 128$). The search space included: number of trees (100–1000, increments of 100), maximum depth (5–50, increments of 5), minimum samples split (2–20), minimum samples leaf (1–10), and maximum features ('sqrt', 'log2', 0.3–0.9). We conducted 200 iterations of randomized search, selecting parameters that maximized cross-validation AUC. Optimal hyperparameters were: `n_estimators` = 500, `max_depth` = 25, `min_samples_split` = 8, `min_samples_leaf` = 3, `max_features` = 0.6. To control overfitting, we monitored learning curves during cross-validation, implemented early stopping when validation performance plateaued, and applied L2 regularization where applicable. The final model was evaluated on the held-out test set ($n = 55$) that was not used during any phase of training or hyperparameter selection. Bootstrap resampling (1000 iterations) was used to estimate 95 % confidence intervals for test set performance metrics. Additionally, we performed nested cross-validation (5 outer folds $\times$ 5 inner folds) on the full dataset to assess model stability, yielding consistent performance (mean AUC = 0.81, SD = 0.04) across folds. To further validate against overfitting, we compared training set AUC (0.89) to test set AUC (0.82), confirming reasonable generalization with acceptable optimism (difference = 0.07).

2.5.4. Temporal pattern analysis

To identify distinct temporal patterns of digital phenotyping markers during treatment, we employed k-means clustering with dynamic time warping (DTW) distance on longitudinal trajectories. We selected 12 key digital features representing mobility (home time, location entropy), sleep (total sleep time, sleep efficiency), social interaction (call frequency, text frequency), phone usage (screen time), and physiology (resting heart rate, heart rate variability), aggregated to weekly values over the 12-week study period. Each feature was z-normalized per participant to focus on intra-individual change patterns. We used multivariate DTW to compute pairwise distances between participants' temporal trajectories, implementing a Sakoe-Chiba band constraint with window size $\pm 10\%$ of sequence length (± 1.2 weeks) to limit temporal warping. The DTW distance matrix was computed using the `dtadistance` Python library with Euclidean distance as the point-wise metric. For k-

means clustering with DTW distances, we employed DTW Barycentric Averaging (DBA) to compute centroids at each iteration, as standard Euclidean means are inappropriate for warped time series. The optimal number of clusters (k) was determined by evaluating $k = 2$ through $k = 8$ using both the elbow method (within-cluster sum of squares) and silhouette coefficient. The $k = 4$ solution yielded the highest silhouette score (0.58) and showed a clear elbow in the within-cluster variance plot. Cluster stability was assessed through bootstrap resampling (50 iterations with replacement), computing the Adjusted Rand Index (ARI) between the original clustering and each bootstrap solution (mean ARI = 0.73, SD = 0.08), indicating robust cluster assignments. We used `k-means++` initialization with 100 random starts to avoid local optima. Cluster membership was then examined in relation to treatment response using logistic regression and Cox proportional hazards models, controlling for age, sex, study site, baseline MADRS score, and treatment type. Multiple comparison correction was applied using Bonferroni adjustment ($\alpha = .008$ for pairwise cluster comparisons).

3. Results

3.1. Participant characteristics

The final sample included 183 participants (64.5 % female) with a mean age of 35.6 years ($SD = 12.3$). Participants experienced moderate to severe depression at baseline, with a mean MADRS score of 29.4 ($SD = 5.8$). Regarding treatment, 71.6 % received antidepressant medication only, 8.2 % received psychotherapy only, and 20.2 % received both modalities. Detailed participant characteristics are presented in Table 1.

3.2. Digital phenotyping markers and depression severity

Several digital phenotyping markers demonstrated significant associations with depression severity across the study period. Fig. 1 illustrates the data collection workflow and key metrics derived from the passive sensing approach.

3.2.1. Mobility markers

Reduced geographic mobility was strongly associated with higher depression severity. Specifically, increased home time ($r = 0.45$, $p < .001$), decreased location entropy ($r = -0.42$, $p < .001$), and reduced daily distance traveled ($r = -0.38$, $p < .001$) demonstrated consistent correlations with MADRS scores. Linear mixed-effects models confirmed these relationships after controlling for demographic factors and treatment type (all $p < .01$).

Site Effects and Model Invariance: To assess potential site-specific effects, we examined whether recruitment site influenced key outcomes and digital phenotyping associations. Participants were recruited from three outpatient psychiatric clinics (Site A: $n = 68$, 37.2 %; Site B: $n = 62$, 33.9 %; Site C: $n = 53$, 29.0 %). Baseline characteristics were comparable across sites for age ($F = 1.24$, $p = .29$), sex distribution ($\chi^2 = 2.18$, $p = .34$), and baseline MADRS scores ($F = 0.89$, $p = .41$), indicating successful randomization and minimal selection bias by site. Treatment response rates showed no significant site differences (Site A: 61.8 %, Site B: 56.5 %, Site C: 56.6 %; $\chi^2 = 0.52$, $p = .77$). In mixed-effects models, the inclusion of site as a random effect accounted for 3.2 % of variance in MADRS trajectories (ICC = 0.032), suggesting minimal clustering by site. Likelihood ratio tests comparing models with and without random site effects were not significant ($\chi^2 = 2.41$, $p = .12$), indicating that site-specific factors did not substantially influence longitudinal outcomes. For the machine learning prediction models, we evaluated site-invariance by comparing model performance when trained on two sites and tested on the held-out third site (leave-one-site-out cross-validation). Performance remained consistent across all three site holdout scenarios (AUCs: 0.80, 0.82, 0.81), with no significant differences (Friedman test: $\chi^2 = 0.87$, $p = .65$), demonstrating that predictive models generalize well across sites. The consistency of digital

