Original Paper

Variability in Self-reported Depression Symptomology and Associated Mobile-Sensed Behavioral Patterns in Digital Phenotyping: An Observational Study

Arsi Ikäheimonen[1], Nguyen Luong[1], Ilya Baryshnikov[2,3], Ti John[1], Annasofia Martikkala[2,3], Erkki Isometsä[2,3], Talayeh Aledavood[1]

[1] Department of Computer Science, Aalto University, Espoo, Finland
[2] Department of Psychiatry, University of Helsinki, Helsinki, Finland
[3] Helsinki and Uusimaa Hospital District, Helsinki, Finland

Corresponding Author:
Arsi Ikäheimonen, MSc
Department of Computer Science
Aalto University
FI-02150 Espoo, Finland
Email: arsi.ikaheimonen@aalto.fi

Abstract

Background: Digital phenotyping studies using smartphone-sensed data have identified several behavioral markers associated with depression. However, the generalizability of these markers is limited by the variability in depressive symptoms and associated behaviors, both between and within individuals. While existing research acknowledges this variability, its in-depth analysis remains underexplored.
Objective: This study investigates variability from two angles: (1) variability in self-reported depression symptoms and (2) variability in associated smartphone-sensed behavioral markers. To this end, we examine depression at the symptom level, across diagnostic groups, and by severity level. Associations between depression and behavioral markers are analyzed at the study population, diagnostic group, and individual levels.
Methods: We analyzed a dataset of smartphone-sensed data from 64 patients diagnosed with major depressive episodes belonging to three different subgroups: major depressive disorder (MDD, n=42), borderline personality disorder (BPD, n=12), and bipolar disorder (BD, n=10). Depression severity was assessed by self-reported 9-item Patient Health Questionnaire (PHQ-9) scores. Differences in depression were analyzed at the item and group levels using non-parametric distributional testing, and at the severity level through exploratory data analysis. Variability in behavior–depression associations were analyzed using correlation analysis, and multilevel modeling was used to identify behavioral markers of depression at both between-person and within-person levels, while controlling for demographics and contextual factors.
Results: Our analysis identified group-level differences in five PHQ-9 items: appetite changes, low self-worth, difficulty concentrating, psychomotor disturbances, and suicidal

thoughts. Sample level analysis revealed weak associations between 3.1% (3/98) of the features and depression severity. At the individual level, we found mixed correlations for 98% of the features (96/98) across participants, being positive for some and negative for others. Multilevel modeling indicated that between-person differences explained 58.5% of depression severity variance. Adding behavioral features, demographic background, and contextual factors increased the explained variance to 63.7%. Significant within-person behavioral predictors included nighttime communication application usage (β = 0.31, 95% CI = [0.01-0.61], P = .042), morning battery level variability (β = 0.39, 95% CI = [0.06-0.72], P = .019), morning screen usage (β = 0.35, 95% CI = [0.04-0.65], P = .027), average sleep duration (β = 0.39, 95% CI = [0.11-0.67], P = .006), and morning incoming SMS count (β = 0.36, 95% CI = [0.05-0.66], P = .022), all associated with higher depression severity. Daily outgoing SMS count was associated with lower depression severity (β = -0.36, 95% CI = [-0.-0.66], P = .022).
Conclusions: We found differences in self-reported depression severity and associations between behavioral markers and depression severity. The results point out the importance of accounting for the differences in depression symptoms and associated behaviors in analysis; ignoring these differences can lead to misleading conclusions and poor generalization.

Keywords: *digital phenotyping; digital behavioral data; depressive symptoms; mHealth; smartphone; statistical modeling*

1. Introduction

Depressive disorders rank among the most prevalent mental disorders and are among the most burdensome health conditions worldwide [1]. In 2021, an estimated 330 million people worldwide were living with depressive disorders, affecting approximately 4.4% of the global population [2]. Clinical assessment of depression relies on structural interviews, observer scales and questionnaires [3], which are prone to biases such as recall bias [4]. Additionally, the heterogeneous nature [5] of depressive symptoms add further complexity to diagnosis.

The ubiquity of smartphones and other personal smart devices has inspired research to use data from these devices to detect and predict depressive symptoms continuously, objectively, and in real-life settings. Digital phenotyping is a field of science that utilizes personal digital devices to measure human behavior and physiology unobtrusively and continuously [6]. Collecting data on individuals' behavior using digital devices can provide valuable insights into depression and associated behavioral patterns. Previous studies have utilized such data to monitor and predict depression by focusing on specific behavioral aspects. For example, in a study by Masud et al [7] physical activity and location patterns were measured through smartphone accelerometer and GPS data. A study by Cao et al [8] inferred sociability from the frequency of SMS messages and the average duration of phone

calls, and sleep duration and interruptions were measured using devices like Fitbit activity trackers in a study by Xu et al[9]. Ultimately, digital devices would improve symptom monitoring and prediction, work as an early warning system for symptom worsening, be easily integrated with interventions, and supplement care management [10].

Research on digital phenotyping of depression has mainly focused on identifying associations between smartphone-sensed behavior and depression symptoms, predicting depression symptom severity and diagnostic classification, and forecasting future mood [11]. The studies typically rely on behavioral data representing participants' physical activity, mobility, sociability, sleep, and smartphone usage patterns[12], and utilize methods ranging from simple linear modeling to more advanced machine learning and deep learning techniques [13]. Though these approaches have achieved promising results, the reported associations between behavior and depression vary across studies in both statistical significance and direction (positive or negative) (e.g.,[12,14]). These discrepancies imply that the results may not generalize well across different samples and contexts. The differences may arise partly from differences in study objectives, samples, methodologies, and reporting practices but also from the differences in depressive symptoms and behavioral manifestation between and within the patients [15]. Although these differences are acknowledged, their effect on research outcomes is rarely explored in depth or reported. To address this gap, we evaluate and characterize the variability of depression symptoms and associated behaviors, between and within our study participants. Understanding these variability is crucial for identifying relevant behavioral markers that are truly indicative of depression severity.

This work aims to explore variability in depression and in behavioral markers connected with depression as measured through digital phenotyping, and the importance of considering these variability when choosing the methodology and modeling approach. We investigated variability from two angles. The first angle looked at differences in self-reported depression symptoms through answers to questionnaires. The second angle focused on the smartphone-sensed behavioral markers associated with depression by performing correlation and multilevel analyses. The first angle was represented by our first study objective where the questionnaire answers were compared at the symptom level, across diagnostic groups, and by severity level. The second angle led to our second and third study objectives. The second objective was to study correlations between behavioral markers and depression across study population (pooled analysis), diagnostic groups, and at the individual participant level by calculating pairwise and rolling-window correlations. The third objective was to use multilevel modeling to assess within- and between-person variability in behavioral features and identify behavioral markers indicative of depression while controlling for participants' background and contextual factors.

The study analyzed smartphone-collected behavioral data and self-reported depression scores from the 9-item Patient Health Questionnaire (PHQ-9) [16]. We quantified behavioral features from mobile-sensed data, reflecting physical activity and mobility, sociability, phone usage, and sleep. These features served as objective proxies for broader patterns associated with depression. For example, we used phone-sensed acceleration as a proxy for physical activity, where reduced activity might reflect increased fatigue. Similarly, reduced sociability (approximated via phone call durations and SMS message counts) might reflect social withdrawal and reduced engagement. Finally, disturbances in sleep (hypo- and hypersomnia and irregular sleep timing, which are considered common depressive symptoms) were inferred from smartphone usage.

Our work provides an overview of the variability in depressive symptoms and associated behavioral markers within a sample of outpatients diagnosed with depression. We found notable differences in depressive symptom presentations, variability in behavior–depression associations, and key behavioral features linked to depression severity. These findings benefit future theory building, identification of potential behavioral markers, selection of data collection methods, modeling techniques, and study design considerations, such as a sufficient sample size. Further, the observed between- and within-participant differences suggest using personalized modeling instead of group-level models for depression monitoring and prediction. We encourage future research to report the distributions and variability of key depressive symptoms and behavioral markers to facilitate result evaluation and enhance generalizability.

2. Methods

Setting

This study analyzed behavioral data from the Mobile Monitoring of Mood (MoMo-Mood) study [17], which includes continuously sensed smartphone behavioral data (passive data), self-reported biweekly depression questionnaire responses (active data), and background demographics and contextual information. The study involved 164 participants across three patient groups, major depressive disorder (MDD, n=85), bipolar disorder (BD, n=21), borderline personality disorder (BPD, n=27), and a healthy control group (N=31). Patients were voluntarily recruited from primary care and psychiatric outpatient treatment clinics in Finland, while healthy controls were recruited among university students and healthcare personnel. The patients were under treatment, and diagnosed as having ongoing major depressive episodes with structured interviews, the Mini-International Neuropsychiatric Interview [18]and the Structured Clinical Interview for DSM-IV Axis II Personality Disorders [19]. The study used continuous enrollings, allowing the participants to enter or

exit the study at different times, with a recommended participation period of one year. For more information, refer to earlier work[15,17,20].

