Original Paper

Variability in self-reported depression symptomology and associated behavioral markers in digital phenotyping

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:
Talayeh Aledavood
Department of Computer Science
Aalto University
FI-02150 Espoo, Finland
Email: talayeh.aledavood@aalto.fi

Abstract

Digital phenotyping studies using smartphone-sensed data have identified several behavioral markers associated with depression. However, the generalizability of these markers is constrained by multiple factors, including variability in depressive symptoms and associated behaviors, both between and within individuals over time. This study examines heterogeneity in depression and aims to identify behavioral markers indicative of depression in smartphone-sensed data collected from participants diagnosed with depression. We analyzed smartphone-sensed behavioral data from 62 patients with major depressive episodes across three subgroups: major depressive disorder (MDD, n=41), borderline personality disorder (BPD, n=12), and bipolar disorder (BD, n=9). Depression symptoms were assessed with the 9-item Patient Health Questionnaire (PHQ-9). Symptoms varied between subgroups and across severity levels. Association analysis revealed variability in correlations between depression severity and behavioral markers, both between participants and over time. Multilevel modeling revealed demographic predictors: employment status (β = -4.79, 95% CI = [-7.65, -1.80], P = .004) and age (β = -0.12, 95% CI = [-0.25, -0.00], P = .050) and lower nighttime movement (β = -0.79, 95% CI = [-1.29, -0.29], P = .024), as behavioral markers of depression.

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

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 of depressive symptoms [5] adds 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 depression symptoms objectively. Digital phenotyping is a field of research that utilizes personal digital devices to measure human behavior and physiology unobtrusively and continuously in real-life settings [6]. Collecting and analyzing data on individuals' behavior using these devices can provide new information about mental health disorders, including depression and behavioral patterns associated with them [7]. Ideally, the knowledge produced by digital phenotyping research could be applied in practice as new methods and tools for symptom monitoring and prediction, supplementing current means of mental healthcare. Furthermore, these methods could help design early warning systems, which can be integrated with interventions for symptom worsening [8].

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 [9]. The studies typically rely on behavioral data representing participants' physical activity, mobility, sociability, sleep, and smartphone usage patterns [10]. Although 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) (see, e.g., [10,11]), implying that the results may not generalize well across different samples and contexts. The differences may arise partly from heterogeneity in study objectives, samples, used methodologies, and reporting practices [10], device usage differences between the study participants [12], but also from the differences in depressive symptoms between participants [13] and temporal variability of symptoms within the participants [14–18]. The heterogeneity in depression symptoms and related observable behavior characterizes the depressive disorder; for example, both insomnia and hypersomnia, decrease or increase in appetite, are regarded as depression symptoms [19]. Further, research has identified heterogeneous temporal characteristics of individual depression symptoms [20] and variability in long-term trajectories of depression severity and stability [21].

This study aims to explore variability in depression symptoms and related behavioral markers as measured through digital phenotyping. Moreover, it assesses the underexplored contribution of between- and within-participant effects in behavior-depression relationships, aiming to identify markers associated with depression. The distinction between these effects is important for evaluating how both long-term, between-participant differences in behavior and short-term fluctuations within participants' behavior reflect depression severity. 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 and their associations with depression severity. 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 explore correlations between behavioral markers and depression across the study population, at the participant level, and over time. The third objective was to assess both within- and between-participant variability in behavioral features using multilevel modeling and identify behavioral markers indicative of depression while controlling for participants' background and contextual factors.

The study analyzed smartphone-collected behavioral data collected from outpatients diagnosed with depression. Self-reported scores from the 9-item Patient Health Questionnaire (PHQ-9)[22] were used to assess the severity of depression. We extracted features from smartphone data, reflecting daily behavioral aspects: *physical activity*, *mobility*, *sociability*, *phone usage*, and *sleep*. These features served as proxies for behavioral patterns, possibly associated with depression.

Exploratory analysis found heterogeneity in depressive symptom representations between the patient groups and in behavior-depression associations across the study population. The association exhibited variability in strength and direction (positive and negative) over time within the participants. Furthermore, multilevel modeling demonstrated that the variability over time in behavior within individual participants was more predictive of depression severity than the behavioral differences between participants. These results suggest that the predictive power of smartphone-sensed behavioral markers may be limited when simple, unimodal linear models are applied at the population level (assuming that the same behavioral markers have similar associations with depression across participants). On the contrary, results indicate that modeling should focus on the behavioral changes over time within the participants and on individual symptoms for more accurate depression monitoring and prediction.

Results

PHQ-9 questionnaire analysis

Our first objective was to assess the variability in depression symptoms through PHQ-9 answers at item (symptom), diagnostic group, and severity levels. The summary statistics showed that the MDD patients had, on average, the lowest average depression level (mean=11.6, SD = 6.0), while BPD had the highest average (mean=14.9, SD=4.8). At the study population level, moderate depression severity (PHQ-9 scores from 10 to 14) was the most frequent (27.5%). For a group-level summary of depression severity, refer to Table S1 in Multimedia Appendix 1. Considering PHQ-9 score variability, the BD group showed a wider distribution than others, reflecting higher within-group differences in depression variability. Nonetheless, the Kruskal-Wallis test [23] indicated that these differences were not statistically significant. Figure 1 shows the distribution of PHQ-9 mean scores and standard deviations for each group.

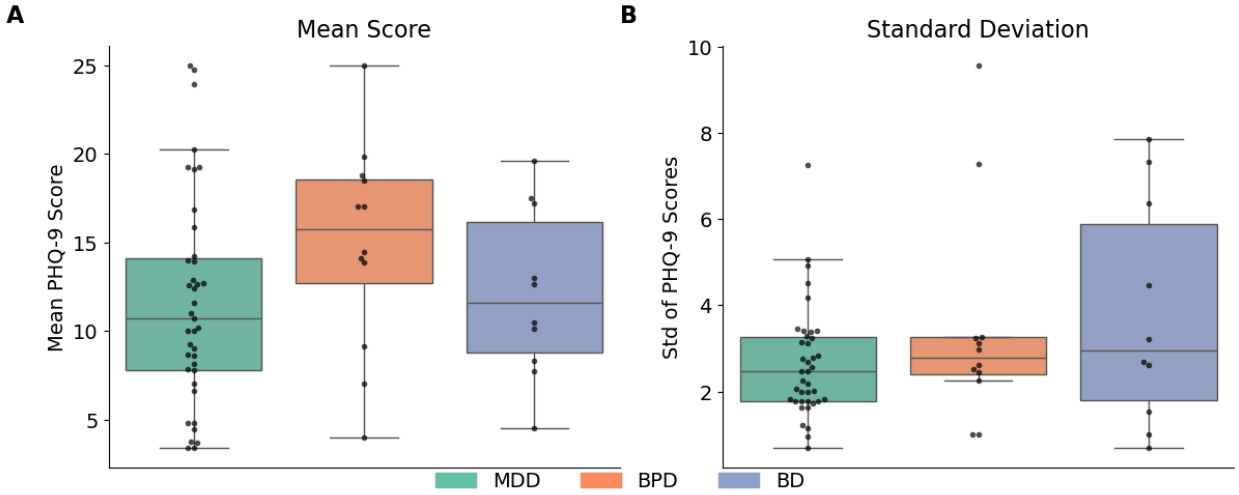

Figure 1. PHQ-9 score group-level summary statistics. The scores were first averaged within participants and then visualized. A) Boxplots of group-level mean scores: MDD (μ=11.65, σ=5.96), BPD (μ=14.89, σ=5.87), and BD (μ=12.12, σ=4.83), B) Boxplots of standard deviations of scores: MDD (μ=2.65, σ=1.28), BPD (μ=3.44, σ=2.49), and BD (μ=3.78, σ=2.61). A wider interquartile range for BD suggests greater fluctuations in depression severity among BD patients compared to others.

