

Can 1-Year Smartphone and Wearable Data allow Individual-level Modeling of Depressive Symptom Severity?

Ulrich Hegerl

`ulrich.hegerl@deutsche-depressionshilfe.de`

Goethe University Frankfurt

Simon Schreynemackers

German Foundation for Depression and Suicide Prevention

Johannes Leimhofer

Goethe University Frankfurt

Milica Petrovic

German Foundation for Depression and Suicide Prevention

Sascha Ludwig

Leipzig University

Andreas Dominik

University of Applied Sciences

Dominik Heider

University of Münster

Hanna Reich

German Foundation for Depression and Suicide Prevention

the MONDY Consortium

adesso SE

Research Article

Keywords: digital phenotyping, mobile health (mHealth), mobile sensing, depression, machine learning, smartphones, wearable devices, passive monitoring

Posted Date: April 2nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9083929/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Objective markers of depression severity in people with Major Depression (MD) would be highly relevant for monitoring disease trajectories and treatment effects. Passive sensing via mobile devices offers a means to capture continuous behavioral and physiological data patterns related to depression. Shortcomings of most prior studies include short observation periods, infrequent symptom assessments, limited data sources, or group-level analyses that obscure within-person variation. Therefore, this study tested whether individual-level models built from 1-year daily mobile sensing and self-report data can provide markers of daily self-rated depression severity (Patient Health Questionnaire-2, PHQ-2) in patients with MD, and whether non-linear ensemble models outperform linear models in detecting such markers.

Methods

Daily PHQ-2 ratings over up to one year, together with continuous smartphone and smartwatch data capturing daily physical activity, sleep–wake patterns, heart-rate, phone-based social interaction, app usage, and brief prompted speech samples, were collected from 15 adults with recurrent MD. Four patients were excluded due to excessive missing PHQ-2 scores. Linear elastic net (EN) regression was applied on the remaining 11 patients to estimate PHQ-2 scores from the mobile sensing data. For each individual, the first 70% of the time series was used to train the model, and the later 30% was kept separate and used only for testing. Model performance on the unseen test period was evaluated using predefined criteria ($R^2 > 0.30$ or mean absolute error (MAE) < 1). In addition, non-linear random forest models were tested to determine whether they increased the number of participants for whom PHQ-2 severity could be estimated using the same performance criteria.

Results

Using individual-level elastic net models, none of the 11 participants met the predefined performance criteria on unseen test data. Non-linear random forest models identified one participant (9.1%) meeting the performance threshold based on MAE. No participant reached the R^2 -based performance threshold under either modeling method.

Conclusions

1-year mobile sensing does not allow the identification of valid indicators of depression severity to support MD patients with self-management of depression. A 1-year observation period might not provide enough variance in depression severity to identify such indicators at the individual level.

Trial registration:

German Clinical Trials Register (DRKS), DRKS00032618. Registered 12 September 2023.

Background

Major depression (MD) is prevalent and associated with recurrent depressive episodes leading to substantial disability and functional impairment [1–3]. Although effective treatments such as psychotherapy and pharmacotherapy are available [4], a substantial proportion of individuals with depression remain untreated [5–7]. This treatment gap largely stems from the fact that access to specialized care by psychotherapists and psychiatrists constitutes a significant bottleneck even in countries with developed health care systems [8].

In this context, supporting and facilitating self-management is an important strategy to improve the situation for people suffering from depression. Evidence across chronic illnesses indicates that self-management interventions reduce depressive and anxiety symptoms, in part through improvements in self-efficacy and adherence to ongoing care [9, 10].

Objective markers of depression severity would facilitate self-management, considering self-ratings were found to have a low sensitivity to symptom change and lower sensitivity than clinician-administered ratings, potentially underestimating improvement or early deterioration [11, 12]. Objective continuously collected behavioral and physiological measures could provide objective markers of depression severity and strengthen self-management.