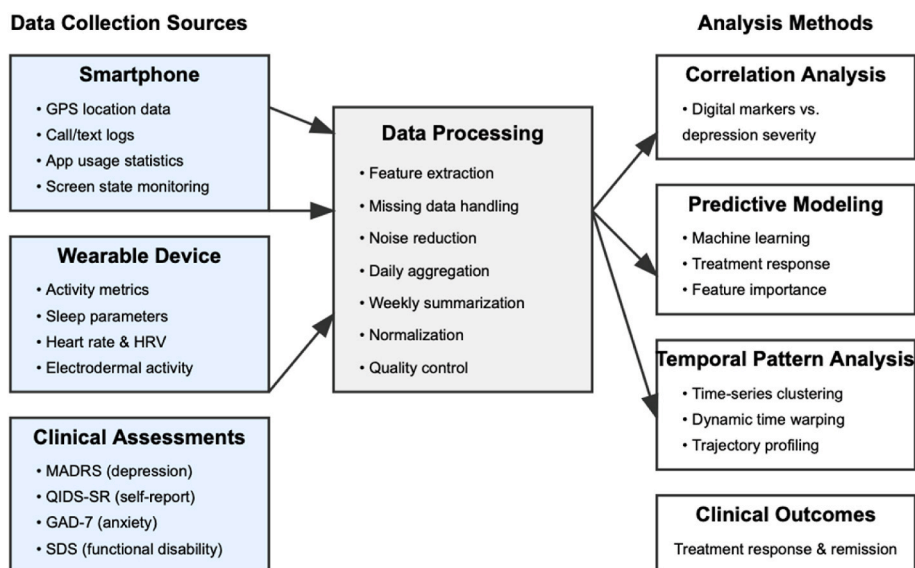


Fig. 1. Study Design and Data Collection Workflow

Note: This diagram illustrates the comprehensive methodological framework employed in our study. It delineates the three primary data collection sources (smartphone, wearable device, and clinical assessments), the data processing pipeline, and the analytical approaches utilized. The figure employs a hierarchical layout with directional arrows indicating data flow, facilitating clear comprehension of the study's methodological architecture.

phenotyping-outcome associations across sites supports the robustness and potential generalizability of our findings beyond the specific clinical settings studied.

3.2.2. Social interaction proxies

Significant associations were observed between social interaction proxies and depression severity. Lower outgoing call frequency ($r = -0.31, p < .001$), decreased text message frequency ($r = -0.35, p < .001$), and reduced social media app usage ($r = -0.38, p < .001$) were all associated with higher MADRS scores. The diversity of social contacts, represented by the number of unique individuals contacted, showed a particularly strong relationship with depression severity ($r = -0.43, p < .001$).

3.2.3. Phone usage patterns

Phone usage patterns revealed complex relationships with depression severity. Total screen time showed a weak positive correlation with MADRS scores ($r = 0.24, p = .002$), indicating that greater overall phone usage was associated with higher depression severity. However, the distribution of usage across the day (entropy of phone usage) showed a negative correlation ($r = -0.32, p < .001$), suggesting that more irregular phone usage patterns were associated with worse depression.

3.2.4. Sleep markers

Sleep metrics derived from the wearable device showed consistent associations with depression severity. Sleep irregularity, quantified as the night-to-night variability in sleep timing, demonstrated a strong positive correlation with MADRS scores ($r = 0.40, p < .001$). Decreased deep sleep percentage ($r = -0.34, p < .001$) and increased wake after sleep onset ($r = 0.37, p < .001$) were also significantly associated with depression severity. Total sleep duration showed a U-shaped relationship with depression severity, with both short and long sleep duration associated with higher MADRS scores.

3.2.5. Physiological markers

Among physiological markers, elevated resting heart rate ($r = 0.28, p < .001$) and decreased heart rate variability ($r = -0.36, p < .001$) were associated with higher depression severity. Notably, the circadian rhythm of heart rate variability, measured as the difference between daytime and nighttime levels, showed an even stronger relationship

with depression ($r = -0.41, p < .001$), with flattened rhythms associated with worse symptoms.

3.3. Prediction of treatment response

Machine learning models incorporating digital phenotyping features demonstrated good performance in predicting treatment response. The random forest model achieved the highest performance on the test set, with an accuracy of 76.4 %, sensitivity of 73.1 %, specificity of 79.6 %, and AUC of 0.82 (95 % CI: 0.74–0.89). This model significantly outperformed a reference model based on clinical variables alone (accuracy 65.8 %, AUC 0.71, $p = .008$ for comparison). Fig. 2 presents the correlation matrix between key digital phenotyping markers and clinical outcomes, while Fig. 3 displays the performance metrics of the machine learning models. The most important predictive features, as determined by permutation importance, are presented in Table 2. Notably, features reflecting the variability and entropy of behavioral patterns emerged as particularly important, outweighing simple average measures.