Dataset description

Passive data encompasses five aspects of daily behavior: physical activity and mobility, sociability, phone interactions, and sleep. We operationalized these behavioral aspects through specific features extracted from raw smartphone data, serving as proxies for real-world behaviors. The active data has participants' self-reported PHQ-9 questionnaires, which assess the depression severity. PHQ-9 is a tool used in screening for diagnosis and monitoring of depressive symptoms [21]. The survey comprises nine items, each addressing the main symptoms of depression, such as having little interest or pleasure in doing things. The patients are asked to respond based on their experience over the last two weeks, with each item rated from 0 (not at all) to 3 (nearly every day). The severity of depression is measured by the total sum of independent items, ranging from 0 (no depression) to 27 (severe depression). For the details, refer to Table S1 in Multimedia Appendix 1.
Background demographics include categorical variables, including participants' age and sex, used as control variables. Contextual information includes details about full-time work, shift work, cohabitation status, the number of children, and whether the participants have pets. We selected these factors since they may affect participants' daily rhythms by introducing higher or lower regularity. Additionally, the COVID-19 stringency index [22] was used as a time-varying external factor to control the impact of pandemic-related restrictions that occurred partially during the study.

Preprocessing

Of the 164 participants, we excluded the control participants from the analysis since they exhibited low levels and variability (mean = 1.2, std = 1.8) of self-measured depression (Ikäheimonen et al., 2024). Further, we excluded the participants who provided less than one month of passive data (equaling at least two biweekly PHQ-9 responses), resulting in total 64 (39%) participants from three different groups: (1) patients with BD (n=10), (2) patients with BPD (n=12), and (3) patients with MDD (n=42).

We used the Niimpy behavioral analysis toolbox [23] to preprocess the raw sensor data and extract behavioral features. The raw data originate from various smartphone sensors and logs, including accelerometer, application usage log, battery level, GPS location, phone call log, SMS log, and screen usage. We extracted three types of features: volume-based, temporal, and contextual features. The volume based features represent counts and durations of specific activities, such as smartphone screen usage time. Temporal features were derived by segmenting the volume-based features into 6-hour bins based on the time of day (12:00 AM-6:00 AM, 6:00 AM-12:00 PM, 12:00 PM-6:00 PM, and 6:00 PM-12:00 AM), similar to our previous study [20]. Contextual features were inferred from GPS data,

reflecting behavioral patterns such as time spent at home. We used these features as proxies for behavioral aspects of physical activity, mobility, sociability, phone interactions, and sleep. Physical activity was captured through phone accelerometer data, providing temporal information about movement intensity. Mobility was assessed using GPS location data, providing spatial and contextual behavior information, such as total distance travelled and the number of unique locations visited daily. Sociability was measured by phone call counts, durations, and the number of messages sent and received. Phone interactions were inferred from application usage logs (e.g., application category, usage count, and duration), screen usage counts (on and off), usage duration, and battery charge level. Sleep duration and timing were approximated using screen activity data, following a process similar to that demonstrated by Aledavood et al[24]. For further details, refer to Table S2 in Multimedia Appendix 1.

Extracted behavioral features were aligned with biweekly depression questionnaires and aggregated by calculating the arithmetic mean over the 14 days preceding each PHQ-9 measurement. Further, we pre-filtered redundant features based on prior knowledge (for example, screen off count, since it reflects the same information as screen on count). We applied additional filtering by removing features with a high missingness ratio (threshold of 75% or more missing data points across the participants), highly correlated features (threshold Pearson correlation 0.7 or higher), features showing no or minimal variance (threshold of 0.01 or lower standard deviation across all participants) and excluded the features with a variance inflation factor (VIF) greater than 10 to address predictor collinearity. After the filtering, missing intermittent passive data were filled using median imputation. The resulting data contained 790 biweekly observations from 98 behavioral features, two background demographic, five contextual, and one external variable. Figure 1 shows the data preprocessing and aggregation schematics.

For the first study objective, to assess how depressive symptoms differ between patient groups, we used the raw values from the PHQ-9 questionnaire items, removing missing or incomplete records. For the second objective, behavior–depression association analysis, we focused on the within-person variability to examine how participants' behaviors fluctuate over time and how those fluctuations relate to changes in their depression severity. This method involves standardizing the behavioral data features within each participant by calculating their mean and standard deviation and then transforming each data point by subtracting the individual's mean and dividing by their standard deviation. We merged passive (behavioral features) and active (self-assessed PHQ-9 scores) data using 14-day frequency, keeping the behavioral data for each participant from the days preceding the observed PHQ-9 scores. We selected only participants completing at least two PHQ-9 questionnaires and provided matching passive data. For the third objective, we applied multilevel modeling to examine within- and between-person effects by disaggregating passive data features using person-mean and grand-mean centering [25]. The within-person components (reflecting the fluctuations relative to an individual's own average) were obtained by standardizing the behavioral data features within each participant, as detailed above. To model the between-person effects, we constructed the between-person components by first calculating the mean value of each behavioral feature within each participant. These individual means were then standardized across all

participants by subtracting the overall grand mean and dividing by the standard deviation of these means. These between-person effects reflect how differences in average behavior relate to depression severity (e.g., whether participants sleeping longer on average have higher or lower depression levels). Finally, categorical background demographics and contextual variables for multilevel modeling were encoded using one-hot encoding.

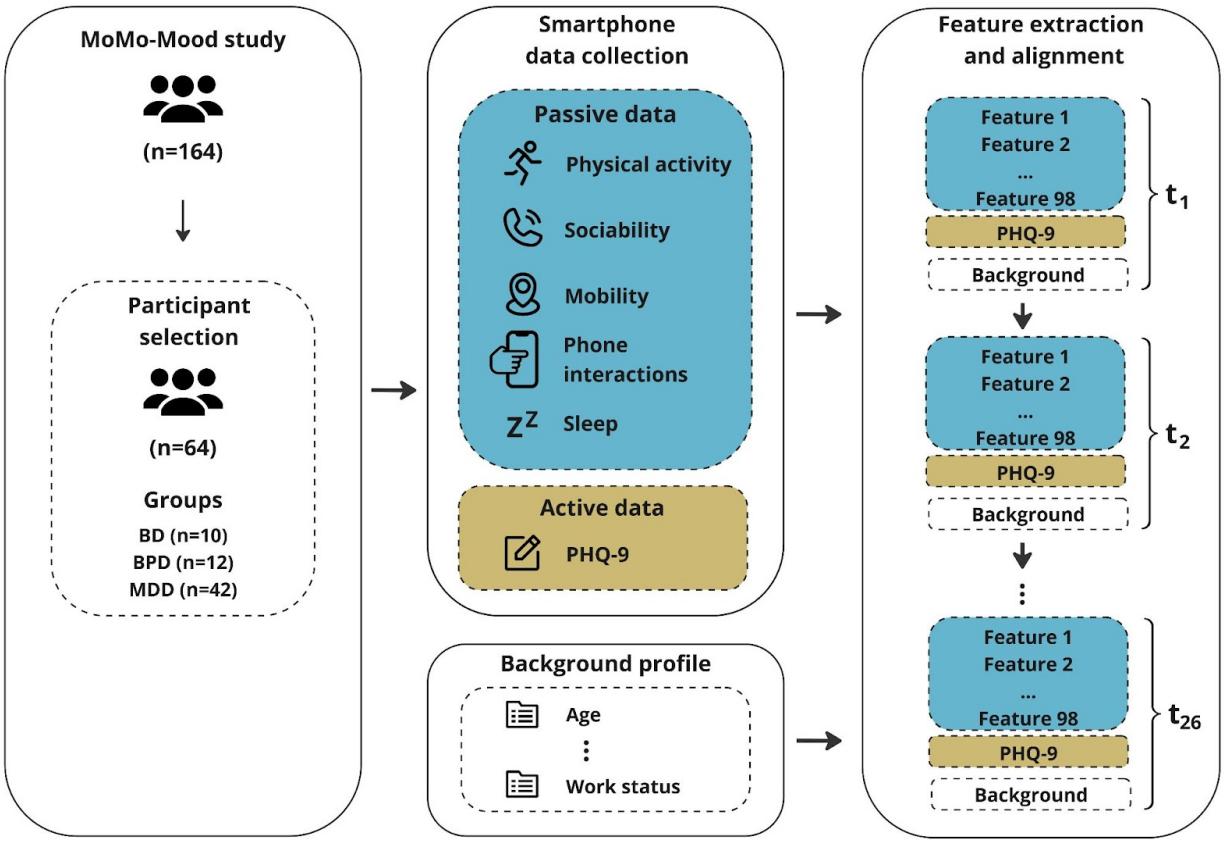

Figure 1. Participant selection and data preprocessing schema. For this study, we selected participants from three patient groups (BD, BPD, and MDD). Further, we selected individuals who answered at least two bi-weekly self-assessed PHQ-9 questionnaires and provided passive data for two weeks prior to each PHQ-9 score. Passive data consisted of raw smartphone sensed data, including accelerometer readings, phone call and SMS logs, app and screen usage log, battery level logs, and GPS location data. Additionally, demographics information, such as, participants age and work status, were collected at the time of enrollment. Behavioral features were extracted from raw passive data, with a total of 98 behavioral features included in the analysis. As the passive data was sampled at various frequencies, all features were aggregated over two-week periods, and aligned with bi-weekly PHQ-9 scores, reported at the end of a given two-week period.