Multimodal Digital Data Sources for Depression

Smartphone and wearable devices enable the passive collection of data on activity, sleep–wake patterns, autonomic functioning, communication behavior, and speech characteristics. Prior research has shown that several of these signals are associated with depressive symptom severity. For example:

- Communication-based markers such as reductions in call frequency, shorter call duration, and lower contact diversity have repeatedly been linked to higher depressive symptom burden [13–16]. However, these studies typically used monitoring periods of days to weeks, applied group-level models, and despite being multimodal in some cases, did not combine high-frequency smartphone and wearable data over long durations.
- Similarly, speech-based acoustic features such as prosodic variability, perturbation indices, and spectral descriptors have been associated with psychomotor slowing and suicidal ideation [17–20]. These studies predominantly analyzed speech as a single data modality, used cross-sectional or short-term recordings, and relied on group-level classification approaches.
- Wearable-derived activity features including circadian rhythm regularity and activity fragmentation have also been associated with depressive symptoms [21–25]. These datasets, however, commonly involve short or medium-length actigraphy recordings, rarely exceeding several weeks, and the

resulting analyses are performed at the group level. App-usage studies similarly rely on short-term smartphone monitoring and do not integrate multimodal physiological data [26, 27].

- Heart-rate variability (HRV), measured via photoplethysmography or ECG, has been consistently associated with depression severity and treatment response [28–30]. Yet these findings are derived largely from short-duration HRV recordings and cross-sectional group comparisons, not from continuous longitudinal monitoring.
- Finally, several studies have reported that behavioral markers from communication logs, activity patterns, speech features, or circadian metrics may provide early indicators of symptom worsening [31–33]. Taken together, these investigations also relied on weeks- to a few months-long datasets, used group-level machine-learning models, and did not incorporate daily symptom assessments or n-of-1 analytic approaches.

Taken together, while multiple sensing domains show consistent associations with depressive symptoms, the existing evidence is predominantly based on short-term, group-level, and single- or limited-source analyses. To our knowledge, no prior work has combined multimodal smartphone and wearable sensing with daily symptom ratings over a 1-year period using individualized modeling frameworks.

Considerations for Identifying Depression Severity in Individual Patients

For the identification of data patterns linked to depression severity the following aspects are relevant:

- Depression exhibits substantial heterogeneity, with individuals differing in the prominence of psychomotor retardation, agitation, circadian disruption, or social withdrawal [34]. Group-level analyses often obscure these differences, and important person-specific symptom patterns may be lost when data are averaged across participants. Therefore, longitudinal, idiographic approaches are needed to detect data patterns indicating depression severity for a certain patient [35].
- To identify links between person-specific biosensors and other data to depression severity, long observation periods with fluctuations in depression severity are required [36]. Only if there is sufficient variance in depression severity, individual data patterns can be identified that show covariation with depression severity.
- Combining several data sources such as physical activity, speech, app usage, smartphone interaction patterns, and physiological signals is likely to increase the ability to detect reliable indicators of depression severity. Prior research suggests that no single passive sensing modality reliably captures depression severity on its own, and that integrating multiple data sources tends to improve identification of symptom-related patterns [24, 31, 32, 36].
- In a recent systematic review, Amin and colleagues [36] identified very few studies that collected passive sensing data for longer than six months. Most available work relied on monitoring periods of only days to several weeks, used infrequent symptom assessments, and applied group-level analytic frameworks that do not allow individualized inference. Although a small number of year-

long investigations have since demonstrated the feasibility of extended monitoring, these studies typically involved non-clinical or mixed samples, biweekly or monthly symptom ratings, or group-level analyses that do not resolve person-specific associations [37–39]. As such, the ability of daily passive sensing features to reflect day-to-day depressive symptom severity in clinically diagnosed individuals over long durations remains poorly understood.

Missing data in smartphones or wearable data is a problem often encountered in such long-term studies. They often arise from behavioral changes rather than random technical failures. Missingness not at random (MNAR) may itself reflect clinical phenomena such as withdrawal or reduced engagement. Prior studies have shown that disengagement from digital monitoring predicts symptom severity or relapse risk [40, 41].