An exploratory comparison of the PHQ-9 item distributions at different depression severity levels exhibited differences across the patient groups. However, conducting the Kruskall-Wallis test indicated that these differences were not statistically significant. At the lowest depression level, items 3 (trouble with sleep) and 4 (tiredness or low energy) are more pronounced in the BPD group, while item 7 (trouble concentrating) 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 (suicidal

ideation) scores higher in the BD group, while item 1 (anhedonia) is pronounced in the MDD group, and the BPD group also exhibits higher scores for item 6 (low self-worth). Finally, at the highest severity level, the BD group differs from the BPD group, particularly in items 8 (psychomotor retardation) and 9 (suicidal ideation). Figure 2 illustrates these differences in distributions across groups and depression severity levels. For detailed results, refer to Appendix 1 Table S2, and Table S3 for a summary of PHQ-9 questions.

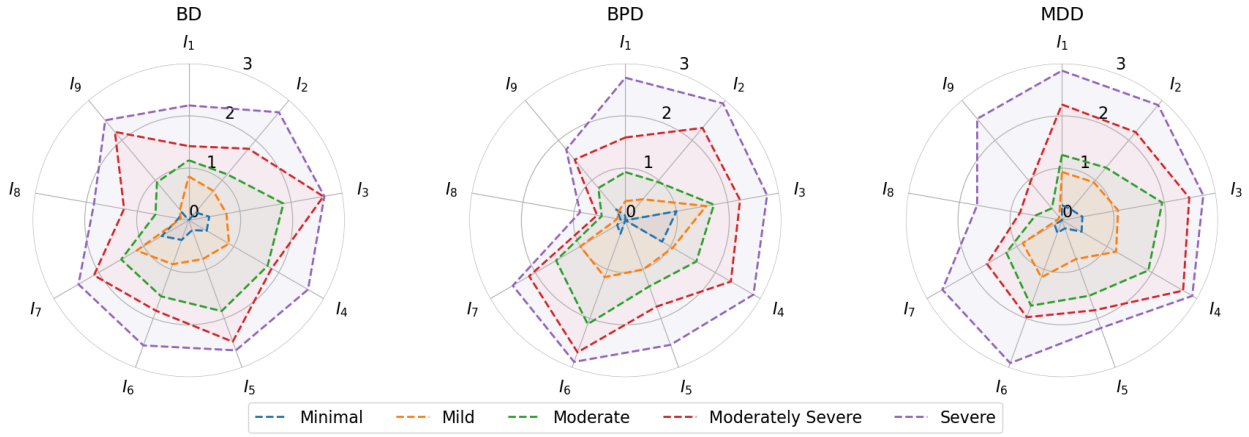

Figure 2. Average PHQ-9 item scores by diagnosis group and depression severity. For visualization purposes, we treated PHQ-9 items (ordinal) as continuous variables. The radar charts show the average of PHQ-9 item scores across different depression severity levels for BD, BPD, and MDD groups. Letters from I_1 to I_9 denote the items, and different colors represent severity levels. The scale, ranging from 0 (center) to 3 (outer circle), denotes item values.

Further exploration of the item score distributions across depression severity levels revealed that, on average, PHQ-9 items do not increase at the same rate or linearly with the total score. Items 3 (trouble with sleep) and 4 (tiredness or low energy) constantly contributed to the total score across all the severity levels. Also, item 6 (low self-worth) had high scores across the severity levels. On the other hand, item 8 (psychomotor retardation) stayed relatively low across all depression severity levels. Items 3, 4, 5 (poor appetite or overeating), and 6 showed the highest relative changes, transiting between mild depression and moderate depression. These changes suggest that these items (and corresponding symptoms) might be particularly informative for detecting depression severity above the clinical threshold (a PHQ-9 score of 10). Further, item 9 (suicidal ideation) only emerged at higher depression levels. Figure 3 illustrates the PHQ-9 item distributions across the severity levels.

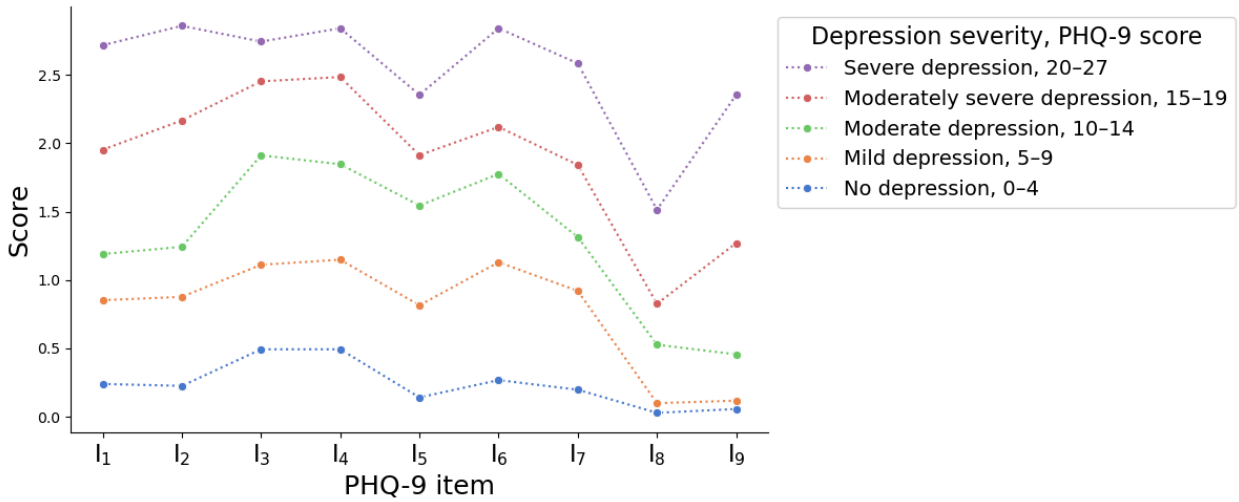

Figure 3. PHQ-9 item average score distributions across depression severity levels, combining the data from all patient groups. For exploratory analysis purposes, we treated the (ordinal) items as continuous variables and visualized the distribution shapes. The comparison shows that the items neither increase at the same rate nor in a linear manner with total depression severity. Notably, items 8 and 9 have higher values only at higher severity levels. Additionally, comparing the mild (scores from 5 to 9) and moderate (scores from 10 to 14) levels of depression severity, items 3 to 6 exhibit the highest differences.

Association analysis

To address the second objective, we examined the associations between behavioral markers and depression severity. We conducted an exploratory correlation analysis, calculating pairwise Kendall rank correlation coefficients between the behavioral features and PHQ-9 scores by pooling data from participants, and at the individual participant level, thus comparing general, study population-level associations and individual participant-level variability. The results revealed that pooling the participants, all behavioral data features had weak to moderate correlations (Kendall rank correlation coefficient between -0.3 and 0.3) with PHQ-9 depression scores. Nonetheless, these correlations were not significant after controlling for false positives with false discovery rate (FDR) correction at a significance level of α = .05. On the other hand, inspecting the correlations at the individual participant level resulted in a broader range of coefficients, summarized in Table 1.