PHQ-9 questionnaire analysis

To assess the variability in self-reported depression, we conducted a descriptive analysis of PHQ-9 scores and a distribution analysis for each item across the groups and different

depression severity levels. Descriptive analysis included comparing PHQ-9 score summary statistics, including mean and standard deviation across the patient groups with the Kruskall-Wallis test [26]. Distribution analysis assessed differences in PHQ-9 item score compositions by comparing the average scores for each item across patient groups and depression severity levels. We employed the Kruskal-Wallis test to assess group-wise differences and the pairwise Mann-Whitney test between groups for individual PHQ-9 items. To control for multiple comparisons, we applied false detection rate (FDR) correction using the Benjamini-Hochberg (BH) method [27]with a significance level of α=0.05.

Association analysis

To study how behavioral features are associated with self-reported depression among participants, we conducted a correlation analysis between the features and PHQ-9 scores. We focused on within-person effects, specifically how deviations from a participant's average behavior (e.g., average daily sleep duration in hours) were associated with their PHQ-9 scores. We utilized two approaches: pooled- and within-participant analysis. For pooled-participants analysis, we calculated pairwise Pearson correlation coefficients between the passive behavioral features and PHQ-9 scores by using the data from all participants. For within-participant analysis, correlations were calculated independently for each participant. To provide an overview of behavior–depression associations, we classified within-participant correlations. We labeled each Pearson correlation coefficient as weak if it was between ±0.2, as positive if it was above 0.2, and as negative if it was below -0.2, and summarized the labels across sensors and participant groups.

We conducted a rolling-window correlation analysis to explore the temporal dynamics of behavior–depression associations using a window length corresponding to four bi-weekly observations. For an overview, we counted the proportions of labeled correlations across each participant and sensor.

To evaluate the statistical significance of the within-person correlations, we conducted permutation tests with 10,000 iterations, randomly shuffling PHQ-9 scores across participants and time points. We used the Kolmogorov-Smirnov two-sample test [28]for each behavioral feature to compare the empirical distribution of observed correlation coefficients against the permuted correlation distributions. To control for multiple comparisons, we applied FDR correction using the BH method at a significance level of α=0.05.

Multilevel modeling

To assess the effect of within- and between-person behavioral features on PHQ-9 scores, we used multilevel linear mixed modeling to account for the data's hierarchical structure. Demographic background, and contextual predictors were introduced as control variables to account for their possible effects on depression severity. We first built a baseline model with only a random intercept for each participant to evaluate the total unexplained variance

in PHQ-9 scores (unconditional variance) and variability between participants. We then compared three different model configurations: 1) a model with random intercepts and demographic and contextual factors, including age, sex, full-time employment, shift work, dog ownership, and the COVID-19 Stringency Index, 2) adding within-person effects into the model, and 3) adding between-person effects into the model. These models allowed us to assess how within- and between-effects explained the variability in self-reported depression and how much of the variability can be attributed to demographic and contextual factors. Finally, we compared the models with analysis of variance (ANOVA), assessing both their explanatory power and parsimony.

3. Results

PHQ-9 questionnaire analysis

Our first objective was to assess the variability in depression symptoms through subjective data and analyze PHQ-9 answers at symptom (questionnaire item), diagnostic group and severity levels. This section presents PHQ-9 summary statistics, group-level comparison of items, and item distributions across severity levels.

The summary statistics showed that the MDD patients had, on average, the lowest average depression level at group-level, while the scores standard deviation has similar magnitude across the groups. Inspecting the standard deviation at individual participant level, the BD group showed a wider range of values compared to others, reflecting higher within-group differences in depression variability. Nonetheless, the Kruskal-Wallis test suggested that these differences were not statistically significant. Figure 2 shows the distribution of PHQ-9 mean scores and standard deviations for each group.

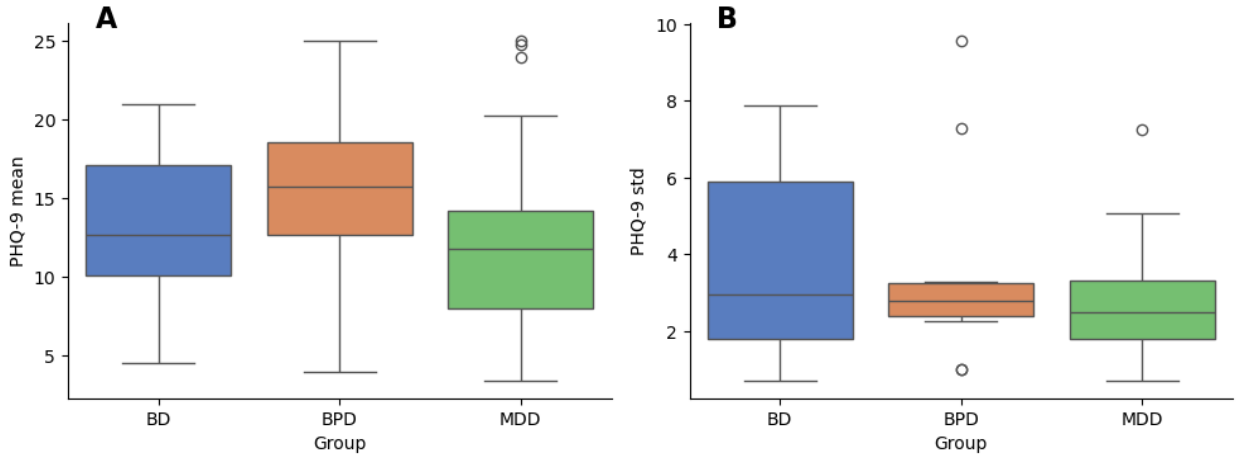

Figure 2. PHQ-9 score summary statistics: group-level means and individual-level variability. A) Boxplots of group-level PHQ-9 mean scores with standard deviations for BD (mean = 14.1, SD = 6.5), BPD (mean = 13.8, SD = 6.5), and MDD (mean = 12, SD = 6.6), showing no

statistically significant differences between groups. B) Boxplots of individual-level standard deviations in PHQ-9 scores within each group, indicating higher variability in the BD group compared to the others, suggesting greater fluctuations in depression severity among BD patients. However, these differences in individual-level variability are not statistically significant.

A comparison of depression severity levels across the groups showed that moderate depression severity (PHQ-9 scores from 10 to 14) is most frequent (27.5%). Moderately severe (PHQ-9 scores from 15 to 19) and severe levels (PHQ-9 scores from 20 to 27) were more frequent among BD (23.9% and 23.9%) and BPD (26.1% and 23.4%) participants compared to MDD participants (18.2% and 14.7%). For details, refer to Table S1 in Multimedia Appendix 2, which summarizes the frequencies of depression severity by patient groups.

A between-group comparison revealed how individual items contributed to the aggregated score, summarizing all responses from all participants in the group, as shown in Figure 3. The Kruskal-Wallis test, comparing average PHQ-9 item scores, identified items 5, 6, 7, 8, and 9 differing between the groups. Further, applying Mann-Whitney test and FDR correction using the BH-method at a significance level of α=0.05 reveals the differing groups for each item, as summarized in Table 1. For full details, refer to Table S2 in Multimedia Appendix 2.