Aim of the Present Study

The present study aims to test the hypothesis that individual-level models built from 1-year daily mobile sensing and self-report data provide markers of daily self-rated depression severity (Patient Health Questionnaire-2, PHQ-2) in patients with major depression, and that non-linear ensemble models outperform linear models in detecting these markers.

Research question 1 (RQ1): Can individual-level linear models built from longitudinal daily mobile sensing and self-report data collected over up to one year provide markers of daily self-rated depression severity (Patient Health Questionnaire-2, PHQ-2) in at least a subset of patients with major depression (MD)?

Research question 2 (RQ2): Do non-linear ensemble models, compared with individual-level linear models, improve the identification of reliable depression severity indicators and increase the number of patients in whom daily PHQ-2 symptom severity can be estimated with sufficient accuracy on previously unseen data according to predefined performance criteria?

Methods

Study Design

Observational data were prospectively collected as part of the MONDY trial (German Clinical Trials Register DRKS00032618, registered on September 12, 2023 [42]). The work was carried out in accordance with the World Medical Association Declaration of Helsinki [43].

Ethical approval

has been granted by the Ethics Committee of the Department of Medicine at Goethe University Frankfurt, Germany (date: 24/08/2023, reference number: 2023–1309).

Participants

15 adults (age > 18) with recurrent major depression participated in the study and self-monitored their symptoms for up to one year using self-report questionnaires and smartphone and smartwatch data. Table 1 summarizes participant characteristics by sex (sample size and age).

Table 1
Participant characteristics by sex
(sample size and age).

Participants	N	Age mean ± SD
Male	6	41.67 ± 10.43
Female	9	35.67 ± 8.64

Participants were recruited within clinic premises where a study promotional flyer was presented on the information board and via social media channel. The recruitment took place in person, and all participants signed the consent form on site. The baseline data collected included basic sociodemographic information, PHQ2 and PHQ9. Once the baseline has been established, the team member provided support in installing the iTrackDepression smartphone application concluding the inclusion in the study.

Data Sources and Feature Extraction

Self-ratings

Participants used the iTrackDepression smartphone application to complete daily and weekly self-ratings of depressive symptoms. The PHQ-2 was administered each evening, assessing two core symptoms of depression: loss of interest and depressed mood. Responses were collected on a visual analogue scale from 0 (never) to 10 (all the time) [44–46], resulting in a total score range of 0–20 for the two PHQ-2 items. In addition, the Patient Health Questionnaire-9 (PHQ-9) was completed once per week, capturing the presence and frequency of nine depressive symptoms over the past two weeks, using the standard 4-point Likert scale from 0 (not at all) to 3 (nearly every day) [47]. An overview of PHQ-2 recording duration, number of observations, and symptom severity per participant is provided in Table 2.

Table 2
Participant-level summary of PHQ-2 evening scores.

Participant	Recording duration (weeks)	PHQ-2 observations (n)	PHQ-2 mean $\pm$ SD	PHQ-2 missing (%)
A	37.7	236	13.23 $\pm$ 2.24	10.9
B	44.4	58	4.67 $\pm$ 3.33	81.4
C	51.7	316	3.45 $\pm$ 1.4	12.9
D	47	74	11.69 $\pm$ 4.25	76.7
E	13	68	11.19 $\pm$ 3.24	26.1
F	54	213	7.61 $\pm$ 3.06	43.5
G	54.3	361	10.15 $\pm$ 3.28	5.2
H	54	228	6.22 $\pm$ 3.97	39.8
I	55	192	4.42 $\pm$ 3.43	50.3
J	53.4	315	7.68 $\pm$ 3.48	15.8
K	53.6	300	16.17 $\pm$ 1.74	20.2
L	23	19	6.89 $\pm$ 2.02	85.3
M	55.9	157	10.91 $\pm$ 3.39	59.9
N	55.7	226	6.43 $\pm$ 4.52	42.2
O	59.1	294	6.51 $\pm$ 2.58	29.2
Mobile Sensing				

Mobile sensing was conducted using a Samsung Galaxy Watch 5 and the iTrackDepression app, which continuously recorded call behavior, prompted audio segments, accelerometry, step counts, app usage, message traffic, and optical heart rate variability (HRV). Further technical details can be found in the Mobile Sensing section of the study protocol [42].