Correlation type	Range	Pooled participants (%)	Individual participants (%)
Very strong negative	0.5 to 1.0	0.0	5.7
Strong negative	0.3 to 0.5	0.0	10.8

Moderate negative	0.1 to 0.3	29.7	17.5
No/Weak correlation	-0.1 to 0.1	53.6	23.0
Moderate positive	-0.3 to -0.1	16.7	20.6
Strong positive	-0.5 to -0.3	0.0	11.8
Very strong positive	-1 to - 0.5	0.0	10.6

Table 1. Association analysis summary, the proportion of correlation types across the behavioral features. The correlation was measured by Kendall rank correlation coefficients, calculated by pooling the participants and at the participant level. In general, participant-level coefficients exhibited broader distributions than pooled-level coefficients.

Further, we examined how the associations varied over time within the participants with exploratory rolling-window correlation analysis. We counted the proportion of users and features exhibiting both positive and negative correlations during the data collection period across different window sizes. The results indicated that correlation directions (positive or negative) might change over time. In general, the proportions of mixed correlation, reflecting that participants had both positive and negative correlations, are higher with shorter windows. These findings suggest that the associations are not static and should be considered dynamic in modeling tasks. Appendix 1 Figure S1 summarizes the average proportion of features exhibiting both positive and negative correlations across participants.

Multilevel modeling

To address the third objective, we used multilevel modeling to examine both within- and between-participant variability in behavior and to identify markers indicative of depression. We employed a stepwise modeling strategy, starting with a baseline model that included only the random intercepts for each participant and subsequently adding background information, between- and within-effects as predictors. The baseline model, including only random intercepts, explained 72.0% of the total variance. The first model, which included demographic and contextual information, increased the conditional R^2-score (accounting for both fixed and random effects) to 0.737 and the marginal R^2-score (reflecting only fixed effects) to 0.162. The second model, adding time-varying within-participant features, further increased the conditional R^2 to 0.743 and the marginal R^2-score to 0.172. Finally, the complete model combines background, contextual, and within- and between-participant behavioral features, resulting in a conditional R^2-score of 0.797 and a marginal R^2-score of 0.207.

The results from multilevel models showed that the intraclass correlation coefficient (ICC) was relatively high across models (ranging from 0.69 to 0.73), 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 (Model 2) improved model fit over the baseline model (Model 1) (χ^2 = 26.22,, df = 16, P = .05; AIC = 3613.76, BIC = 3721.02). However, incorporating between-participant features (Model 3) only led to marginal improvement over Model 2 (χ^2 = 54.78, df = 16, P < .001; AIC = 3590.9, BIC = 3769.74). Therefore, Model 2 represents the most parsimonious model.

In Model 2, employment status (β = -4.79, 95% CI: -7.65 to -1.80, P = .004) and age (β = -0.12, 95% CI: -0.25 to -0.00, P = .050) emerged as significant demographic predictors of depression severity, with unemployment and younger age linked to higher symptom levels. Among behavioral features, higher nighttime movement (β = -0.35, 95% CI: -0.63 to -0.06, P = .024) was linked to reduced symptom severity. Notably, when taking into account between-participant effect, outgoing SMS count emerge as a significant predictors. The effect is positive at between-participant level (β = 2.20, 95% CI: 1.12 to 3.32, P = <.001) but negative at within-participant level (β = -.56, 95% CI: -1.00 to -0.17, P = .004). Table 2 shows the condensed model summaries, displaying only the features with P-values below 0.05. For the detailed results, refer to Table S4 in Multimedia Appendix 1.

		Model 1			**Model 2**			**Model 3**	
	Estimate	*CI*	*p*	*Estimate*	*CI*	*p*	*Estimate*	*CI*	*p*
Age	-0.1	-0.22 – 0.03	.142	-0.12	-0.25 – -0.00	**.05**	-0.09	-0.24 – 0.05	.218
Gender (Male)	1.05	-2.28 – 4.34	.534	1.3	-2.08 – 4.48	.458	1.53	-2.15 – 5.22	.404
Children (Yes)	-1.9	-5.05 – 1.09	.23	-2.39	-5.69 – 0.95	.162	-2.60	-5.58 – 1.047	.166
Pet (Yes)	-0.67	-3.53 – 2.29	.656	-1.13	-4.14 – 1.83	.442	-1.48	-4.99 – 1.87	.400
Employed (Yes)	-4.18	-7.01 – -1.38	.004	-4.79	-7.65 – -1.80	**.004**	-5.40	-8.75 – -1.93	**.002**
Call count out W				-0.03	-0.38 – 0.35	.872	-0.16	-0.69 – 0.42	.618
SMS count out W				0.01	-0.38 – 0.38	.954	-0.56	-1.00 – -0.17	**.004**
Acc STD night W				-0.38	-0.68 – -0.07	**.012**	-0.79	-1.29 – -0.29	**.002**
SMS count out B							2.20	1.12 – 3.32	**<.001**
Random Effects									
σ2	11.85			11.15			10.89		
τ00	25.95 user			27.36 user			29.49 user		
ICC	0.69			0.71			0.73		
N	57 user			57 user			57 user		

Observations	645	645	645
Marginal R^2 / Conditional R^2	0.162 / 0.737	0.172 / 0.743	0.201 /0.797

Table 2. Multilevel model summaries. For brevity, the table shows only the statistically significant behavioral predictors. Model 2 is the best-fitting and most parsimonious, showing within-participant behavioral features: nighttime accelerometer variability with lower severity. Additionally, higher age and employed status are associated with lower depression severity.
W denotes the within-participant effect, σ2 residual variance, τ00 intercept variance, and ICC intra-class correlation coefficient.
Model 1 includes background demographics and contextual predictors, with random intercepts for each participant.
Model 2 adds within-participant effects (features) into model 1
Model 3 adds between-participant effects (features) into model 2.

Discussion

This study investigated the variability in self-reported symptoms of depression and associated smartphone-sensed behaviors. We observed variability in PHQ-9 items between the patient groups and across different levels of depression severity. Associations between behavioral markers and depression severity varied considerably between the study participants and over time within participants. The multilevel model results indicate that approximately 70% of the variability in self-assessed depression is attributable to between-participants differences in depression severity. Adding demographic background, context, and time-varying within-participant behavioral features improves the model's explanatory power by up to 76.3%. Multilevel modeling identified that demographic factors, age and employment status, and within-participant behavioral markers of daily outgoing SMS count and physical nighttime activity are associated with depression severity.

Inspection of PHQ-9 item distributions revealed differences in the representation of depression across the diagnostic subgroups. The differences suggest that accommodating individual differences in depression manifestations at the symptom level and modeling each participant individually or pooling together participants with similar symptoms might benefit the modeling accuracy. Additionally, these differences in symptom profiles 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, we observed the highest difference in PHQ-9 score distributions between mild and moderate depression on items 3 (trouble with sleep), 4 (tiredness or low energy), 5 (poor

appetite or overeating), and 6 (low self-worth). These differences suggest that the transition from lower depression into moderate level may be identifiable through monitoring behavior related to these items, e.g., sleep disturbances, daily activity and mobility, and sociability. Additionally, we noticed that item 8 (psychomotor retardation) exhibits 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 [19,24]. Similarly, item 9 (suicidal ideation) gets higher scores only at higher depression levels. This finding is consistent with other studies and is a clinically important point - the alleviation of depression results in the alleviation of suicidal thoughts [25].