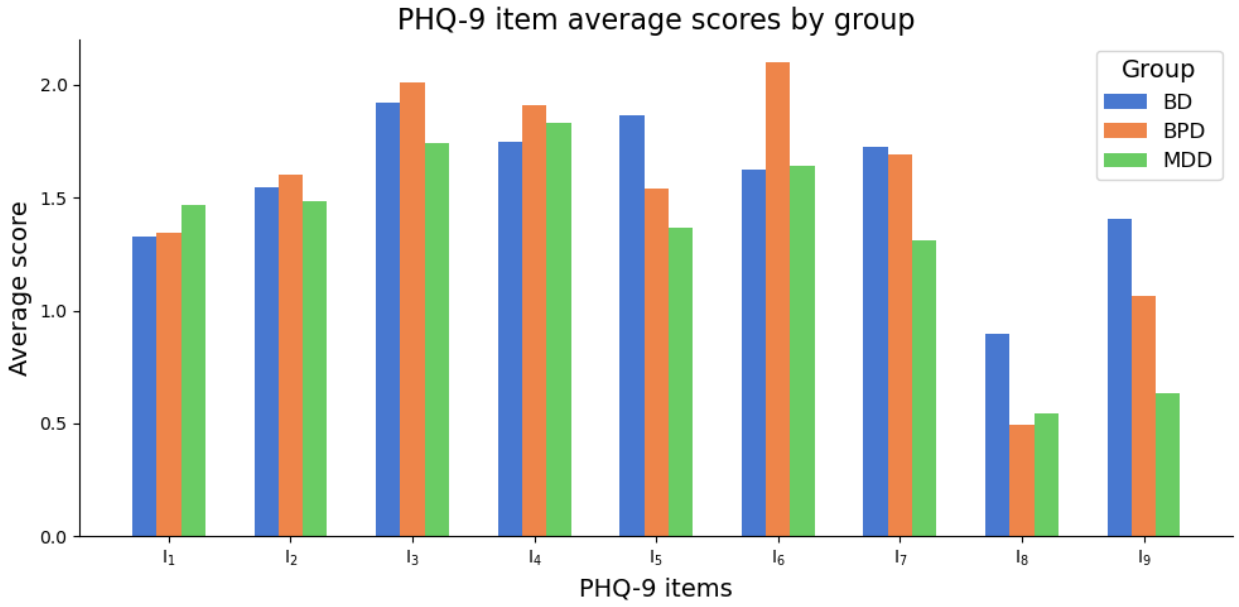

Figure 3. Average scores of individual PHQ-9 items by patient group. This summary includes responses from all participants in the groups without considering depression severity. The most notable differences between the groups are observed in items 5, 6, 7, 8, and 9.

		MDD vs. BD	MDD vs. BPD	BD vs. BPD
PHQ-9 item	**Item description**	**Corrected P value**	**Corrected P value**	**Corrected P value**
5.	Poor appetite or overeating	**.001**	.232	.085
6.	Feeling bad about oneself	.977	**<.001**	**.001**
7.	Trouble concentrating	**.001**	**.002**	.977
8.	Psychomotor changes	**.001**	.977	**.004**
9.	Suicidal thoughts	**<.001**	**<.001**	.083

Table 1. Mann-Whitney test P values for differences in PHQ-9 item scores between diagnosis groups. The table presents PHQ-9 item verbatim as phrased in the full questionnaire. Statistically significant differences between groups are observed for items 5, 6, 7, 8, and 9. The corrected P values correspond to FDR correction using the Benjamini-Hochberg method.

A comparison of the item distributions at different depression severity levels across the patient groups revealed notable differences. At the lowest depression level, items 3 and 4 are more pronounced in the BPD group, while item 7 has higher scores for the BD group than others. At the mild and moderate depression levels, the items are more evenly distributed across all groups. At the moderately severe level, item 9 scores higher in the BD group, while item 1 is pronounced in the MDD group, and the BPD group also exhibits higher scores for item 6. Finally, at the highest severity level, the BD group differs from the BPD group, particularly in items 8 and 9. Figure 4 illustrates these differences in distributions across groups and depression levels.

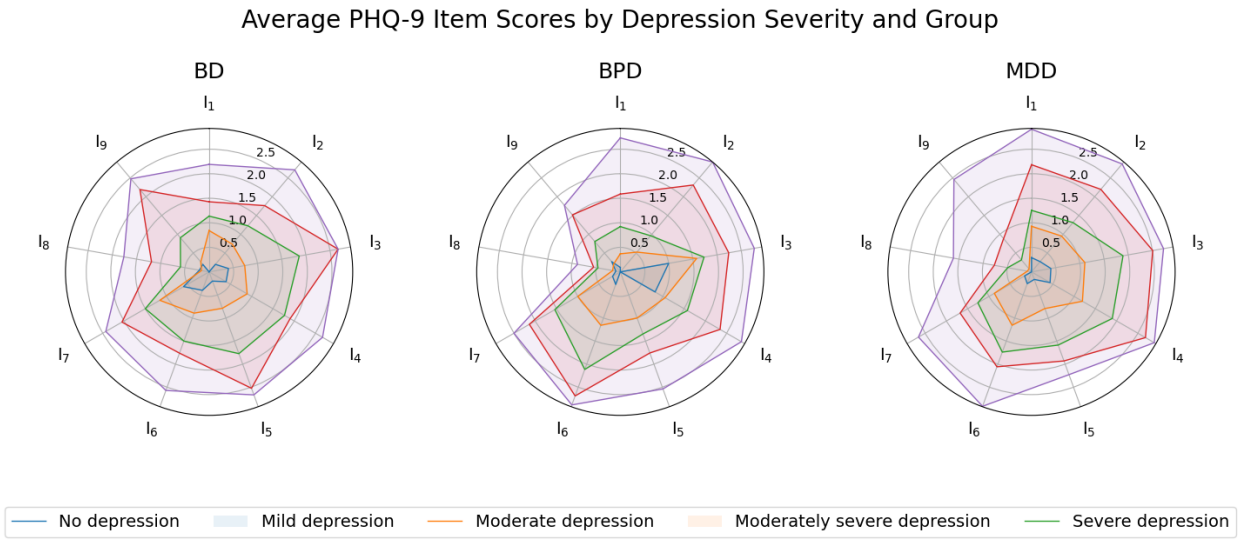

Figure 4. Average PHQ-9 item scores by diagnosis group and depression severity. The radar charts show the distribution of PHQ-9 item scores across different depression severity levels for BD, BPD, and MDD groups. Letters from I_1 to I_9 denote the items and different colors represent severity levels. The figure illustrates how symptom patterns vary across groups and severity levels.

Further exploration of the item score distributions across depression severity levels revealed that average scores for individual PHQ-9 items do not increase at the same rate or linearly with the total score. Item 3 (relating to sleeping problems) and item 4 (feeling tired and having low energy), constantly contributed to the total score across all the severity levels. Also, item 6 (feeling bad about yourself) had high scores across the severity levels. On the other hand, item 8 (regarding moving or speaking unusually fast or slow) stayed relatively low across all depression severity levels. It is worth noting that items 3, 4, 5, and 6 showed higher relative changes, transiting between mild depression and moderate depression. Further, item 9 (regarding suicidal thoughts) only emerged at higher depression levels. For a detailed view, refer to Figure 5.

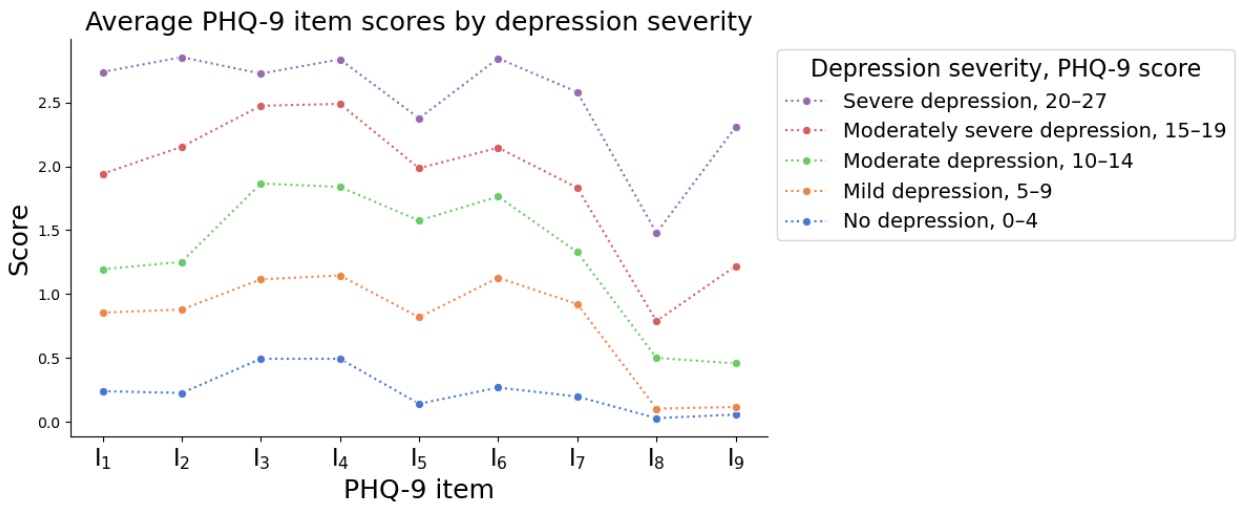

Figure 5. PHQ-9 item average score distributions across depression severity levels, combining the data from all patient groups. Comparison shows that the items neither increase at the same rate nor linearly, when depression severity increases. Notably, items 8 and 9 have higher values only at higher severity levels. Also, comparing the mild (scores from 5 to 9) and moderate (scores from 10 to 14) depression severity levels, items 3 to 6 exhibit the highest differences.