Features derived from Mobile Sensors

Daily Segmentation

To examine within-person relationships between passive sensing data and depressive symptoms, we aligned all sensor-derived features to two fixed daily intervals based on local time. Each day, participants completed the PHQ-2 self-rating in the evening, starting from 18:00. We therefore defined the “night” period as the 12 hours preceding the rating (18:00–06:00) and the “daytime” period as the 12 hours directly before the start of rating (06:00–18:00). This segmentation ensured that all sensor features were

temporally aligned with the PHQ-2 outcome, reflecting the participant's current day. In addition, the 12-hour windows capture well-established diurnal patterns in mood, physiology, and behavior, such as differences between sleep-related and wake-related activity patterns [23, 24]. To reduce redundancy among derived features, we applied a two-step collinearity control. First, within homogeneous feature families (e.g., heart rate variability indices or accelerometry-based metrics), we computed pairwise Spearman rank correlations and removed one feature from any pair with very high correlation ($|\rho| \geq 0.90$), keeping the more interpretable and statistically robust one [48]. Second, we applied iterative pruning based on variance inflation, a method to reduce multicollinearity: we removed the least informative feature among those exceeding a conservative threshold (inflation factor > 10), repeating until all remaining values were acceptable or predefined, domain-relevant features remained [49]. Finally, we applied domain rules to avoid mathematically redundant combinations (e.g., raw values and their normalized ratios) and retained features consistent with accepted measurement standards in heart rate variability and activity research [50, 51]. This combination of statistical filtering and domain-based reasoning produced a compact, interpretable feature set.

Time Features

To account for basic calendar-related effects, we included three-time features: study day, day of week, and month. These were treated as simple baseline block outside of sensor data, to control for routine-related variation in mood and behavior.

Speech & Acoustic Features

For each 30–60 second Daily-Dialog recording collected in the evening, we extracted acoustic features using the openSMILE toolkit [52]. We selected a compact, interpretable set covering key voice dimensions: Mean loudness (overall vocal intensity) and loudness variability (emphasis/dynamism over time), Jitter (cycle-to-cycle pitch variability) and shimmer (amplitude instability), reflecting vocal control and the first four mel-frequency cepstral coefficients (MFCCs), summarizing the spectral envelope and timbral qualities of speech. These features capture prosodic, spectral, and stability-related aspects of vocal expression, which have been linked to depressive and suicidal states in prior research [17–19, 53–55].

Phone and Communication Features

We summarized daily phone behavior using four interpretable metrics, separately for day and night segments. Total call duration reflected the intensity of interpersonal engagement. Call frequency captured overall communication activity. The number of unique contacts indexed the breadth of the caller's social network. Missed calls served as a proxy for responsiveness or social availability.

Activity Features

We used wrist-worn accelerometer data to extract movement features that were grouped into three categories: fragmentation, intensity, and rhythmicity. Raw tri-axial signals were resampled, gravity-corrected, and reduced to a single activity signal using the Euclidean Norm Minus One (ENMO). Feature extraction followed the SciKit-Digital-Health (SKDH) framework [50]. Only features with valid data in day and night segments were used. Fragmentation features describe how burst-like movement is over the day. We included power-law exponents from moderate-to-vigorous, light, and sedentary bouts, and the fluctuation scaling exponent (DFA alpha) that captures temporal regularity [22, 25]. Intensity features reflect the amount and duration of physical activity. These include daily step count, total active minutes, and time spent in light, moderate, or vigorous movement. We also included average bout durations and activity during sleep windows, which act as proxies for daytime activity peaks [21, 23]. Rhythmicity features capture how regular the daily activity pattern is. These include spectral power (60-minute bands), entropy measures, balance between day and night activity, explained variance from a 24-hour rhythm model, and the slope of activity accumulation over the day [56, 57].