Correlation analysis results indicated that most behavioral markers had weak correlations with PHQ-9 depression scores at the population level. In contrast, at the individual participant level, the correlations exhibited a wider range of both positive and negative values, which may cancel each other out when aggregated at the population level. Notably, the number of data points provided by each participant varies. Thus, some of the higher correlations might stem from the relatively low observation count by change. On the other hand, the low variability in PHQ-9 scores could explain the low correlations for some participants. Likewise, considering application usage-related markers, low correlations may occur because some participants use smartphone applications infrequently. Finally, the complementary rolling-window correlation analysis explored the within-participant variability in time, revealing periods with varying correlation strengths and directions. This observation suggests that the relationship between behavioral markers and depression severity should be considered time-dependent rather than static, aligning with existing literature that emphasizes the need for models that can accommodate the time-varying dynamics of the depression-behavior relationship [14,16].

The multilevel modeling results regarding the explanatory power of random and fixed effects align with other studies using hierarchical linear modeling (e.g., [26–28]). Adding behavioral features increased the marginal R^2 -score by 2.1%, indicating modest explanatory power of passively sensed, time-varying participant-level behavioral features. Further, the significant portion of PHQ-9 variance is explained by between-participant variability in depression severity. These findings suggest focusing on within-participant effects in behavior and accounting for participant's depression severity history in detection and future prediction.

The validity and generalizability of this study's findings are primarily limited by the amount of available data; the sample size, participants dropping out of the study, and missing intermediate observations (the dropout rate and data missingness is more thoroughly discussed in our previous work [29,30]). Limited yet high dimensional data is a common

issue across studies involving personal digital devices for data collection [31,32]. While the quantity of passively collected raw data is relatively high in this study, more bi-weekly sampled active data for each participant would have benefited the analysis, leading to higher statistical power and making modeling less prone to overfitting. Also, combining data sets across different studies may provide a feasible approach to obtaining larger and more heterogeneous samples.

Additionally, the temporal nature of the data might bias the analysis results. Used feature standardization and imputation techniques assume stationarity of variance over time, which might not hold with behavior in participants with mood disorders. These issues could be mitigated by considering time-aware standardization and thoroughly analyzing the nature of missing observations (whether they are missing completely at random, at random, or not at random) and thorough comparison between different imputation techniques, such as multiple imputation using chained equations (MICE) [33].

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 biases in depression, such as changes in speech prosody or the linguistic content of written texts (for example, in emails and social media platforms), consumed media content, or facial expressions.

Due to the limited explanatory power of linear modeling, we recommend using multiple features in future modeling, exploring feature interactions and non-linear relationships in the data with modeling techniques, such as gradient-boosting trees, and personalizing the models. Previous research has addressed the problem of between-participant variability through personalized modeling that adapts to the differences between participants. The models can be personalized, for example, by using a participant-specific (idiographic) approach (e.g., [28]), personalizing the depression symptom profiles (e.g., [34]), hierarchical modeling (e.g., [35,36]), symptom similarity-based grouping (e.g.,[37]), combining the training data from participants with similar behavioral feature distributions (e.g., [38]), or collaborative filtering modeling (e.g.,[39]).

This study exemplifies how self-reported depressive symptoms and associated mobile-sensed behaviors vary between and within patients clinically diagnosed as experiencing major depressive episodes. Our multilevel modeling revealed that a substantial portion of depression variability is due to individual differences. In summary, these findings suggest that smartphone-sensed behavioral markers may have limited clinical utility in monitoring depression using digital phenotyping within this study population. Hence, we propose that future research focus on improving depression prediction and forecasting accuracy by utilizing data from multiple sources using advanced techniques capable of modeling non-linear and time-varying relationships, feature

interactions, and differences between participants. Through these results, we emphasize the importance of addressing and reporting heterogeneity within the study population, as it may provide critical information for future research design, identifying useful behavioral markers, and study sample size optimization with power analysis.

Methods

Setting

This study analyzed behavioral data from the Mobile Monitoring of Mood (MoMo-Mood) study [29], which includes continuously sensed smartphone behavioral data (passive data), self-reported responses to biweekly depression questionnaire (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 [40], and the Structured Clinical Interview for DSM-IV Axis II Personality Disorders [41]. The study used continuous enrolling, allowing the participants to enter or withdraw from the study at different times, with a recommended participation period of one year. Information about the study, including the option to leave the study at any point, and the data collected prior to enrolment was provided to the participants, who provided written consent for the study. The participants were compensated with four movie tickets. The Helsinki and Uusimaa Hospital District's Ethics Committee approved the study research protocols (including the data stream and collection platform security), and Helsinki and Uusimaa Hospital District Psychiatry granted a research permit (approval number § 125/2018). A written data security statement was authorized by the local research ethics committee and IT support. For more details, refer to earlier work [29,30,42].

Dataset description

The passive data used in our analysis originated from various smartphone sensors and logs, collected using the AWARE app [43] and curated using the Niima platform [44]. The data encompasses five aspects of daily behavior: physical activity and mobility, sociability, phone interactions, and sleep. Behavioral features related to these aspects have been identified as being associated with depression severity in previous research (e.g., for physical activity [45–47], mobility [48,49], sociability [50,51], phone interactions [52,53], and sleep [39,54]). 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. The PHQ-9 is a tool used for screening for diagnosis and monitoring of depressive symptoms [55]. The PHQ-9 survey comprises nine items, each addressing the main symptoms of

depression, such as having little interest or pleasure in doing things. According to the survey, patients were 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 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 behaviors.

Data preparation and 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 [30]. Further, we excluded the participants who provided less than one month of passive data (equaling at least two biweekly PHQ-9 responses), resulting in a total of 62 (39%) participants from three different groups: (1) patients with BD (n=9), (2) patients with BPD (n=12), and (3) patients with MDD (n=41).

The raw passive data comprised accelerometer readings, application usage logs, battery level, GPS locations, phone call logs, SMS logs, and screen usage. We used the Niimpy behavioral analysis toolbox [56] to preprocess the raw sensor data and extract behavioral features. 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) for more precise information in daily rhythms, similar to our previous work [30]. Considering the behavioral aspects, physical activity was captured through phone accelerometer data, providing temporal information about movement intensity and variability. Mobility was assessed using GPS location data, which provided spatial (e.g., total distance traveled) and contextual information (e.g., daily time spent at home) on behavior. 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 the longest period of inactivity measured through screen activity data, following a process similar to that demonstrated in [29]. This method gives comparable results with other methodologies for measuring sleep in this population [57]. For further details, refer to Table S1 in Multimedia Appendix 1.

For active data, we used scores from individual items and assessed the depression severity with the total sum of item score, with the result ranging from 0 (no depression) to 27 (severe depression). Further, categorical background demographics and contextual variables were used as control variables. We used demographics, including age and gender, as well as contextual variables that might influence daily behavior: such as having children, owning a pet, and being employed. The variables were used in multilevel modeling, encoded using one-hot encoding.

Extracted temporal and volume-based behavioral features were aligned with biweekly depression questionnaires and aggregated by calculating the arithmetic mean over the 14 days preceding the days the PHQ-9 measurement took place. Location data features (including distance-based and significant places-based features) were calculated directly over 14-day windows, thus requiring no averaging. 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). The resulting data comprised 671 biweekly observations covering 224 behavioral features, two background demographics, five contextual variables, and one external variable. Figure 4 shows the details of 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 used aligned passive (behavioral features) and active (self-assessed PHQ-9 scores) data, removing observations where either the feature or questionnaire was missing. For the third objective, we imputed the missing intermittent passive data observations. Behaviors that do not tend to occur daily, such as sociability, were imputed with zeros. Within-participant median imputation was used for the remaining variables. Further, to examine within- and between-participant effects of behavior using multilevel modeling, we disaggregated passive data features using participant-mean and grand-mean centering [58]. The within-participant components reflect fluctuations relative to an individual's average and their relationship to the severity of depression. Contrarily, between-participant 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). For within-participant centering, we centered each variable around the participant's mean to account for variability over time. Second, for between-participant centering, the variables were averaged within participants and then centered across the study population's mean. This centering approach enables the detection of both individual-level behavioral fluctuations and broader population-level patterns.