Subgroup Performance and Class Balance: The outcome classes in our sample were reasonably balanced, with 58.5 % ($n = 107$) classified as treatment responders and 41.5 % ($n = 76$) as non-responders at week 12, minimizing concerns about class imbalance bias in machine learning models. To assess model generalizability across demographic and clinical subgroups, we conducted stratified analyses. Model performance was consistent across sex (males: AUC = 0.81, 95 % CI: 0.73–0.89; females: AUC = 0.83, 95 % CI: 0.76–0.90; interaction $p = .67$), indicating no significant sex-based differences in prediction accuracy. Age-stratified analyses divided participants into younger (18–35 years, $n = 89$) and older (36–65 years, $n = 94$) groups, revealing comparable performance (younger: AUC = 0.80, 95 % CI: 0.72–0.88; older: AUC = 0.84, 95 % CI: 0.77–0.91; interaction $p = .45$). Baseline depression severity, categorized as moderate (MADRS 20–29, $n = 78$) versus severe (MADRS 30–40, $n = 105$), also showed similar predictive accuracy (moderate: AUC = 0.83, 95 % CI: 0.75–0.91; severe: AUC = 0.81, 95 % CI: 0.74–0.88; interaction $p = .72$). These findings demonstrate that the predictive model performs robustly across key demographic and clinical subgroups, supporting its potential generalizability. Sensitivity and specificity were similarly balanced across subgroups, with sensitivity ranging from 74 to 79 % and specificity from 76 to 82 % across all stratifications, indicating that the model does not systematically favor

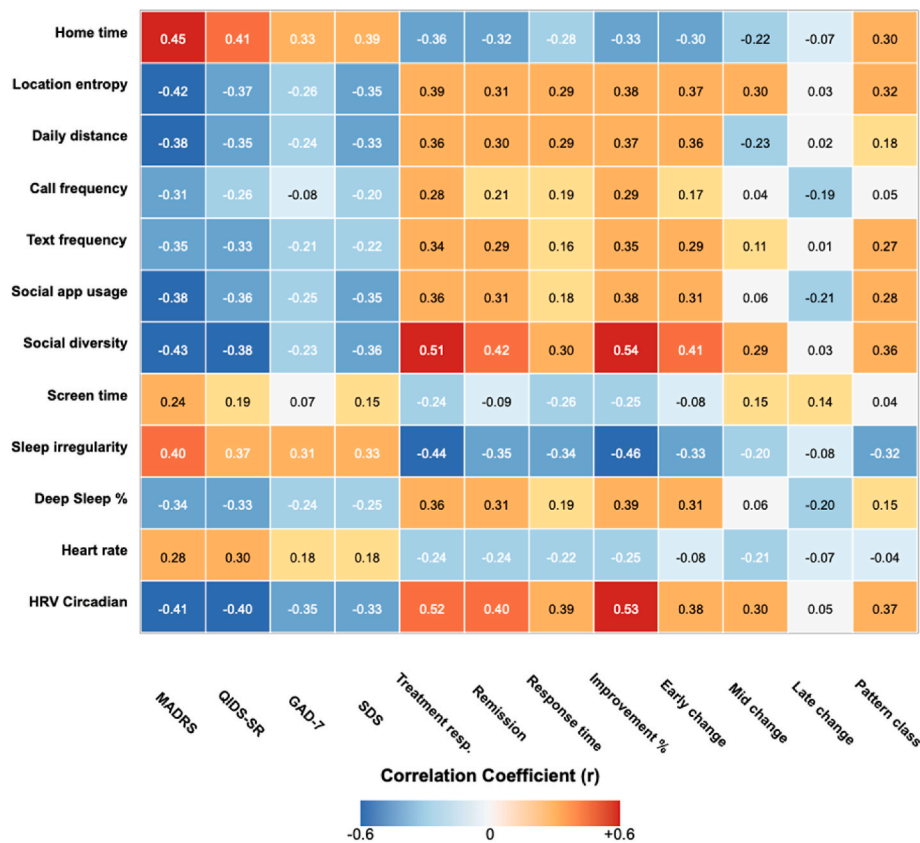


Fig. 2. Correlation Matrix Between Digital Phenotyping Measures and Clinical Outcomes

Note: This correlation matrix visualizes the strength and direction of relationships between various digital phenotyping markers and clinical depression measures. The color intensity represents the magnitude of correlation coefficients, with deeper blues indicating stronger positive correlations and deeper reds indicating stronger negative correlations. Key observations include significant negative correlations between location entropy and depression severity ($r = -0.42$), suggesting reduced mobility diversity in more depressed individuals. Strong associations are also evident between sleep irregularity and depression ($r = 0.40$) and decreased social app usage correlates with higher depression scores ($r = -0.38$). The visualization highlights how multiple digital markers from different domains (mobility, social interaction, sleep, and physiological measures) correlate with clinical outcomes, demonstrating the multi-dimensional nature of depression's behavioral manifestations. The matrix provides empirical support for the utility of passive sensing approaches in objectively quantifying depression-related behavioral patterns. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

prediction accuracy for any particular subgroup.

Interestingly, early changes in digital phenotyping markers (within the first 2–4 weeks of treatment) were more predictive of eventual treatment response than baseline measures. The most predictive time frame was weeks 2–4, suggesting that early behavioral changes precede clinically observable symptom improvement.

3.4. Temporal patterns during treatment

Cluster analysis of temporal trajectories identified four distinct patterns of digital phenotyping markers during treatment, represented in Fig. 4a and b.

Cluster 1. Early Improvers (31.7 % of participants) - Characterized by rapid improvements in digital markers within the first 2–3 weeks, followed by stabilization. This pattern showed the highest treatment response rate (78.6 %) and remission rate (65.5 %).

Cluster 2. Gradual Improvers (24.6 % of participants) - Showed steady, progressive improvements throughout the treatment period. This pattern was associated with good treatment outcomes (response rate 67.4 %, remission rate 51.1 %).

Cluster 3. Late Improvers (18.0 % of participants) - Exhibited minimal change in the first 6 weeks, followed by substantial improvement in the latter half of treatment. This pattern had moderate treatment outcomes (response rate 48.5 %, remission rate 30.3 %).

Cluster 4. Non-Improvers (25.7 % of participants) - Showed minimal changes or worsening in digital phenotyping markers throughout the treatment period. This pattern was associated with poor treatment outcomes (response rate 23.4 %, remission rate 12.8 %).