App Usage & Social Interaction Features

We used two passive Android sensors to measure smartphone interaction: app traffic (data usage) and app usage (screen time). Daily incoming and outgoing bytes were summed separately for communication and social media apps. We also calculated total screen time (foreground duration) for these app types. App categories were based on Google Play metadata and included only those with at least 10,000 installs and present on participants' phones ($n = 43$). Values over 24 hours per day were removed as implausible. App traffic reflects the volume of digital interaction [27], while app screen time indicates active engagement [26]. Both have been linked to mental health outcomes, including stress, anxiety, and depression [15].

Heart-Rate Variability (HRV)

We derived HRV features from optical pulse signals collected by the smartwatch. After cleaning and quality checks using NeuroKit2 [51], we extracted time-domain (e.g., variability), frequency-domain (e.g., spectral power), and nonlinear (e.g., complexity) features from day and night segments. Redundant variables were removed using statistical and physiological criteria. We also included the number of valid heartbeats per day as a proxy for data quality.

Sleep Features

We estimated daily sleep duration using a validated deep-learning model for sleep staging based on heart rate variability [58]. Cleaned inter-beat-interval (IBI) data were segmented and classified as sleep if non-wake stages dominated. Total daily sleep time in hours was then calculated. Preprocessing used NeuroKit2 [51].

Adherence (MNAR) Features

To capture clinically relevant sensor dropout, we derived simple Missing Not at Random (MNAR) indicators across all sensor types. Each flag reflects absence, imbalance, or coverage drop likely tied to user behavior rather than noise. For heart rate, we flagged missing or low-quality data, extreme day–night imbalance, and low total coverage (< 10 h), based on individual thresholds. For activity, we marked days with diurnal imbalance. For phone calls and audio, we flagged full-day absence and tracked consecutive days without data. For sleep, we marked missing sleep, extremely short durations (below personal 10th percentile), and night-only recordings. For app use and traffic, we flagged near-zero use days and selectively low traffic despite phone activity, indicating social disengagement.

Exclusion Procedure Due to Insufficient Data

Participants were anonymized using single-letter identifiers (e.g., A, B, C) and excluded if their time series did not meet the minimal length needed to construct at least one temporally valid train–test split. This requirement was determined by the sum of the minimum training window (60 days) and the minimum test-window constraints (30-day holdout window). Thus, four of the fifteen participants with fewer than 90 valid daily assessments were excluded and not used for model training.

Data Analysis

Data Preprocessing

Daily features were first aligned to the corresponding evening PHQ-2 score, and days without a valid outcome were removed. Because mobile and wearable sensor data naturally contain extreme or physiologically implausible values, often due to recording artefacts, intermittent device use, or measurement noise, we applied a conservative outlier-handling procedure. Values exceeding 3 times the interquartile range beyond the first or third quartile were set to missing [59]. This prevented unrealistic observations from disproportionately influencing downstream modeling. An overview of all features, including summary statistics and missingness, is provided in Additional file 1.

Missing feature values (with up to 80% missingness per participant) were then imputed using multivariate imputation by chained equations (MICE) [60], implemented with scikit-learn's `IterativeImputer` (maximum 50 iterations, convergence tolerance of 1×10^{-2}) [61]. Features were retained despite high missingness because the proportion of missing data alone is not a valid criterion for exclusion; [62] demonstrated that MICE can yield unbiased estimates even with up to 90% missingness when auxiliary information is available. In mobile sensing and mental health contexts, missingness is often Missing Not At Random (MNAR) and reflects user adherence or disengagement, making omission of such variables potentially biasing. Each imputation model was trained solely on the training portion and subsequently applied to the test window, ensuring that no future information leaked into past observations. This multivariate approach maintains the covariance structure among features and is well suited for high-dimensional behavioral time series.