To conclude the PHQ-9 score composition analysis, we found statistically significant differences between the patient groups in how individual items contributed to the total score. This finding exemplified individual differences on depression symptom expression. We noticed that between mild and moderate depression, the highest differences were observed in items 3, 4, 5, and 6, thus these items might be particularly informative for detecting depression severity above the clinical threshold (corresponding to PHQ-9 score of 10).

Association analysis | Correlation analysis

To address the second objective, we examined how the associations between behavioral markers and depression severity varied across the study population, diagnostic group, and individual participant levels. This section summarizes the results of correlation analysis followed by exploration of temporal dynamics calculating rolling-window correlations.

Correlation analysis at the participant-level revealed a wider range of correlation values (mean=0.013, std=0.35), while analysis using full study population resulted in a narrower range of weak correlations (mean= 0.022, std=0.046). Comparing these approaches, pooling the data across the participants averages the individual differences and value fluctuations, leading to conservative correlation estimates. Contrarily, participant-level

correlations are affected by smaller observation counts and individual temporal dynamics (for example, autocorrelation), leading to a more diverse distribution. Figure 6 shows the distributions for pooled and participant-level approaches.

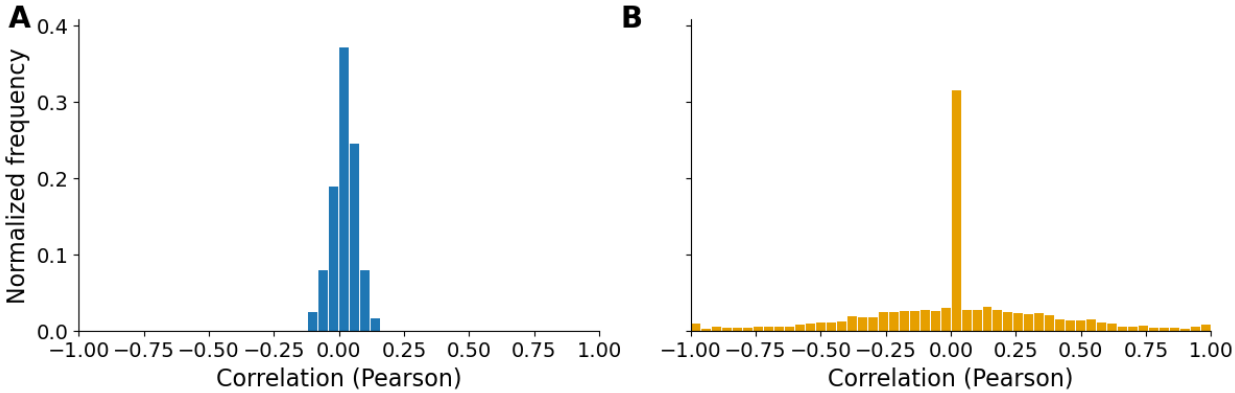

Figure 6: Distribution of Pearson correlation coefficients for (A) pooled (B) and participant-level correlations. We calculated both distributions employing within-person standardizing for behavioral features. (A) shows a distribution concentrated near zero (mean = 0.011, std = 0.021). In contrast, (B) shows a broader distribution (mean = 0.013, std = 0.35) with a peak near zero.. Additionally, a spike around zero reflects that many participants exhibit weak correlations between behavior and depression.

Pooled-participant analysis resulted in statistically significant correlation coefficients for three features out of 98 (2.4%), after controlling for multiple comparisons using FDR correction at a significance level of α=0.05. Specifically, battery level variability in the morning (r = 0.126, FDR-corrected p = 0.024) and in the evening (r = 0.118, FDR-corrected p = 0.038), as well as incoming call duration in the morning (r = 0.129, FDR-corrected p = 0.024), all features showing weak positive correlations.

Within-person correlation analysis yielded 8,128 individual correlation coefficients (each representing a unique participant-feature-depression score combination). Overall, 23.3% (1,890 out of 8,128) of correlations were positive, 20.6% (1,671 out of 8,128) were negative, and 56.2% (4,567 out of 8,128) showed low or no correlation. Further, we saw that the proportion of positive and negative correlations did not differ noticeably across the sensors. Smartphone app-related features had the highest proportion (65.4%) of low/no correlation. Battery features showed the highest proposition (37.7%) of positive correlations, and accelerometer features the highest proportion (33.5%) of negative correlations. Further, we compared the proportion of correlations across the participant groups. Generally, the proportion of low/no correlation was the highest among most sensors and groups. The BPD group showed the highest difference in negative and positive correlation proportions between groups, especially for battery, call, and screen usage-related features. The BD group exhibited the highest difference between negative and positive associations with sleep features. Figure 7 summarizes the correlation proportions across the sensors and groups.

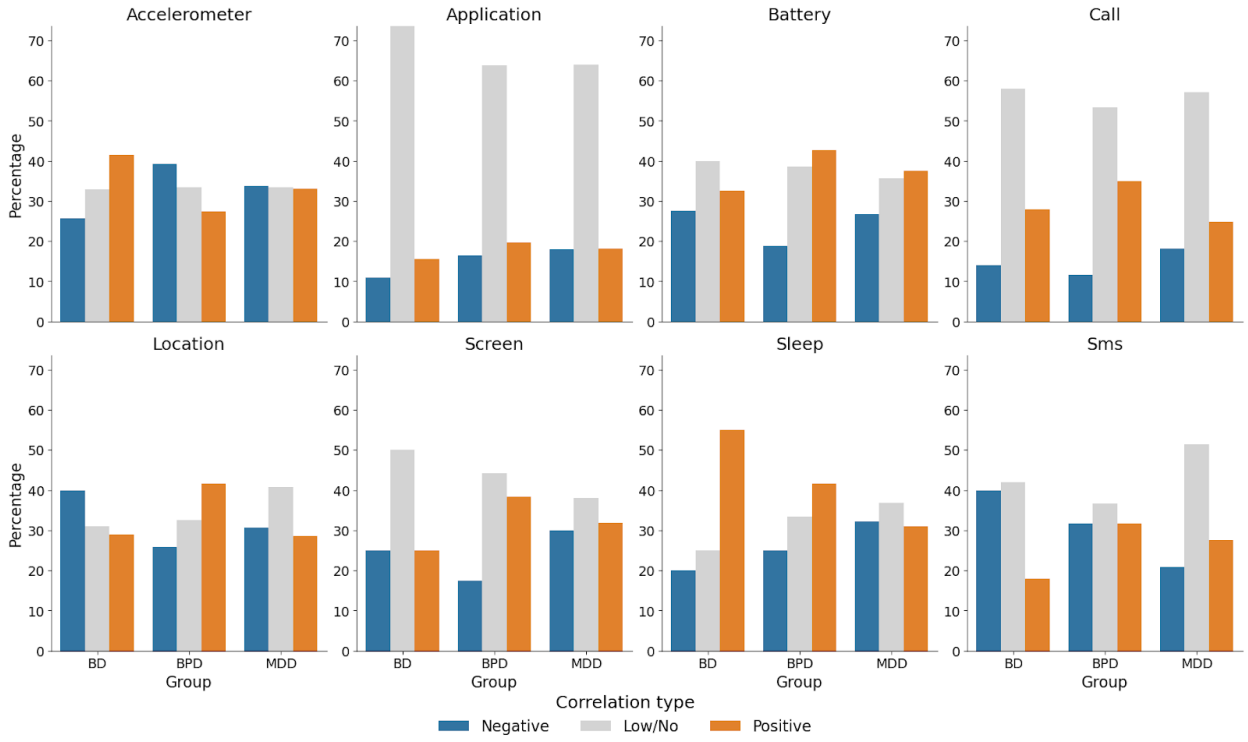

Figure 7. Correlation proportions across different behavioral data sensors. The BD group generally shows a higher proportion of low/no correlations across the sensor than other groups. Application-related features show relatively even distributions of negative and positive correlations across the groups. The BPD group shows the highest proportion of positive correlations for battery, phone calls, and screen usage-related features. For accelerometer features, BD shows a higher proportion for positive and BPD a higher proportion for negative correlations. For location features, BPD shows a higher proportion of positive correlations compared to negative. For SMS count-related features, the BD group shows a higher proportion of negative than positive correlations.

Finally, we ranked the features showing the highest proportion of positive, negative, and no or low correlations. Features with the least positive or negative correlations included app usage count and call durations at night. The highest positive correlations included call durations in the morning, daily outgoing call durations and counts, afternoon battery standard deviation, daily incoming SMS count, average sleep duration, and accelerometer standard deviation. The highest negative correlations included afternoon outgoing call duration, time spent at the single location and home, daily screen use and screen on durations, daily outgoing call count, and daily incoming SMS count. For more details, refer to Tables S1, S2, and S3 in Multimedia Appendix 3.