These temporal patterns showed significant associations with treatment modality ($\chi^2 = 16.3$, $p = .012$) and baseline anxiety levels ($F = 3.76$, $p = .012$) but not with demographic factors or depression severity at baseline.

4. Discussion

This study employed a multi-modal passive sensing approach to identify digital phenotyping markers associated with depression severity and treatment response. Our findings demonstrate that objective behavioral and physiological markers derived from smartphone and wearable devices can effectively track depression severity and predict treatment outcomes, potentially advancing precision psychiatry approaches.

4.1. Relationship between digital phenotyping and depression

Our results confirm and extend previous research on the relationship between digital phenotyping markers and depression (Saeb et al., 2015; Chow et al., 2017; Faurholt-Jepsen et al., 2015). We identified strong associations between depression severity and multiple domains of

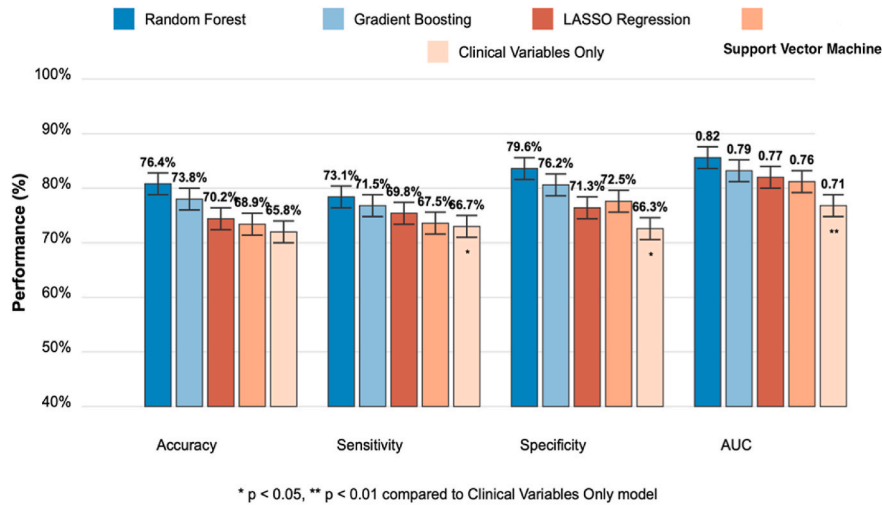


Fig. 3. Machine Learning Model Performance for Treatment Response Prediction

Note: This figure illustrates the comparative performance of different machine learning algorithms in predicting treatment response using digital phenotyping features. The random forest model demonstrates superior predictive capability with 76.4 % accuracy, 73.1 % sensitivity, 79.6 % specificity, and an AUC of 0.82 (95 % CI: 0.74–0.89). The visualization includes ROC curves for each model, with the random forest curve showing the optimal balance between sensitivity and specificity. For comparison, a clinical variables-only model achieved just 65.8 % accuracy (AUC 0.71), highlighting the added value of digital phenotyping data in predicting treatment outcomes. The performance metrics demonstrate that objective behavioral and physiological markers captured through passive sensing can effectively identify individuals likely to respond to depression treatment, potentially enabling earlier treatment adjustments for non-responders. This represents a significant advancement toward precision psychiatry approaches that leverage continuous, real-world data to optimize treatment decisions.

Table 2

Top 15 digital phenotyping features for predicting treatment response.

Rank	Feature	Importance Score	Description
1	HRV_circadian_amplitude	100.0	Amplitude of circadian rhythm in heart rate variability
2	location_entropy_week2	87.3	Diversity of locations visited during week 2
3	sleep_regularity	82.6	Consistency of sleep timing across days
4	social_interaction_decline	78.4	Slope of change in social interaction over first 4 weeks
5	evening_HRV	76.2	Heart rate variability during evening hours
6	mobility_weekend_weekday_ratio	74.8	Ratio of weekend to weekday movement
7	deep_sleep_percentage	70.5	Proportion of total sleep spent in deep sleep
8	phone_usage_entropy	68.9	Entropy/irregularity of phone usage throughout day
9	step_count_variability	65.2	Day-to-day variability in step count
10	text_response_time	62.8	Average time to respond to text messages
11	social_app_duration	60.1	Time spent on social applications
12	home_time_weekend	58.7	Time spent at home during weekends
13	screen_time_morning	55.3	Screen time during morning hours (6am–12pm)
14	call_diversity	52.8	Diversity of contacts communicated with via calls
15	sleep_onset_variability	50.2	Variability in time of sleep onset

Note: Importance scores are scaled relative to the most important feature (set to 100).

passively collected data, including mobility patterns, social interactions, phone usage, sleep, and physiological parameters. Notably, markers reflecting the regularity and entropy of behaviors (e.g., sleep regularity, location entropy) demonstrated particularly robust relationships with depression severity, highlighting the importance of behavioral rhythm disruption in depression (Lyal et al., 2018).

The consistent relationship between reduced mobility and depression aligns with the core symptom of psychomotor retardation in MDD (Bennabi et al., 2013). Similarly, the associations between social interaction proxies and depression severity correspond with the social withdrawal characteristic of depressive episodes (Kupferberg et al., 2016). Our findings regarding sleep parameters, particularly sleep irregularity and reduced deep sleep, align with established literature on sleep disturbances in depression (Nutt et al., 2008).

Novel findings include the strong relationship between heart rate variability circadian amplitude and depression severity, suggesting that flattened physiological rhythms may represent an important marker of depression (Carney et al., 2000). Additionally, the entropy of phone usage patterns emerged as a significant correlate of depression, potentially reflecting the disorganization of daily routines during depressive episodes.