Continuous predictors were transformed using a two-stage scaling process, designed both to harmonize feature distributions and to create optimal conditions for elastic net (EN) regularization. In the first stage, each continuous feature underwent one of three possible transformations depending on its distributional properties: right-skewed variables were stabilized using a natural logarithm transformation ($\log_{10}$), variables with bounded ranges were normalized to the [0, 1] interval using min–max scaling, and approximately symmetric variables were standardized to have mean 0 and standard deviation 1 (see Additional file 2 for the mapping between features and applied scaling methods). These transformations ensured that features spanning different physical units (e.g., app usage seconds, HRV indices, accelerometer-based counts) were brought to a common numerical scale (the scaler per feature is described in the appendix). In the second stage, all features, regardless of the transformation applied in the first stage, were again z-standardized. This second global standardization step ensured uniform, centered, and variance-normalized inputs, which is essential for stable EN fitting and also beneficial for random forest (RF) training in high-dimensional spaces.

Model Training and Validation

For each participant, daily depressive symptom severity was assessed using the evening PHQ-2 sum score, which served as the dependent variable in the models. Two complementary regression model types were considered: an elastic net (EN) regression [63], which has been applied in similar studies [64, 65], and a random forest (RF) regressor [66]. All models were implemented using scikit-learn [61] and tuned using the hyperparameter grids described in Additional file 3.

Hyperparameters for elastic net (EN) and random forest (RF) models were tuned using grid search with R^2 as the optimization metric. For each participant, the available time series was first divided using a forward-in-time split: the first 70% of observations served as the training set and the later 30% as a temporally subsequent, fully unseen test set. Hyperparameter tuning was performed exclusively on the training set to prevent information leakage. Within this training portion, tuning used forward-in-time cross-validation with 5 sequential splits, so that models were always trained on earlier data and validated on later data while preserving temporal order.

To address RQ1 and RQ2, it was examined whether EN and RF models could provide individual-level markers of daily PHQ-2 symptom severity that met predefined performance criteria on unseen future data. After hyperparameters were selected, final EN and RF models were refit using the complete 70% training data and then evaluated once on the held-out 30% future test period, which was not used during model selection or tuning. This procedure reflects a prospective prediction setting, where a model trained on past data is applied to later, unseen observations.

Performance on the held-out test set was quantified using the coefficient of determination (R^2), representing explained variance in daily PHQ-2 scores, and the mean absolute error (MAE), representing the average absolute deviation between predicted and observed scores. A model was considered to yield an objective PHQ-2 marker for a given participant if it met predefined out-of-sample performance criteria on this unseen test set ($R^2 > 0.30$ or $MAE < 1.0$).

Protocol Deviations

Compared with the preregistered analysis protocol [67], several adjustments were made to better reflect the longitudinal, individual-level structure of the data. Participants with fewer than 90 valid daily observations were excluded to ensure a temporally valid training–test configuration. Instead of the preregistered random 70/30 split with shuffled observations and k-fold cross-validation, we used a strictly forward-in-time 70/30 split, reserving the last 30% of each participant’s time series as a fully unseen future test set. Hyperparameter tuning was performed only on the training portion using time-series cross-validation (5 splits).

Preprocessing was adapted by adding IQR-based outlier handling and feature-specific scaling with subsequent z-standardization instead of uniform min–max scaling. Planned explainability analyses were deferred to follow-up work.

Results

RQ 1 - Identification of Individual PHQ-2 Markers with Linear Models

EN models were evaluated at the individual level to determine whether they met the predefined accuracy criteria for representing an objective PHQ-2 marker based on mobile sensing data. Across the 11 participants included in the final analyses, overall prediction performance was low and varied considerably between individuals. R^2 values calculated on the held-out 30% future test data were below zero for all participants (see Table 3), indicating that linear models did not explain a substantial proportion of variance in daily PHQ-2 scores at the individual level.

Table 3
Individual elastic net model performance for estimating daily PHQ-2 scores from mobile sensing data. Reported are R^2 and mean absolute error (MAE) calculated on the held-out 30% future test segment for each participant.