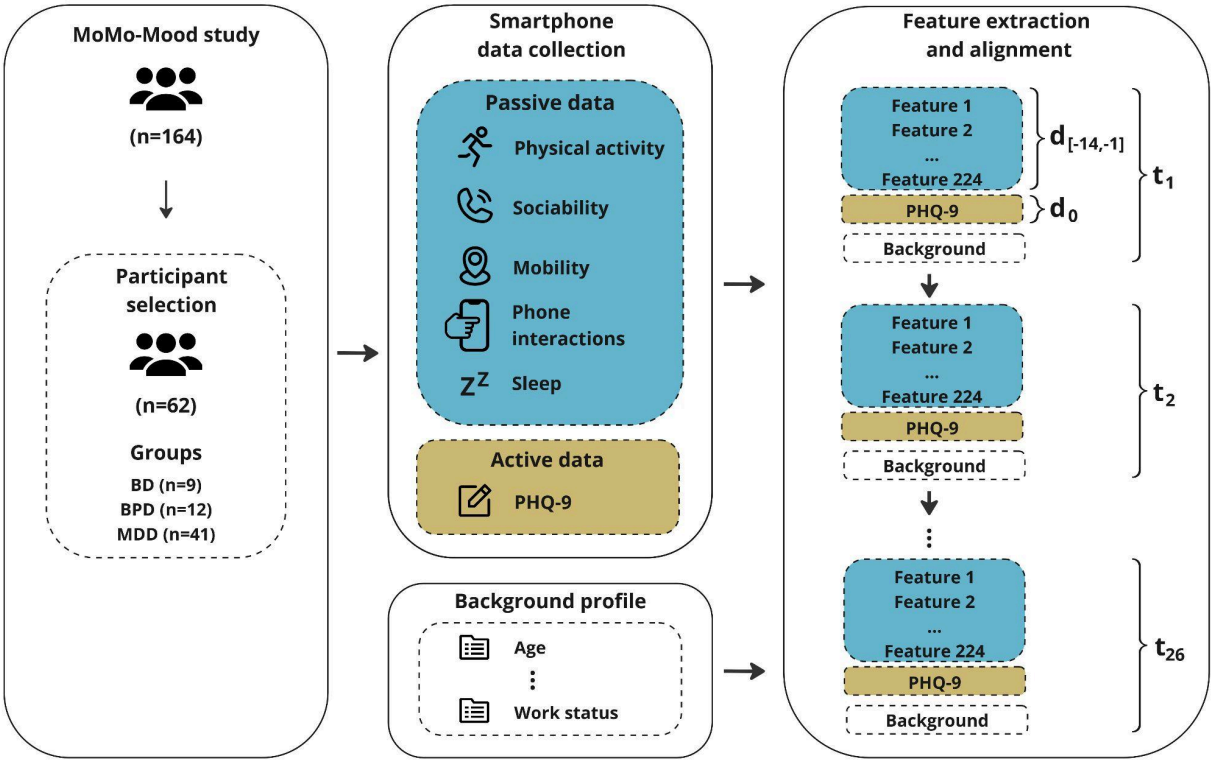

Figure 4. Participant selection, data collection, and processing schema. For this study, we selected participants from three patient groups (BD, BPD, and MDD) who answered at least two bi-weekly (14-day) PHQ-9 questionnaires and provided passive data for 14 days prior to each PHQ-9 answer. Smartphone-collected data encompassed: 1) passive data, consisting of raw smartphone-sensed data, including accelerometer readings, phone call and SMS logs, app and screen usage logs, battery level logs, and GPS location data; 2) active data, consisting of PHQ-9 questionnaires; and 3) demographic information, such as participants' age and work status. The merged datapoints are denoted by t_1 to t_{26}, and day d_0 represents the day the PHQ-9 was reported, and d_{-14} to d_{-1} the preceding 14-day period passive data was collected. Data collection lasted up to one year.

PHQ-9 questionnaire analysis

To assess the heterogeneity in self-reported depression, we conducted a descriptive analysis of PHQ-9 scores, applied the Kruskal-Wallis test [23] to evaluate group-wise differences in score means and standard deviations, and performed an exploratory analysis for each item across the groups and different depression severity levels. Additionally, we utilized exploratory data analysis to examine the scores across the groups and severity levels

Association analysis

To investigate the associations between behavioral features and self-reported depression among participants, we conducted exploratory correlation analysis, calculating pair-wise

Kendall rank correlation coefficients between the behavioral features and PHQ-9 scores. We selected Kendall rank correlation since it does not require the compared variables to be normally distributed or linearly related, and it can be used with ordinal data. For analysis, we used a function implemented in the *SciPy* library [59]. We employed two approaches: pooling data from all participants and calculating the coefficients for each participant individually. Due to the multiple comparisons, we applied the FDR correction with the Benjamini-Hochberg procedure [60] to control Type I errors. Additionally, we examined whether the associations between behavioral features and depression symptoms varied within participants over time by conducting an exploratory rolling-window correlation analysis. We tested various window sizes, ranging from 28 days to one year. Within each window, we computed the Kendall rank correlation coefficient, shifting the window forward by 14 days to calculate successive correlations. We excluded participants who did not have enough data to fill at least one complete window for a given window size (and subsequent longer window sizes) from the analysis. For each window size, we used 1,000 bootstrapping samples to generate more robust distributions of correlation coefficients.

Multilevel modeling

To assess the within- and between-participant effects of behavioral features on PHQ-9 scores, we used multilevel modeling with random intercepts to account for the data's hierarchical structure. Independent variables consisted of features representing behavioral aspects, including sleep, activity, mobility, sociability, and device usage. PHQ-9 scores were used as the dependent variable. Participants were used as grouping variables. Generalized variance inflation factor (GVIF) was used to assess multicollinearity, keeping only variables with GVIF < 10. We employed a stepwise modeling strategy. First, we started with a baseline model that included only random intercepts for each participant. Second, we included demographic and contextual variables. Next, within-participant variables were added to the model. Finally, between-participant variables were included. The *lme4* package [61] in R was used for fitting the models, and 95% confidence intervals were estimated using bootstrapping (1000 iterations) with *the sjPlot* package [62]. For model comparison, we used the likelihood ratio tests to assess explanatory power and ensure model parsimony. The model formulas and the final list of predictors are shown in Appendix 3.

Data availability

Due to the sensitive nature of patient data it cannot be shared publicly.

Code availability

The code used for preprocessing the data is available on GitHub [63].