Finally, rolling-window correlation analysis revealed that most participants exhibited both negative and positive correlations across the features. Highest proportion of mixed

correlations 92.2% (59 / 64 participants) were found on screen usage duration in the afternoon, battery level and variability in the afternoon, screen usage duration, and screen on duration in the evening. Lowest proportions of mixed associations were found within application-usage features, travel application duration at evening and at morning having 4.7% (3 / 64) participants exhibiting mixed associations. Figure 8 exemplifies the rolling correlation between mean sleep duration and PHQ-9 score for two participants.

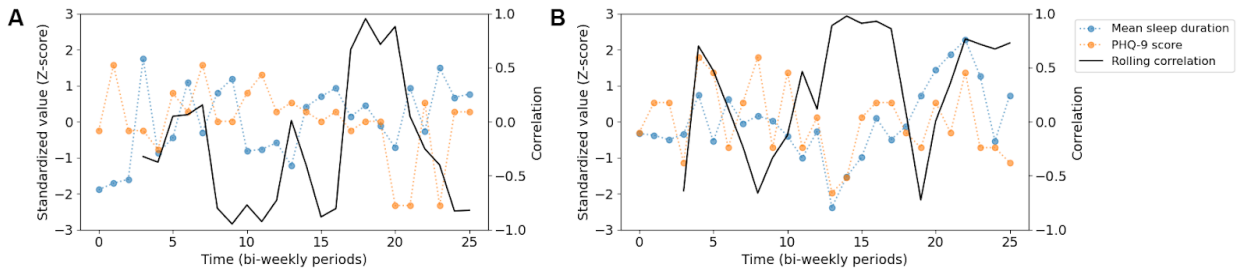

Figure 8. Rolling window correlations between mean sleep duration and PHQ-9 scores for two participants. The data is normalized within a participant, values showing the difference in standard deviations to mean value of 0. Both participants have provided full year (26 bi-weekly periods) of data. A) The Pearson correlation calculated by pooling all the data points yields moderate positive correlation coefficient of 0.34. Rolling window correlations revealed periods of both negative and positive correlations. B) The Pearson correlation calculated by pooling all the data points yields moderate negative correlation coefficient of -0.30. Rolling window correlations revealed periods of both negative and positive correlations. This example shows that calculating correlations across all the data points can mask the temporal dynamics in correlations, and lead into misleading conclusions.

In summary, the results implied that most observed behavioral data features across the sensors were not highly or consistently correlated with PHQ-9 depression scores. In general, the proportions of positive and negative correlations were not markedly different. However, group-wise inspection revealed differences in correlation proportions between the groups, implying between-participant variability. Finally, rolling window analysis exemplified time-varying dynamics, fluctuating periods of positive and negative correlations between behavior and depression severity.

Association analysis | Permutation tests

Permutation tests and the Kolmogorov-Smirnov test for correlation distributions initially identified 11 behavioral features with distributions that differed from randomized data. However, after applying FDR correction, none of these remained statistically significant. Despite this, their broader correlation distributions suggest that these features may still capture meaningful associations with depression and warrant further investigation. These features included incoming and outcoming call durations at different times of day, acceleration variability in the evening, screen use time during the morning, time spent at the most frequently visited place and home, battery level variability in the morning and

evening, mean sleep duration, communication application usage during the night, and incoming SMS count at morning. Figure S1 in Multimedia Appendix 4 displays the original and permuted correlation distributions.

Multilevel modeling

To address the third objective, we used multilevel modeling to examine both within- and between-participant variability in behavioral features and identify potential behavioral markers indicative of depression. This section outlines the development and comparison of multilevel models, starting with a baseline model and stepwise addition of demographic, contextual, and behavioral predictors.

For modeling, we selected behavioral features identified by a permutation test to examine within- and between-participant variability in behavior–depression associations. The baseline model, which included only random intercepts for each participant, yielded a R^2-score of 0.585, suggesting that 58.5% of the variability in depression severity (PHQ-9 score) was attributable to between-participant differences. The first model added background and contextual information, increasing the conditional R^2-score (accounting for both fixed and random effects) to 0.615 and the marginal R^2-score (reflecting only fixed effects) to 0.093. The second model, added time-varying within-person features, further increased the conditional R2 to 0.637 and the marginal R^2-score to 0.118. Finally, the complete model combines background, contextual, and within- and between-person behavioral features, resulting in a conditional R^2-score of 0.679 and a marginal R^2-score of 0.163.

The results from multilevel models showed that the intraclass correlation coefficient (ICC) was relatively high across models (ranging from 59% to 62%), indicating that a substantial proportion of the variability in self-reported depression severity was attributed to the between-participant differences. Model comparisons with ANOVA revealed that adding within-participant behavioral features improved model fit over the baseline model ($\chi 2$ = 52.46, df=14, P<.001), demonstrating the importance of time-varying predictors. However, incorporating between-person features did not lead to further improvement ($\chi 2$=9.03, df=12, P=.70). Based on the comparison results, we concluded that the second model provided the best fit.

The best performing model coefficients suggested the following within-participant associations in the sample: 1) higher communication application usage during night compared to individuals' average was associated with higher depression (β = 0.31, 95% CI = [0.01-0.61], P = .042), 2) higher battery level variability at morning compared to individuals' average was related associated with higher depression (β = 0.39, 95% CI = [0.06-0.72], P = .019), 3) higher screen usage duration at morning compared to individuals'

average was associated with higher depression (β = 0.35, 95% CI = [0.04-0.65], P = .027), 4) higher average sleep duration compared to individuals' average was associated with higher depression (β = 0.39, 95% CI = [0.11-0.67], P = .006), 5) higher incoming SMS count at morning compared to individuals' average was associated with higher depression (β = 0.36, 95% CI = [0.05-0.66], P = .022), and 6) higher daily SMS count (sent and received) compared to individuals' average was associated with lower depression (β = -0.36, 95% CI = [-0.-0.66], P = .022). Additionally, the estimated intercept for self-assessed depression is 14.36 (95% CI =[8.31-20.40], P<.001), indicating that, on average, the participants reported moderate levels of depressive symptoms on the PHQ-9 scale.

The identified features can be grouped into sociability, smartphone usage, and sleep-related features in the context of behavioral aspects of interest. Communication app usage at night, SMS count, and call duration may be used as a proxy for sociability. Compared at the participant level, increased temporal features, nighttime communication application usage, and incoming SMS count were both associated with higher depression. Additionally, increased outgoing SMS count (a volume-based feature) was associated with lower depression. For sleep, increased sleep duration compared to the individual's average was associated with higher depression. Phone interactions showed that increased phone usage during the morning was associated with higher depression. Battery level variability implies that the phone is used, charged, or both, thus may be used as a proxy for participants' phone usage. In contrast, screen usage duration directly measures the time the phone is interacted with. While background and contextual factors did not show any significant features, regular work shifts and having kids have both negative coefficient estimates. Table 2 shows the condensed model summaries; for the details, refer to Table S1 in Multimedia Appendix 5. For parsimony, table 2 only displays the features with P values below 0.05.

	Model 1[a]			**Model 2[b]**			**Model 3[c]**		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept)	14.37	8.32 – 20.41	**<0.001**	14.36	8.31 – 20.40	**<0.001**	13.41	6.11 – 20.70	**<0.001**
shift [regular]	-3.32	-6.66 – 0.02	0.051	-3.32	-6.66 – 0.02	0.051	-2.23	-6.06 – 1.59	0.252
kids [Yes]	-2.37	-5.38 – 0.64	0.122	-2.37	-5.38 – 0.63	0.121	-3	-6.92 – 0.92	0.133
Communications application usage duration night W	—	—	—	0.31	0.01 – 0.61	**0.042**	0.31	0.01 – 0.61	**0.042**

Battery level std morning W	—	—	—	0.39	0.06 – 0.72	**0.019**	0.39	0.06 – 0.72	**0.019**
Screen usage duration morning W	—	—	—	0.35	0.04 – 0.65	**0.027**	0.35	0.04 – 0.65	**0.027**
Average sleep duration W	—	—	—	0.39	0.11 – 0.67	**0.006**	0.39	0.11 – 0.67	**0.006**
Incoming SMS count morning W	—	—	—	0.36	0.05 – 0.66	**0.022**	0.36	0.05 – 0.66	**0.022**
Outgoing SMS count daily W	—	—	—	-0.36	-0.66 – -0.06	**0.018**	-0.36	-0.66 – -0.06	**0.018**
Random Effects									
σ2	15.86			15.04			15.05		
τ00	21.50 participant			21.57 participant			24.14 participant		
ICC	0.58			0.59			0.62		
N	63 participant			63 participant			63 participant		
Observations	783			783			783		
Marginal R2 / Conditional R2	0.093 / 0.615			0.118 / 0.637			0.163 / 0.679		