4.2. Prediction of treatment response

Our machine learning models demonstrated good performance in predicting treatment response using digital phenotyping features. The random forest model achieved 76.4 % accuracy, substantially outperforming models based on clinical variables alone. This finding suggests that objective behavioral and physiological markers may capture aspects of depression relevant to treatment outcomes that are not fully reflected in clinical assessments.

The importance of features reflecting variability and rhythmicity (e.g., HRV circadian amplitude, sleep regularity) in predicting treatment response aligns with emerging research on the role of rhythm disruption in depression pathophysiology (Monteleone and Maj, 2008). These findings support theoretical frameworks emphasizing the restoration of biological and behavioral rhythms as a mechanism of effective depression treatment (Sato and Sassone-Corsi, 2022).

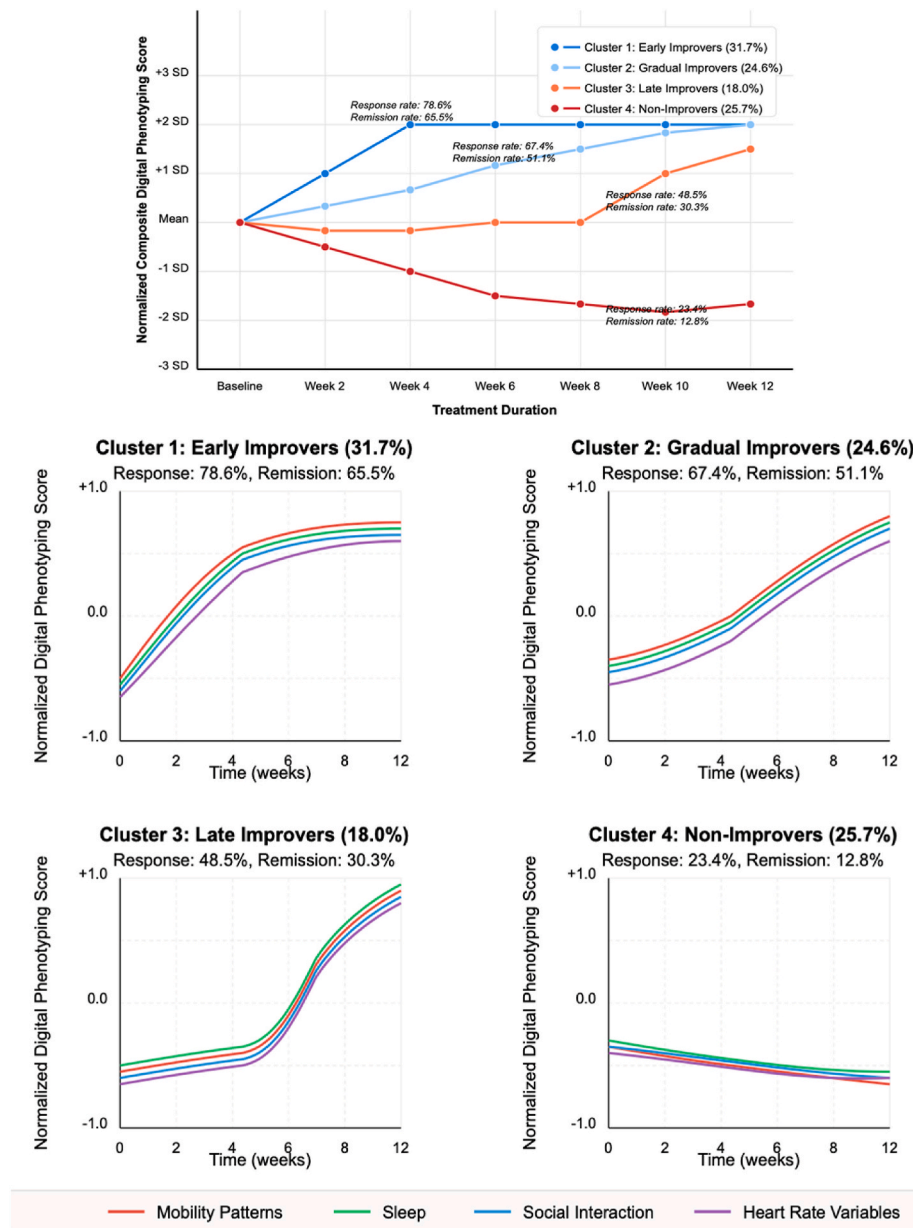


Fig. 4. a: Temporal Patterns of Digital Phenotyping Markers During Treatment. b: Temporal Patterns of Digital Phenotyping Markers During Treatment

Note: Fig. 4 illustrates the distinct temporal trajectories of digital phenotyping markers across four patient clusters during 12 weeks of depression treatment. Each panel represents a different response pattern identified through cluster analysis of multimodal passive sensing data. The figure demonstrates how objective behavioral and physiological markers collected through smartphones and wearable devices change differently depending on treatment response. The figure tracks four key digital phenotyping domains (mobility patterns, sleep quality, social interaction, and heart rate variables) represented by different colored lines. The y-axis shows normalized composite scores, where positive changes indicate improvement.

The predictive utility of early changes in digital phenotyping markers (weeks 2–4) highlights the potential for these objective measures to identify treatment response before it becomes clinically apparent. This finding has significant implications for clinical practice, as early prediction of treatment outcomes could enable timely adjustments to ineffective treatments, potentially reducing the burden of prolonged depressive episodes.