Acknowledgments

The computational resources provided by the Aalto Science-IT project are gracefully 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

AIC: Akaike information criterion

ANOVA: Analysis of Variance

BD: Bipolar Disorder

BH: Benjamini-Hochberg

BIC: Bayesian information criterion

BPD: Borderline Personality Disorder

CI: Confidence Interval

FDR: False Discovery Rate

GPS: Global Positioning System

GVIF: Generalized Variance Inflation Factor

HUS: Helsinki and Uusimaa Hospital District

ICC: Intraclass Correlation Coefficient

MDD: Major Depressive Disorder

MICE: Multivariate Imputation by Chained Equations

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. Huckvale K, Venkatesh S, Christensen H. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. *Npj Digit Med*. 2019;2(1):1-11. doi:10.1038/s41746-019-0166-1

8. 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

9. 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

10. 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

11. 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

12. Lee K, Lee TC, Yefimova M, et al. Using digital phenotyping to understand health-related outcomes: A scoping review. *Int J Med Inf*. 2023;174:105061. doi:10.1016/j.ijmedinf.2023.105061

13. Fried EI, Nesse RM. Depression is not a consistent syndrome: An investigation of unique symptom patterns in the STAR*D study. *J Affect Disord*. 2015;172:96-102. doi:10.1016/j.jad.2014.10.010

14. Brietzke E, Hawken ER, Idzikowski M, Pong J, Kennedy SH, Soares CN. Integrating digital phenotyping in clinical characterization of individuals with mood disorders. *Neurosci Biobehav Rev*. 2019;104:223-230. doi:10.1016/j.neubiorev.2019.07.009

15. Price GD, Heinz MV, Song SH, Nemesure MD, Jacobson NC. Using digital phenotyping to capture depression symptom variability: detecting naturalistic variability in depression symptoms across one year using passively collected wearable movement and sleep data. *Transl Psychiatry*. 2023;13(1):1-10. doi:10.1038/s41398-023-02669-y

16. Lekkas D, Gyorda JA, Price GD, Jacobson NC. Depression deconstructed: Wearables and passive digital phenotyping for analyzing individual symptoms. *Behav Res Ther*. 2023;168:104382. doi:10.1016/j.brat.2023.104382

17. Bowen R, Peters E, Marwaha S, Baetz M, Balbuena L. Moods in Clinical Depression Are More Unstable than Severe Normal Sadness. *Front Psychiatry*. 2017;8. doi:10.3389/fpsyt.2017.00056

18. Kleiman EM, Turner BJ, Fedor S, Beale EE, Huffman JC, Nock MK. Examination of real-time fluctuations in suicidal ideation and its risk factors: Results from two ecological momentary assessment studies. *J Abnorm Psychol*. 2017;126(6):726-738. doi:10.1037/abn0000273

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

20. van Eeden WA, van Hemert AM, Carlier IVE, Penninx BW, Giltay EJ. Severity, course trajectory, and within-person variability of individual symptoms in patients with major depressive disorder. *Acta Psychiatr Scand*. 2019;139(2):194-205. doi:10.1111/acps.12987

21. Musliner KL, Munk-Olsen T, Eaton WW, Zandi PP. Heterogeneity in long-term trajectories of depressive symptoms: Patterns, predictors and outcomes. *J Affect Disord*. 2016;192:199-211. doi:10.1016/j.jad.2015.12.030

22. 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

23. 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

24. 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

25. 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

26. 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

27. 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

28. Balliu B, Douglas C, Seok D, et al. Personalized mood prediction from patterns of behavior collected with smartphones. *Npj Digit Med*. 2024;7(1):1-14. doi:10.1038/s41746-024-01035-6

29. 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

30. 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

31. Berisha V, Krantsevich C, Hahn PR, et al. Digital medicine and the curse of dimensionality. *Npj Digit Med*. 2021;4(1):1-8. doi:10.1038/s41746-021-00521-5

32. Mohr DC, Zhang M, Schueller SM. Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning. *Annu Rev Clin Psychol*. 2017;13(Volume 13, 2017):23-47. doi:10.1146/annurev-clinpsy-032816-044949

33. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med*. 2011;30(4):377-399. doi:10.1002/sim.4067

34. 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

35. Busk J, Faurholt-Jepsen M, Frost M, Bardram JE, Kessing LV, Winther O. Forecasting Mood in Bipolar Disorder From Smartphone Self-assessments: Hierarchical Bayesian Approach. *JMIR MHealth UHealth*. 2020;8(4). doi:10.2196/15028

36. Zhang Y, Folarin AA, Sun S, et al. Predicting Depressive Symptom Severity Through Individuals' Nearby Bluetooth Device Count Data Collected by Mobile Phones: Preliminary Longitudinal Study. *JMIR MHealth UHealth*. 2021;9(7):e29840. doi:10.2196/29840

37. Kathan A, Harrer M, Küster L, et al. Personalised depression forecasting using mobile sensor data and ecological momentary assessment. *Front Digit Health*. 2022;4. doi:10.3389/fdgth.2022.964582

38. Adler DA, Wang F, Mohr DC, Choudhury T. Machine learning for passive mental health symptom prediction: Generalization across different longitudinal mobile sensing studies. *PLOS ONE*. 2022;17(4):e0266516. doi:10.1371/journal.pone.0266516

39. Xu X, Chikersal P, Dutcher JM, et al. Leveraging Collaborative-Filtering for Personalized Behavior Modeling: A Case Study of Depression Detection among College Students. *Proc*

ACM Interact Mob Wearable Ubiquitous Technol. 2021;5(1):41:1-41:27. doi:10.1145/3448107

40. 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*.

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

42. 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

43. Ferreira D, Kostakos V, Dey AK. AWARE: Mobile Context Instrumentation Framework. *Front ICT*. 2015;2. Accessed March 8, 2022. https://www.frontiersin.org/article/10.3389/fict.2015.00006

44. Aledavood T, Hoyos AMT, Alakörkkö T, et al. Data collection for mental health studies through digital platforms: requirements and design of a prototype. *JMIR Res Protoc*. 2017;6(6):e110.

45. 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

46. Pedrelli P, Fedor S, Ghandeharioun A, et al. Monitoring Changes in Depression Severity Using Wearable and Mobile Sensors. *Front Psychiatry*. 2020;11. Accessed August 21, 2023. https://www.frontiersin.org/articles/10.3389/fpsyt.2020.584711

47. Chikersal P, Doryab A, Tumminia M, et al. Detecting Depression and Predicting its Onset Using Longitudinal Symptoms Captured by Passive Sensing: A Machine Learning Approach With Robust Feature Selection. *ACM Trans Comput-Hum Interact*. 2021;28(1):3:1-3:41. doi:10.1145/3422821

48. Saeb S, Lattie EG, Schueller SM, Kording KP, Mohr DC. The relationship between mobile phone location sensor data and depressive symptom severity. *PeerJ*. 2016;4:e2537.

49. Laiou P, Kaliukhovich DA, Folarin AA, et al. The Association Between Home Stay and Symptom Severity in Major Depressive Disorder: Preliminary Findings From a Multicenter Observational Study Using Geolocation Data From Smartphones. *JMIR MHealth UHealth*. 2022;10(1):e28095. doi:10.2196/28095

50. 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

51. Sverdlov O, Curcic J, Hannesdottir K, et al. A Study of Novel Exploratory Tools, Digital Technologies, and Central Nervous System Biomarkers to Characterize Unipolar

Depression. *Front Psychiatry*. 2021;12. doi:10.3389/fpsyt.2021.640741

52. Sun S, Folarin AA, Zhang Y, et al. Challenges in Using mHealth Data From Smartphones and Wearable Devices to Predict Depression Symptom Severity: Retrospective Analysis. *J Med Internet Res*. 2023;25:e45233. doi:10.2196/45233

53. Zou B, Zhang X, Xiao L, et al. Sequence Modeling of Passive Sensing Data for Treatment Response Prediction in Major Depressive Disorder. *IEEE Trans Neural Syst Rehabil Eng Publ IEEE Eng Med Biol Soc*. 2023;31:1786-1795. doi:10.1109/TNSRE.2023.3260301