Participant	R^2	MAE
A	-0.09	1.09
B	—	—
C	-0.98	1.22
D	—	—
E	—	—
F	-0.07	1.75
G	-0.06	2.03
H	-0.13	3.39
I	-83.46	2.11
J	-0.02	2.35
K	-4.76	1.75
L	—	—
M	-0.24	2.98
N	-0.07	2.96
O	-0.07	1.35

None of the 11 participants met the required performance thresholds on the held-out 30% test data ($R^2 > 0.30$ or $MAE < 1.0$). Both explained variance and prediction error remained outside the acceptable range for all individuals. In most participants, prediction errors exceeded two PHQ-2 scale points, indicating that the linear models were generally not sufficiently accurate to provide reliable individual-level markers of depressive symptom severity.

RQ 2 - Identification of Individual PHQ-2 Markers with Non-Linear Models

RF regression models were evaluated to examine whether allowing non-linear relationships between mobile sensing features and depressive symptom severity increased the number of participants for whom objective PHQ-2 markers could be identified. Overall, performance on the held-out 30% future test data remained limited and varied across participants (see Table 4). Most participants showed R^2 values

at or below zero, indicating little or no explained variance. One participant (J) had a R^2 value greater than zero, but still below the predefined performance threshold.

Using the predefined eligibility criteria, only one participant (participant I; 1/11, 9.1%) met the required performance threshold based on prediction error ($MAE < 1.0$). No participant met the R^2 criterion ($R^2 > 0.30$), and all remaining participants failed to meet either performance threshold. Compared with the linear elastic net models, the non-linear approach therefore identified one additional participant with acceptable prediction accuracy, but overall performance remained low at the individual level.

Table 4
Individual random forest model performance for estimating daily PHQ-2 scores from mobile sensing data. Reported are R^2 and mean absolute error (MAE) calculated on the held-out 30% future test segment for each participant.

Participant	R^2	MAE
A	-0.62	1.4
B	—	—
C	-1.32	1.4
D	—	—
E	—	—
F	-0.52	2.3
G	-0.18	2.1
H	-0.15	3.4
I	-4.82	0.6
J	0.02	2.4
K	-4.13	1.6
L	—	—
M	-0.01	2.5
N	-0.03	3
O	-0.11	1.5
Comparison of Linear and Non-Linear Modeling Approaches		

Direct comparison of elastic net and random forest results revealed no qualitative difference in overall feasibility between linear and non-linear modeling approaches. For the elastic net models, no participants met the predefined eligibility criteria. In the random forest models, eligibility was observed in only a single participant and was driven exclusively by MAE rather than R^2 , as none of the participants in

either approach achieved the R^2 -based performance threshold. This marginal difference did not alter the overall conclusion that objective PHQ-2 markers could be derived for only a small minority of participants. Extremely negative R^2 values occurred in cases with very low variance in the held-out test segment, where small absolute prediction errors resulted in disproportionately large negative R^2 values.

Discussion

Principal Results

This study examined whether long-term mobile sensing data collected up to one year can support the derivation of objective, individual-level markers of daily depressive symptom severity, as measured by PHQ-2 scores. Overall, the results provide little evidence that such markers can be reliably derived at the individual level. Across participants, prediction performance on previously unseen data was low, with negative R^2 values observed in all cases and only one participant meeting the predefined accuracy threshold based on MAE.

The findings suggest that, in the present dataset, passive sensing features did not provide sufficient information to accurately estimate daily depressive symptom severity for most individuals. While prior work has reported associations between behavioral or physiological signals and depressive symptoms [13, 41], our results indicate that such relationships may be weak, inconsistent, or not sufficiently stable over time to support individual-level prediction in longitudinal settings.

Two factors may help explain these predominantly negative findings. First, the availability of up to one year of data did not translate into improved predictive performance. Exploratory analyses using expanding training windows did not show consistent gains as more data became available, suggesting that longer observation periods alone may not be sufficient to capture stable and predictive relationships between sensing features and symptom severity. Second, limited within-person variability in PHQ-2 scores may have constrained model performance. If symptom levels remain relatively stable over time, the signal-to-noise ratio for prediction is low, making it difficult for models to learn meaningful associations.