Table 2. Multilevel model summaries. The predictors found statistically significant at least in one model are shown at the table. Models show association with higher depression severity for increased within-participant 1) communication application usage at night, 2) battery level variability at morning, 3) screen usage at morning, 4) average sleep duration, and 5) incoming SMS count at morning. Conversely, increasing daily outgoing SMS count shows an association with lower depression severity. Notably, between-participant effects did not show any statistically significant associations.
W denotes within-person effect
B denotes between-person effect
[a] *includes background demographics and contextual predictors, and random intercepts for each participant*
[b] *Includes behavioral features, within-participant and between-participant effects, and random intercepts*
[c] *includes predictors from model 1 and model 2*

4. Discussion

Key findings

In this study, we investigated the variability in symptoms of depression through the angle of subjective data. We used self-assessment of depression symptoms from questionnaires. We

also looked at the variability of depressive symptoms by looking at their association with behavioral markers gathered through smartphones. We used multilevel modeling to explore associations at both within- and between-participant levels and to identify potential behavioral markers indicative of depression. We focused on understanding how this variability affects the modeling of depression. We observed differences in responses to five PHQ-9 items between the patient groups. In terms of behavioral markers, associations with depression severity varied considerably within and between participants. Notably, most of the behavioral markers exhibit no association with changes in depression. Furthermore, the proportion of positive and negative correlations was relatively similar across all diagnostic groups and features. The multilevel model results indicate that approximately 58.5% of the variability in self-assessed depression is attributed to differences between participants. The remaining 41.5% reflects within-participant variation, including time-varying factors and other unexplained sources. Adding demographic background, context, and time-varying behavioral features improves the model's explanatory power by up to 63.7%.

Implications

The differences in PHQ-9 item distributions indicated variability in depression representation across the study subgroups. This finding may serve as proof of the principle that distinct subgroups exist within depressive disorder; a unique symptom profile may characterize each subgroup. However, as the subgroups in this study were small, caution should be exercised when generalizing such findings. Further, the differences suggest that depression modeling should accommodate individual differences in depression manifestations related to the symptoms by either modeling each participant individually or pooling together participants with similar symptoms. For example, the depression symptom profiling approach introduced by Akbarova et al. [29] aimed to enhance the model's accuracy by profiling the participants by the importance of each PHQ-9 item for each participant. Further, the highest difference in PHQ-9 score distributions between the diagnostic groups between mild and moderate depression occurs on items 3 (reflecting trouble in sleep), 4 (feeling tired), 5 (eating problems), and 6 (self-worth concerns). These findings suggest that these symptoms most clearly reflect the transition from mild to moderate depression. Thus, the transition may be identifiable through monitoring behavior related to these items, e.g., sleep disturbances, daily activity and mobility levels, and sociability. Additionally, items 8 (reflecting psychomotor retardation) exhibit higher scores only at higher depression severity levels, aligning with clinical knowledge that psychomotor retardation typically occurs only in severe depression and is an important characteristic of the DSM-5 melancholic specifier or subtype of major depressive episodes [30,31]. Similarly, item 9 (reflecting suicidal ideation), gets higher scores only at higher

depression levels. This finding is consistent with other studies, and is a clinically important point - alleviation of depression results in alleviation of suicidal thoughts [32].

Correlation analysis results indicated that most behavioral data features collected within the study were not highly or consistently correlated with PHQ-9 depression scores. For applications, one possible explanation of these results could be the infrequent usage of smartphone applications among the participants. Also, for some participants, the low variability in PHQ-9 scores could lead to low correlations. On the other hand, mixed correlations indicate differing behavior–depression associations between the participants. While the results show no consistent association between any behavioral feature and self-reported depression across the study population, some participants did express constant and high correlations. Rolling correlation analysis further exemplified the within-person variability in associations. While the correlation between a feature and depression severity shows moderate positive or negative correlation across the study range, rolling-window correlation reveals periods with varying correlation strengths and directions. These time-varying correlations further complicate the depression modeling, indicating that the relationship between behavior and depression severity is not static but may change over time.

Permutation test reveals further insight into correlations. While none of the relationships between a behavior and feature remained statistically significant after controlling for false positive discovery ratio, the results suggest focusing on the 11 identified behavioral features in future research. Further, the results hint at conducting a power analysis or simulation study to estimate the sufficient study size to detect the differences suggested by permutation tests. This would be an important step for planning the future studies and participant recruitment.

The multilevel model comparison revealed the differences between the participants, modeled by random intercepts for each individual, explain a substantial portion of the variability in self-reported depression. This is partly due to the stability of depression severity throughout the study, already reported in an earlier study [20]. Adding background and contextual predictors, and time-varying within-participant behavioral features improve the model explanatory power from 58.5% to 63.7%. While adding between-person predictors improved the explanatory power up to 67.9%, ANOVA implied no further improvement. These findings suggest that time-varying , within-person predictors are more important for modeling, than stable between-person predictors. This might be because the within-person predictors already capture some between-person differences, or because standardized participant mean centering may not be ideal for assessing between-person effects.

The modeling results, regarding explanatory power of random and fixed effects, align with other studies using hierarchical models (for example [33,34]). Adding behavioral features increase the explanatory power by 2.2%, indicating weak linear relationship between the passively sensed behavior and self-reported depression.

Finally, mixed behavior–depression correlations and the significant portion of PHQ-9 variance explained by between-person variability, support using personalized modeling for depression detection. Further, these findings suggest focusing on depression at the symptom level and pre-selecting the behavioral features, demonstrating associations for each participant individually or for sub-groups with similar behavior–depression associations. Moreover, between-participant variability supports the importance of model personalization over the generalized study population-based model for more accurate depression prediction.

Limitations

The validity and generalizability of this study's findings are primarily limited by the amount of available data, by sample size, participant dropping out the study, and missing intermediate observations. Limited data is a common issue across digital phenotyping studies. While the quantity of passively collected raw data is relatively high, more bi-weekly sampled active data for each participant would have benefited the analysis. More data would lead to greater statistical power for detecting the correlations, and make modeling less prone to overfitting.

Additionally, the simple median imputation technique used to fill in missing data may bias the analysis results. This issue could be mitigated by thoroughly analyzing the nature of missing observations (whether they are missing completely at random, missing at random, or missing not at random) and thorough comparison between different imputation techniques.

Moreover, the accuracy of the analysis may be affected by the potential biases in the high-level features, extracted from the raw data as proxies for behavior. Comparing the features across other datasets and studies would help to assess their robustness and alignment with findings in the literature. Likewise, PHQ-9 scores have inherent limitations due to subjectivity and recall biases.

Furthermore, the multilevel modeling used for analysis has also limitations, as it captures only linear relationships. Thus, advanced modeling techniques, capable of modeling non-linear relationships, feature interactions and temporal dependencies would be needed to address this.

Future directions

Acknowledging that a relatively small sample size hinders the generalization of most digital phenotyping studies, we emphasize the importance of utilizing the existing data fully. While designing and conducting large-scale digital phenotyping studies is challenging, combining data sets across different studies may offer a feasible approach to obtaining larger and heterogeneous samples. Further, frequently sampled passive data may offer insights into individual differences in depressive symptoms and associated behavior in more detail, as understanding these differences will help develop accurate, personalized models. Additionally, understanding individual variability is an important prerequisite for assessing the sufficient sample size to detect the differences between participant behavior. The ICC, effect sizes, and proportion of missing data reported in this study and others can be utilized to plan future research and run a-priori power analysis for required sample sizes.

Due to the observed relatively weak and inconsistent pairwise behavior–depression correlations, we recommend using multiple features in future modeling, exploring feature interactions and non-linear relationships in the data, and personalizing the models. It is also worth exploring whether the changes in behavior are more common in moderate to severe depression. If sensitive digital phenotyping measures for the milder range of depression are to be explored, they probably must tackle more subjective and internal aspects of depression. Achieving this might require measures that are sensitive to the negative emotional and self-referential cognitive bias in depression, such as changes in speech prosody or linguistic contents of written text (for example in emails, and social media platforms), consumed media contents, or facial expressions.

5. Conclusions

This study offers a comprehensive exploration of self-reported depressive symptoms and associated mobile-sensed behaviors by disentangling both within-person and between-person effects using mobile-sensed data from clinically diagnosed patients experiencing major depressive episodes. The results reveal depressive symptoms that should be focused on when discriminating the depressed and non-depressed patients with digital phenotyping. The association analysis points out that not all behavioral markers are uniformly related to depression, and thus not relevant across the study participants. Finally, our multilevel modeling revealed that a substantial portion of depression variability is due to individual differences. Hence, future research should focus on improving the depression prediction and forecasting accuracy by model personalization. Through these results, we emphasize the importance of addressing the heterogeneity within study population due to the critical information it provides for future research design, identifying useful behavioral markers, and study sample size optimization with power analysis.

Acknowledgments

The computational resources provided by the Aalto Science-IT project are gratefully acknowledged. The collection of data used in this manuscript was made possible with the help of Richard Darst, Jesper Ekelund, Roope Heikkilä, Joel Holmén, Kirsi Riihimäki, and Outi Saleva. We thank them for their valuable efforts at different stages. This research was supported by Wihuri Foundation through a personal research grant to Arsi Ikäheimonen.