54. Bai R, Xiao L, Guo Y, et al. Tracking and Monitoring Mood Stability of Patients With Major Depressive Disorder by Machine Learning Models Using Passive Digital Data: Prospective Naturalistic Multicenter Study. *JMIR MHealth UHealth*. 2021;9(3):e24365. doi:10.2196/24365

55. 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

56. 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

57. Mahir A, Luong N, Baryshnikov I, Martikkala A, Isometsä E, Aledavood T. Multi-Modal Sleep Measurement and Alignment Analysis in Outpatients with Major Depressive Episode. Published online April 30, 2025:2025.04.29.25326308. doi:10.1101/2025.04.29.25326308

58. 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

59. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*. 2020;17(3):261-272. doi:10.1038/s41592-019-0686-2

60. 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

61. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. *J Stat Softw*. 2015;67:1-48. doi:10.18637/jss.v067.i01

62. Lüdecke D, Bartel A, Schwemmer C, Powell C, Djalovski A, Titz J. sjPlot: Data Visualization for Statistics in Social Science. Published online November 29, 2024. Accessed June 11, 2025. https://cran.r-project.org/web/packages/sjPlot/index.html

63. DigiTraces Lab. Niimpy. Published online 2025. https://github.com/digitraceslab/niimpy

Multimedia Appendix 1

Depression severity	No	Mild	Moderate	Moderately severe	Severe	**Total**
Group						
BP	5 (5.7%)	19 (21.6%)	22 (25.0%)	21 (23.9%)	21 (23.9%)	88
BPD	11 (9.9%)	19 (17.1%)	26 (23.4%)	29 (26.1%)	26 (23.4%)	111
MDD	55 (11.5%)	127 (26.6%)	138 (28.9%)	87 (18.2%)	70 (14.7%)	477
Total	71 (10.5%)	165 (24.4%)	186 (27.5%)	137 (20.3%)	117 (17.3%)	676

Table S1. Depression severity levels by groups. The table shows the distribution of depression severity levels based on repeated measurements of PHQ-9, meaning that each user may contribute to multiple severity levels. The table shows that moderate depression severity (PHQ-9 scores from 10 to 15) is most frequent across the groups, and moderately severe and severe levels are more frequent among BD and BPD patients compared to MDD patients.

	PHQ-9 item	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***
Depression severity	***Group***									
Minimal	***bd***	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*
	bpd	*0*	*0*	*1*	*1*	*0*	*0*	*0*	*0*	*0*
	mdd	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*
Mild	***bd***	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*0*	*0*
	bpd	*0*	*1*	*1*	*1*	*1*	*1*	*1*	*0*	*0*
	mdd	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*0*	*0*
Moderate	***bd***	*1*	*1*	*2*	*2*	*2*	*2*	*2*	*0*	*1*
	bpd	*1*	*1*	*2*	*1*	*1*	*2*	*2*	*0*	*1*

	mdd	*1*	*1*	*2*	*2*	*1*	*2*	*1*	*0*	*0*
Moderately Severe	***bd***	*1*	*2*	*3*	*2*	*3*	*2*	*2*	*1*	*2*
	bpd	*2*	*2*	*2*	*3*	*2*	*3*	*2*	*0*	*1*
	mdd	*2*	*2*	*3*	*3*	*2*	*3*	*2*	*1*	*1*
Severe	***bd***	*3*	*3*	*3*	*3*	*3*	*3*	*2*	*2*	*3*
	bpd	*3*	*3*	*3*	*3*	*3*	*3*	*3*	*1*	*2*
	mdd	*3*	*3*	*3*	*3*	*2*	*3*	*3*	*1*	*3*

Table S2. Figure 2 Median PHQ-9 item scores by diagnosis group and depression severity. The item scores range from 0 to 3.

PHQ-9 item no.	**Description**
1.	Little interest or pleasure in doing things
2.	Feeling down, depressed, or hopeless
3.	Trouble falling or staying asleep, or sleeping too much
4.	Feeling tired or having little energy
5.	Poor appetite or overeating
6.	Feeling bad about yourself — or that you are a failure or have let yourself or your family down
7.	Trouble concentrating on things, such as reading the newspaper or watching television
8.	Moving or speaking so slowly that other people could have noticed? Or the opposite — being so fidgety or restless that you have been moving around a lot more than usual
9.	Thoughts that you would be better off dead or of hurting yourself in some way

Table S3: Nine-item patient health questionnaire (PHQ-9) item descriptions, as introduced by Kroenke and Spitzer (2002).

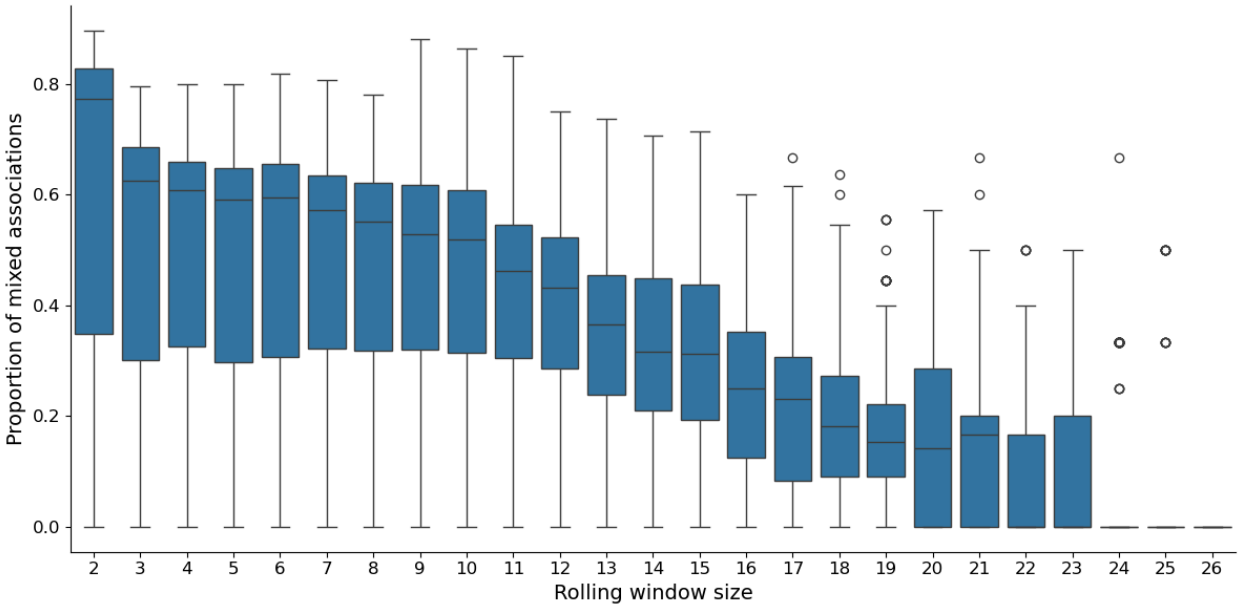

Figure S1. Exploratory rolling window correlation analysis summary, showing the proportion of behavioral features showing both positive and negative Kendall's Tau rank correlation coefficients across the users. In general, shorter window yields higher proportion of mixed correlations. Window size of 2 (equaling 28 days) result in 67% mixed correlations on average.