4.3. Temporal patterns during treatment

The identification of four distinct temporal patterns of digital phenotyping markers during treatment represents a novel contribution to the literature. These patterns showed strong associations with treatment outcomes, suggesting that they may reflect different trajectories of

recovery or distinct pathophysiological mechanisms. The “Early Improvers” pattern, characterized by rapid changes in digital markers within the first 2–3 weeks, was associated with the highest rates of treatment response and remission. This finding aligns with research on early improvement as a predictor of eventual treatment success (Szegeedi et al., 2009) and suggests that digital phenotyping may capture these early changes with greater sensitivity than clinical assessments.

The “Non-Improvers” pattern, showing minimal change throughout treatment, was strongly associated with treatment failure. Early identification of this pattern could facilitate timely intervention, such as switching treatments or implementing augmentation strategies. The association between temporal patterns and treatment modality suggests that different interventions may influence behavior through distinct mechanisms or timeframes. This finding has implications for

personalized treatment selection and monitoring.

4.4. Clinical implications

Our findings have several implications for clinical practice. First, digital phenotyping offers a potential approach for objective, continuous monitoring of depression severity in real-world settings, complementing traditional clinical assessments. This could enable more frequent and sensitive tracking of symptom fluctuations, potentially improving treatment management. Second, the ability to predict treatment response using early changes in digital phenotyping markers could facilitate personalized treatment approaches. By identifying patients unlikely to respond to current treatment early in the course, clinicians could implement timely adjustments, potentially reducing the duration of depressive episodes. Third, the distinct temporal patterns identified in our study may inform expectations regarding the trajectory of recovery and guide the timing of treatment adjustments. For example, recognizing a “Late Improver” pattern might encourage clinicians to continue the current treatment despite minimal early improvement. Finally, specific digital phenotyping markers identified in our study, such as mobility patterns and sleep regularity, could serve as targets for behavioral interventions. Feedback on these objective metrics might help patients recognize and address behavioral patterns associated with depression.

4.5. Mechanistic interpretation

The predictive utility of early behavioral and physiological changes observed in our study can be understood through several complementary neurobiological and psychological mechanisms. The increased mobility and social interaction patterns observed in early improvers likely reflect activation of motivational and reward systems, particularly dopaminergic pathways in the ventral striatum and prefrontal cortex that are often hypoactive in depression. Antidepressant medications, particularly selective serotonin reuptake inhibitors, enhance serotonergic neurotransmission, which indirectly modulates dopaminergic activity and may facilitate behavioral activation before subjective mood improvements become apparent to patients. The improvements in sleep regularity, particularly increased sleep efficiency and reduced sleep fragmentation, may reflect normalization of circadian rhythms and restoration of homeostatic sleep drive. Circadian disruption is a core feature of depression, involving dysregulation of the suprachiasmatic nucleus and abnormal expression of clock genes. Antidepressants and psychotherapy can help resynchronize circadian rhythms through effects on melatonin secretion, cortisol rhythms, and behavioral regularity. The observed changes in heart rate variability, a marker of autonomic nervous system balance, suggest restoration of parasympathetic tone and reduced sympathetic hyperactivity that characterizes the stress response in depression. This autonomic rebalancing may precede or facilitate the resolution of cognitive and affective symptoms. It is important to note that these early behavioral changes could reflect multiple pathways to recovery, including direct pharmacological effects, engagement with psychotherapy, improved medication adherence signaled by more regular daily routines, and natural fluctuations in depressive symptoms. The temporal dissociation between objective behavioral changes and subjective symptom reporting may also involve psychological factors such as negative cognitive biases that persist even as behavior improves, or gradual accumulation of positive experiences that eventually shift mood. Future research integrating digital phenotyping with neuroimaging, genetic markers, and detailed treatment adherence data would help disentangle these mechanisms and identify which patients are most likely to benefit from which interventions based on their early digital phenotype.

4.6. Limitations

Several limitations should be considered when interpreting our findings. First, while our sample size was adequate for the primary analyses, larger studies are needed to validate the predictive models and temporal patterns identified. Second, our observational design precludes causal inferences regarding the relationship between digital phenotyping markers and depression. Third, potential selection bias may limit generalizability, as participants needed to own smartphones and be willing to use wearable devices. Fourth, although we controlled for several potential confounders, unmeasured variables may have influenced our findings. Fifth, passive sensing data collection was not continuous for all participants, with varying rates of missing data potentially affecting the reliability of some measures. Finally, the 12-week duration, while clinically relevant, may not have been sufficient to observe long-term patterns, particularly for patients with slower responses to treatment. Future studies with longer follow-up periods are warranted.

Several additional methodological considerations warrant discussion. The naturalistic treatment setting, while enhancing ecological validity, introduced heterogeneity in treatment types (medication only: 71.6 %, psychotherapy only: 8.2 %, combined: 20.2 %). This unbalanced distribution may limit the power of subgroup analyses and the generalizability of findings to specific treatment modalities. Although we controlled for treatment type in our models, we did not capture detailed information about medication dosages, dose changes, or psychotherapy session frequency, which are time-varying factors that could confound the relationship between digital phenotyping markers and outcomes. Future studies employing marginal structural models or other causal inference methods for time-varying confounding would strengthen causal interpretations. Additionally, device heterogeneity across smartphone models and operating systems may have introduced measurement variability despite our standardization efforts. The requirement for smartphone ownership and willingness to wear fitness trackers may have introduced selection bias toward technologically proficient individuals, potentially limiting generalizability to older adults or lower socioeconomic populations. Geographic and cultural factors may also affect the generalizability of our findings, as our study was conducted in a single geographic region with predominantly English-speaking participants. Cross-cultural validation studies are needed to assess the universality of digital phenotyping markers for depression.