Conflicts of Interest

None to declare.

Abbreviations

BD: Bipolar Disorder

BH: Benjamini-Hochberg

BPD: Borderline Personality Disorder

CI: Confidence Interval

FDR: False Discovery Rate

GPS: Global Positioning System

HUS: Helsinki and Uusimaa Hospital District

ICC: Intraclass Correlation Coefficient

MDD: Major Depressive Disorder

PHQ-9: 9-item Patient Health Questionnaire

SMS: Short Message Service

References

1. Ferrari AJ, Santomauro DF, Aali A, et al. Global incidence, prevalence, years lived with disability (YLDs), disability-adjusted life-years (DALYs), and healthy life expectancy (HALE) for 371 diseases and injuries in 204 countries and territories and 811 subnational locations, 1990–2021: a systematic analysis for the Global Burden of

Disease Study 2021. *The Lancet*. 2024;403(10440):2133-2161. doi:10.1016/S0140-6736(24)00757-8

2. Global Burden of Disease 2021: Findings from the GBD 2021 Study | Institute for Health Metrics and Evaluation. Accessed February 18, 2025. https://www.healthdata.org/research-analysis/library/global-burden-disease-2021-findings-gbd-2021-study

3. Simon GE, Moise N, Mohr DC. Management of Depression in Adults: A Review. *JAMA*. 2024;332(2):141. doi:10.1001/jama.2024.5756

4. James TA, Weiss-Cowie S, Hopton Z, Verhaeghen P, Dotson VM, Duarte A. Depression and episodic memory across the adult lifespan: A meta-analytic review. *Psychol Bull*. 2021;147(11):1184.

5. Spiller TR, Duek O, Helmer M, et al. Unveiling the Structure in Mental Disorder Presentations. *JAMA Psychiatry*. Published online August 7, 2024. doi:10.1001/jamapsychiatry.2024.2047

6. Torous J, Kiang MV, Lorme J, Onnela JP. New Tools for New Research in Psychiatry: A Scalable and Customizable Platform to Empower Data Driven Smartphone Research. *JMIR Ment Health*. 2016;3(2):e16. doi:10.2196/mental.5165

7. Masud MT, Mamun MA, Thapa K, Lee DH, Griffiths MD, Yang SH. Unobtrusive monitoring of behavior and movement patterns to detect clinical depression severity level via smartphone. *J Biomed Inform*. 2020;103:103371. doi:10.1016/j.jbi.2019.103371

8. Cao J, Truong AL, Banu S, Shah AA, Sabharwal A, Moukaddam N. Tracking and Predicting Depressive Symptoms of Adolescents Using Smartphone-Based Self-Reports, Parental Evaluations, and Passive Phone Sensor Data: Development and Usability Study. *JMIR Ment Health*. 2020;7(1):e14045. doi:10.2196/14045

9. Xu X, Liu X, Zhang H, et al. GLOBEM: Cross-Dataset Generalization of Longitudinal Human Behavior Modeling. *Proc ACM Interact Mob Wearable Ubiquitous Technol*. 2023;6(4):190:1-190:34. doi:10.1145/3569485

10. Bufano P, Laurino M, Said S, Tognetti A, Menicucci D. Digital Phenotyping for Monitoring Mental Disorders: Systematic Review. *J Med Internet Res*. 2023;25:e46778. doi:10.2196/46778

11. Leaning IE, Ikani N, Savage HS, et al. From smartphone data to clinically relevant predictions: A systematic review of digital phenotyping methods in depression. *Neurosci Biobehav Rev*. 2024;158:105541. doi:10.1016/j.neubiorev.2024.105541

12. De Angel V, Lewis S, White K, et al. Digital health tools for the passive monitoring of depression: a systematic review of methods. *Npj Digit Med*. 2022;5(1):3.

doi:10.1038/s41746-021-00548-8

13. Moura I, Teles A, Viana D, Marques J, Coutinho L, Silva F. Digital Phenotyping of Mental Health using multimodal sensing of multiple situations of interest: A Systematic Literature Review. *J Biomed Inform*. 2023;138:104278. doi:10.1016/j.jbi.2022.104278

14. Rohani DA, Faurholt-Jepsen M, Kessing LV, Bardram JE. Correlations Between Objective Behavioral Features Collected From Mobile and Wearable Devices and Depressive Mood Symptoms in Patients With Affective Disorders: Systematic Review. *JMIR MHealth UHealth*. 2018;6(8):e9691. doi:10.2196/mhealth.9691

15. Baryshnikov I, Aledavood T, Rosenström T, et al. Relationship between daily rated depression symptom severity and the retrospective self-report on PHQ-9: A prospective ecological momentary assessment study on 80 psychiatric outpatients. *J Affect Disord*. 2023;324:170-174. doi:10.1016/j.jad.2022.12.127

16. Kroenke K, Spitzer RL. The PHQ-9: A New Depression Diagnostic and Severity Measure. *Psychiatr Ann*. 2002;32(9):509-515. doi:10.3928/0048-5713-20020901-06

17. Aledavood T, Luong N, Baryshnikov I, et al. Multimodal Digital Phenotyping Study in Patients With Major Depressive Episodes and Healthy Controls (Mobile Monitoring of Mood): Observational Longitudinal Study. *JMIR Ment Health*. 2025;12(1):e63622. doi:10.2196/63622

18. Sheehan DV. 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*.

19. First MB, Benjamin LS, Gibbon M, Spitzer RL, Williams JB. *Structured Clinical Interview for DSM-IV Axis II Personality Disorders*. American Psychiatric Press; 1997.

20. Ikäheimonen A, Luong N, Baryshnikov I, et al. Predicting and Monitoring Symptoms in Patients Diagnosed With Depression Using Smartphone Data: Observational Study. *J Med Internet Res*. 2024;26:e56874. doi:10.2196/56874

21. Levis B, Benedetti A, Thombs BD. Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis. *BMJ*. Published online April 9, 2019:l1476. doi:10.1136/bmj.l1476

22. Hale T, Angrist N, Goldszmidt R, et al. A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nat Hum Behav*. 2021;5(4):529-538. doi:10.1038/s41562-021-01079-8

23. Ikäheimonen A, Triana AM, Luong N, et al. Niimpy: A toolbox for behavioral data analysis. *SoftwareX*. 2023;23:101472. doi:10.1016/j.softx.2023.101472

24. Aledavood T, Kivimäki I, Lehmann S, Saramäki J. Quantifying daily rhythms with

non-negative matrix factorization applied to mobile phone data. *Sci Rep*. 2022;12(1):5544. doi:10.1038/s41598-022-09273-y

25. Curran PJ, Bauer DJ. The Disaggregation of Within-Person and Between-Person Effects in Longitudinal Models of Change. *Annu Rev Psychol*. 2011;62(1):583-619. doi:10.1146/annurev.psych.093008.100356

26. Kruskal WH, Wallis WA. Use of Ranks in One-Criterion Variance Analysis. *J Am Stat Assoc*. 1952;47(260):583-621. doi:10.1080/01621459.1952.10483441

27. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Ser B Methodol*. 1995;57(1):289-300. doi:10.1111/j.2517-6161.1995.tb02031.x

28. Darling DA. The kolmogorov-smirnov, cramer-von mises tests. *Ann Math Stat*. Published online 1957:823-838.

29. Akbarova S, Im M, Kim S, et al. Improving Depression Severity Prediction from Passive Sensing: Symptom-Profiling Approach. *Sensors*. 2023;23(21):8866. doi:10.3390/s23218866

30. American Psychiatric Association, ed. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5-TR™*. Fifth edition, text revision. American Psychiatric Association Publishing; 2022.

31. Buyukdura JS, McClintock SM, Croarkin PE. Psychomotor retardation in depression: Biological underpinnings, measurement, and treatment. *Prog Neuropsychopharmacol Biol Psychiatry*. 2011;35(2):395-409. doi:10.1016/j.pnpbp.2010.10.019

32. Söderholm JJ, Socada JL, Rosenström TH, Ekelund J, Isometsä E. Borderline personality disorder and depression severity predict suicidal outcomes: A six-month prospective cohort study of depression, bipolar depression, and borderline personality disorder. *Acta Psychiatr Scand*. 2023;148(3):222-232. doi:10.1111/acps.13586

33. Zhang Y, Folarin AA, Sun S, et al. Longitudinal Relationships Between Depressive Symptom Severity and Phone-Measured Mobility: Dynamic Structural Equation Modeling Study. *JMIR Ment Health*. 2022;9(3):e34898. doi:10.2196/34898

34. Stamatis CA, Meyerhoff J, Meng Y, et al. Differential temporal utility of passively sensed smartphone features for depression and anxiety symptom prediction: a longitudinal cohort study. *Npj Ment Health Res*. 2024;3(1):1. doi:10.1038/s44184-023-00041-y