		Model 1			**Model 2**			**Model 3**	
	Estimate	*CI*	*p*	*Estimate*	*CI*	*p*	*Estimate*	*CI*	*p*
Age	-0.1	-0.22 – 0.03	.142	-0.12	-0.25 – -0.00	**.05**	-0.09	-0.24 – 0.05	.218
Gender (Male)	1.05	-2.28 – 4.34	.534	1.3	-2.08 – 4.48	.458	1.53	-2.15 – 5.22	.404
Children (Yes)	-1.9	-5.05 – 1.09	.23	-2.39	-5.69 – 0.95	.162	-2.60	-5.58 – 1.047	.166
Pet (Yes)	-0.67	-3.53 – 2.29	.656	-1.13	-4.14 – 1.83	.442	-1.48	-4.99 – 1.87	.400
Employed (Yes)	-4.18	-7.01 – -1.38	.004	-4.79	-7.65 – -1.80	**.004**	-5.40	-8.75 – -1.93	**.002**
Call count out W				-0.03	-0.38 – 0.35	.872	-0.16	-0.69 – 0.42	.618
Call count in W				0.11	-0.29 – 0.52	.596	-0.08	-0.58 – 0.40	.740
SMS count in W				0.12	-0.28 – 0.49	.51	0.21	-0.28 – 0.74	.386
SMS count out W				0.01	-0.38 – 0.38	.954	-0.56	-1.00 – -0.17	**.004**
Call duration in W				0.21	-0.20 – 0.59	.322	0.13	-0.29 – 0.58	.548
Call duration out W				0.23	-0.11 – 0.57	.182	-0.19	-0.66 – 0.27	.42
Acc STD morning W				-0.08	-0.40 – 0.24	.59	0.00	-0.51 – 0.15	.988
Acc STD afternoon W				0.05	-0.27 – 0.36	.786	-0.11	-0.74 – 0.46	.686

Acc STD evening W				-0.13	-0.45 – 0.17	0.452	-0.11	-0.87 – 0.327	.246
Acc STD night W				-0.38	-0.68 – -0.07	**.012**	-0.79	-1.29 – -0.29	**.002**
Screen time W				-0.19	-0.50 – 0.12	.248	-0.39	-0.92 – 0.13	.132
Battery level variance W				0.3	-0.02 – 0.63	.062	0.27	-0.20 – 0.73	.272
Time at home W				-0.12	-0.42 – 0.17	.42	0.41	-0.19 – 0.96	.184
Location entropy W				0	-0.33 – 0.29	.984	-0.16	-0.77 – 0.39	.552
Sleep duration W				0.13	-0.17 – 0.45	.386	-0.46	-1.05 – 0.14	.132
Midsleep W				0.09	-0.22 – 0.39	.53	-0.29	-0.77 – 0.12	.184
Call count out B							0.28	-0.94 – 1.36	.660
Call count in B							0.68	-0.39 – 1.76	.238
SMS count in B							-1.08	-2.38 – 0.18	.082
SMS count out B							2.20	1.12 – 3.32	**<.001**
Call duration in B							-0.17	-0.90 – 0.48	.648
Call duration out B							0.88	-0.22 – 2.04	.126
Screen time B							0.5	-0.52 – 1.58	.462
Battery level variance B							-0.12	-0.91 – 0.64	.732
Acc STD morning B							-0.35	-1.29 – 0.62	.482
Acc STD afternoon B							0.38	-0.63 – 1.38	.472
Acc STD evening B							0.29	-0.93 – 1.40	.580
Acc STD night B							0.95	-0.11 – 1.93	.078
Time at home B							-0.77	-1.56 – 0.10	.084
Location entropy B							0.33	-0.64 – 1.31	.490
Sleep duration B							1.04	-0.05 – 2.18	**.0484**
Midsleep B							0.62	-0.03 – 1.24	.064
Random Effects									
σ2	11.85			11.15			10.89		
τ00	25.95 user			27.36 user			29.49 user		
ICC	0.69			0.71			0.73		
N	57 user			57 user			57 user		
Observations	645			645			645		
Marginal R^2 / Conditional R^2	0.162 / 0.737			0.172 / 0.743			0.201 /0.797		

Table S4. Multilevel model summaries. Model 2 is the best fitting and parsimonious, showing within-participant behavioral features: (1) daily outgoing SMS count associated higher depression severity, and (2) nighttime accelerometer variability with lower severity.

Additionally, higher age and employed status is associated with lower depression severity. Notably, between-participant effects did not show any statistically significant associations. W denotes within-participant effect, B between-participant effect, σ2 residual variance, τ00 intercept variance, and ICC intra class correlation coefficient.
Model 1 includes background demographics and contextual predictors, with random intercepts for each participant.
Model 2 adds within-particiant effects (features) into model 1
Model 3 adds between-participant effects (features) into model 2.

Multimedia Appendix 2

Aspect	**Sensor**	**Raw features**
Physical activity	Accelerometer	Acceleration magnitude, magnitude max, min, mean and standard deviation, measurement count
Mobility	GPS location	*Distance-based:* total distance, variance, log variance, average speed, speed variance, max speed, location bin count. *Significant place-related:* static point count, moving point count, static bin count, max distance from home, number of significant places, number of rarely visited places, number of transitions between significant places, bin count in top1, top2, top3, top4, and top5 cluster, normalized entropy
Sociability	Phone call and SMS logs	*Phone call-related:* Total call duration, call duration mean, median and standard deviation, call count, type (incoming and outgoing) *SMS-related*: SMS count, type (incoming and outgoing)
Phone interactions	Application and screen usage, battery charge level	*Application-related:* Application class, count, duration *Screen usage-related:* screen-off timestamp, screen event count, screen event duration, screen event duration max, mean, min, and standard deviation, screen first unlock timestamp *Battery level related:* Battery level mean, median and standard deviation, battery shutdown time, battery discharge
Sleep	Inferred from screen	Average sleep duration, sleep start time, sleep end

	usage data	time, midsleep time

Table S1: Summary of extracted features across different behavioral aspects, including physical activity, mobility, sociability, phone interactions, and sleep. Features are derived from smartphone sensor data and logs; accelerometer, GPS, phone call and SMS logs, application usage, and battery levels. Sleep features were inferred from screen usage data.

Multimedia Appendix 3

Variables for the multilevel model were selected to capture aspects of behaviors based on prior evidence linking behavioral markers—such as sleep, activity rhythms, communication patterns, mobility, and device usage—with depressive symptoms. Sleep: duration and midsleep point. Temporal activity rhythm: standard deviation of magnitude from the accelerometer for four segments of days. Communication: count of ongoing and incoming sms/calls, duration of outgoing and incoming sms/calls. Location: time at home and entropy of location. Device usage: screen usage duration, battery mean level. The model specification is presented below.

```
PHQ9-score ~ Age + Gender + Children + Pet + Employed +
OutgoingCallsCount.between + OutgoingCallsCount.within +
IncomingCallsCount.between + IncomingCallsCount.within +
OutgoingSMSCount.between + OutgoingSMSCount.within +
IncomingSMSCount.between + IncomingSMSCount.within +
OutgoingCallDuration.between + OutgoingCallDuration.within +
IncomingCallDuration.between + IncomingCallDuration.within +
ScreenTime.between + ScreenTime.within +
BatteryVar.between + BatteryVar.within +
AccelMorning.between + AccelMorning.within +
AccelAfternoon.between + AccelAfternoon.within +
AccelEvening.between + AccelEvening.within +
AccelNight.between + AccelNight.within +
TimeAtHome.between + TimeAtHome.within +
LocationEntropy.between + LocationEntropy.within +
SleepDuration.between + SleepDuration.within +
Midsleep.between + Midsleep.within +
(1|pid)
```