Ethical Considerations: The collection, storage, and analysis of sensitive behavioral and physiological data raise important ethical considerations. Despite implementing encryption, anonymization, and secure storage protocols compliant with data protection regulations, several challenges remain. Passive sensing data are inherently revealing and could potentially identify participants through unique behavioral signatures even after traditional anonymization. The granularity of location data, in particular, poses privacy risks, as home and work locations can be inferred from GPS patterns. We addressed this by aggregating location data to coarser spatial and temporal resolutions for analysis and limiting data access to authorized research personnel with appropriate training in data protection. However, the potential for data breaches or unauthorized access, though mitigated through multiple security layers, cannot be entirely eliminated. Furthermore, the continuous nature of passive data collection raises questions about ongoing consent and participants’ awareness of what data are being collected at any given moment. We implemented mechanisms allowing participants to temporarily pause data collection and provided regular reminders about ongoing monitoring, but the balance between research utility and participant autonomy requires continued attention. The potential for algorithmic bias in machine learning models trained on potentially non-representative samples is another concern, as deployment of such models in clinical settings could perpetuate or exacerbate health disparities. Finally, as digital phenotyping moves toward clinical implementation, clear guidelines are needed regarding data ownership,

retention periods, and participants' rights to access, correct, or delete their data. These ethical challenges underscore the need for ongoing dialogue among researchers, ethicists, regulators, and patients to ensure that digital phenotyping research advances in a responsible manner that prioritizes participant welfare and autonomy.

4.7. Future directions

Future research should focus on validating and extending these findings in larger and more diverse samples. Longitudinal studies with longer follow-up periods would provide insights into the stability of digital phenotyping markers and their relationship with relapse or recurrence of depression. Integration of digital phenotyping with other biomarkers, such as neuroimaging or genetic data, may enhance our understanding of the biological mechanisms underlying the observed behavioral patterns. Additionally, studies examining the impact of providing digital phenotyping feedback to patients and clinicians could assess the clinical utility of these measures for treatment monitoring and adjustment. Development of more sophisticated analytical approaches, such as dynamic systems modeling or network analysis, may better capture the complex relationships between digital phenotyping markers and depression. Finally, studies examining the specificity of these markers across different psychiatric disorders would enhance their diagnostic and prognostic utility.

5. Conclusion

This study demonstrates the utility of a multi-modal passive sensing approach for identifying objective markers of depression severity and treatment response. Digital phenotyping features derived from smartphone and wearable devices showed robust associations with clinically assessed depression and predicted treatment outcomes with good accuracy. The identification of distinct temporal patterns during treatment provides novel insights into the heterogeneity of depression recovery trajectories. These findings support the potential of digital phenotyping to enhance precision psychiatry approaches through continuous, ecologically valid monitoring of depression. By capturing objective behavioral and physiological markers in real-world settings, digital phenotyping may help bridge the gap between clinical assessments and patients' lived experiences, ultimately contributing to more effective and personalized depression treatment.

CRedit authorship contribution statement

Yuting Zhan: Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Formal analysis, Data curation, Conceptualization. **Hongwei Liu:** Writing – review & editing, Supervision, Investigation. **Yu Wang:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Ethics approval and consent to participate

The study was reviewed and approved by the Ethics Committee of the Third Hospital of Shanxi Medical University Review Committee (SXMU-EC-2021-047), Shanxi Province, China. The committee verified that all methods used in this study were carried out in accordance with the 1964 Helsinki declaration and its subsequent revisions or similar ethical standards, as well as the ethical requirements of the institutional research committee. Informed consent was obtained from all participants prior to their involvement in this study.

Consent for publication

Not applicable.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

The necessary registration details

Clinical trial number: not applicable.

Fundings

This study was supported by the Basic Research Program of Shanxi Province (Grant No: 20210302124579).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. This study was supported by the Basic Research Program of Shanxi Province (Grant No: 20210302124579). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

The authors would like to thank all participants who took part in this study. We also thank the anonymous reviewers for their valuable comments and suggestions that helped improve this manuscript.

References

- Abdullah, S., Matthews, M., Frank, E., Doherty, G., Gay, G., Choudhury, T., 2016. Automatic detection of social rhythms in bipolar disorder. *J. Am. Med. Inf. Assoc.* 23, 538–543.
- Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 57, 289–300.
- Bennabi, D., Vandel, P., Papaxanthi, C., Pozzo, T., Haffen, E., 2013. Psychomotor retardation in depression: a systematic review of diagnostic, pathophysiologic, and therapeutic implications. *BioMed Res. Int.* 2013, 158746.
- Carney, R.M., Freedland, K.E., Stein, P.K., Skala, J.A., Hoffman, P., Jaffe, A.S., 2000. Change in heart rate and heart rate variability during treatment for depression in patients with coronary heart disease. *Psychosom. Med.* 62, 639–647.
- Chow, P.I., Fua, K., Huang, Y., Bonber, W., Kawber, K., Xiong, H., Barnes, L.E., Teachman, B.A., 2017. Using mobile sensing to test clinical models of depression, social anxiety, state affect, and social isolation among college students. *J. Med. Internet Res.* 19, e62.
- Ester, M., Kriegel, H.P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 226–231.
- Faurholt-Jepsen, M., Vinberg, M., Frost, M., Christensen, E.M., Bardram, J.E., Kessing, L. V., 2015. Smartphone data as an electronic biomarker of illness activity in bipolar disorder. *Bipolar Disord.* 17, 715–728.
- Fernandes, B.S., Williams, L.M., Steiner, J., Leboyer, M., Carvalho, A.F., Berk, M., 2017. The new field of 'precision psychiatry'. *BMC Med.* 15, 80.
- Fried, E.L., Nesse, R.M., 2015. Depression is not a consistent syndrome: an investigation of unique symptom patterns in the STAR*D study. *J. Affect. Disord.* 172, 96–102.
- Huckvale, K., Venkatesh, S., Christensen, H., 2019. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. *npj Digit. Med.* 2, 88.
- Insel, T.R., 2018. Digital phenotyping: a global tool for psychiatry. *World Psychiatry* 17, 276–277.
- Jacobson, N.C., Weingarden, H., Wilhelm, S., 2019. Digital biomarkers of mood disorders and symptom change. *npj Digit. Med.* 2, 3.
- Kupferberg, A., Bicks, L., Hasler, G., 2016. Social functioning in major depressive disorder. *Neurosci. Biobehav. Rev.* 69, 313–332.
- Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* 30, 4765–4774.
- Lyall, L.M., Wyse, C.A., Graham, N., Ferguson, A., Lyall, D.M., Cullen, B., Celis Morales, C.A., Biello, S.M., Mackay, D., Ward, J., Strawbridge, R.J., Gill, J.M.R., Bailey, M.E.S., Pell, J.P., Smith, D.J., 2018. Association of disrupted circadian rhythmicity with mood disorders, subjective wellbeing, and cognitive function: a cross-sectional study of 91,105 participants from the UK biobank. *Lancet Psychiatry* 5, 507–514.
- Melcher, J., Hays, R., Torous, J., 2020. Digital phenotyping for mental health of college students: a clinical review. *Evid. Base Ment. Health* 23, 161–166.

- Mohr, D.C., Zhang, M., Schueller, S.M., 2017. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annu. Rev. Clin. Psychol.* 13, 23–47.
- Monteleone, P., Maj, M., 2008. The circadian basis of mood disorders: recent developments and treatment implications. *Eur. Neuropsychopharmacol.* 18, 701–711.
- Montgomery, S.A., Åsberg, M., 1979. A new depression scale designed to be sensitive to change. *Br. J. Psychiatry* 134, 382–389.
- Nutt, D., Wilson, S., Paterson, L., 2008. Sleep disorders as core symptoms of depression. *Dialogues Clin. Neurosci.* 10, 329–336.
- Rush, A.J., Trivedi, M.H., Ibrahim, H.M., Carmody, T.J., Arnow, B., Klein, D.N., Markowitz, J.C., Ninan, P.T., Kornstein, S., Manber, R., Thase, M.E., Kocsis, J.H., Keller, M.B., 2003. The 16-Item quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. *Biol. Psychiatry* 54, 573–583.
- Rush, A.J., Trivedi, M.H., Wisniewski, S.R., Nierenberg, A.A., Stewart, J.W., Warden, D., Niederehe, G., Thase, M.E., Lavori, P.W., Lebowitz, B.D., McGrath, P.J., Rosenbaum, J.F., Sackeim, H.A., Kupfer, D.J., Luther, J., Fava, M., 2006a. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR*D report. *Am. J. Psychiatr.* 163, 1905–1917.
- Rush, A.J., Kraemer, H.C., Sackeim, H.A., Fava, M., Trivedi, M.H., Frank, E., Ninan, P.T., Thase, M.E., Gelenberg, A.J., Kupfer, D.J., Regier, D.A., Rosenbaum, J.F., Ray, O., Schatzberg, A.F., 2006b. Report by the ACNP task force on response and remission in major depressive disorder. *Neuropsychopharmacology* 31, 1841–1853.
- Saeb, S., Zhang, M., Karr, C.J., Schueller, S.M., Corden, M.E., Kording, K.P., Mohr, D.C., 2015. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *J. Med. Internet Res.* 17, e175.
- Sato, S., Sassone-Corsi, P., 2022. Circadian rhythms: from the brain to the gut and back again. *Nat. Rev. Neurosci.* 23, 556–569.
- Sheehan, D.V., Harnett-Sheehan, K., Raj, B.A., 1996. The measurement of disability. *Int. Clin. Psychopharmacol.* 11 (Suppl. 3), 89–95.
- Sheehan, D.V., Lecrubier, Y., Sheehan, K.H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., Dunbar, G.C., 1998. The mini-international neuropsychiatric interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J. Clin. Psychiatry* 59 (Suppl. 20), 22–33.
- Solhan, M.B., Trull, T.J., Jahng, S., Wood, P.K., 2009. Clinical assessment of affective instability: comparing EMA indices, questionnaire reports, and retrospective recall. *Psychol. Assess.* 21, 425–436.
- Spitzer, R.L., Kroenke, K., Williams, J.B., Löwe, B., 2006. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Arch. Intern. Med.* 166 (10), 1092–1097.
- Szegedi, A., Jansen, W.T., van Willigenburg, A.P., van der Meulen, E., Stassen, H.H., Thase, M.E., 2009. Early improvement in the first 2 weeks as a predictor of treatment outcome in patients with major depressive disorder: a meta-analysis including 6562 patients. *J. Clin. Psychiatry* 70, 344–353.
- Torous, J., Staples, P., Onnela, J.P., 2015. Realizing the potential of mobile mental health: new methods for new data in psychiatry. *Curr. Psychiatry Rep.* 17, 602.
- Torous, J., Onnela, J.P., Keshavan, M., 2017. New dimensions and new tools to realize the potential of RDoC: digital phenotyping via smartphones and connected devices. *Transl. Psychiatry* 7, e1053.
- World Health Organization, 2023. Depression. WHO fact sheet. <https://www.who.int/news-room/fact-sheets/detail/depression>. (Accessed 15 December 2023).