Furthermore, the similarity in performance between linear and non-linear models indicates that increasing model flexibility did not meaningfully improve predictive accuracy. This suggests that the primary limitation may lie in the underlying data rather than in the choice of modeling approach. Taken together, the findings highlight the challenges of deriving reliable individual-level markers of depressive symptom severity from passive sensing data in naturalistic settings.

Strengths and Limitations

A key strength of this study is the use of longitudinal, multimodal mobile sensing data collected over an extended period of up to one year, enabling the evaluation of predictive models at the individual level

using temporally separated training and test data. This design provides a stringent test of generalization to future observations.

Several limitations should be considered. First, the observation period, although relatively long, may not have captured sufficient variability in depressive symptom severity within individuals. Limited within-person variation in PHQ-2 scores reduces the ability of models to detect meaningful associations and may partly explain the low predictive performance observed.

Second, the relationship between passive sensing features and depressive symptoms may be intrinsically weak, indirect, or highly dynamic over time. Prior research has demonstrated substantial intraindividual heterogeneity in person-specific relationships between depressive symptoms and longitudinal predictors. In some individuals, these associations can shift within a matter of weeks [68]. Such temporal instability may limit the ability of models trained on one observation window to generalize to unseen data. In addition, behavioral and physiological signals captured via smartphones and wearable devices may not align closely with daily fluctuations in self-reported symptom severity, limiting their usefulness for individual-level prediction.

Third, the small sample size limits the generalizability of the findings and increases the uncertainty around individual-level estimates. Given the heterogeneity of depressive disorders, larger samples are needed to better characterize the conditions under which reliable prediction may be possible.

Finally, although models were evaluated using a temporally held-out test set, the overall low performance indicates that the current modeling approach and available data were insufficient to achieve reliable prediction. Future work should further investigate the temporal dynamics and variability of both symptom measures and sensing features to better understand the conditions required for successful individual-level modeling.

Conclusions

In this 1-year longitudinal mobile sensing study in patients with MD, individual-level machine learning models did not produce reliable indicators of daily depression severity as measured by PHQ-2. Neither linear nor non-linear approaches consistently met predefined performance criteria when tested on new data. These findings indicate that passive smartphone and wearable data alone may be insufficient to estimate short-term symptom severity at the individual level. For the development of artificial intelligence tools in mental health care, the findings underscore the need for cautious interpretation and thorough testing before such models are considered for clinical decision support.

Abbreviations

DFA
Detrended Fluctuation Analysis
ECG

Electrocardiogram
EN
Elastic Net
HRV
Heart Rate Variability
IBI
Inter-Beat Interval
IQR
Interquartile Range
MAE
Mean Absolute Error
MD
Major Depression
MFCC
Mel-Frequency Cepstral Coefficient
MICE
Multivariate Imputation by Chained Equations
MNAR
Missing Not at Random
PHQ-2
Patient Health Questionnaire-2
PHQ-9
Patient Health Questionnaire-9
RF
Random Forest
 R^2
Coefficient of Determination
SD
Standard Deviation
SKDH
SciKit-Digital-Health
WMA
World Medical Association

Declarations

Ethics approval and consent to participate

Ethical approval has been granted by the Ethics Committee of the Department of Medicine at Goethe University Frankfurt, Germany (date: 24/08/2023, reference number: 2023–1309).

Consent for publication

Not applicable

Funding

This work was supported by the German Federal Ministry of Education and Research (BMBF, grant numbers: 13GW0576D, 13GW0576A, 13GW0576B, 13GW0576C, date awarded: November 01, 2021).

Author Contribution

UH: Conceptualization, Funding acquisition, Supervision, Writing – original draft, Writing – review & editing. SS: Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. JL: Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. MP: Writing – original draft, Writing – review & editing, Project administration. SL: Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing – review & editing. AD: Supervision, Writing – review & editing. DH: Supervision, Writing – review & editing. HR: Conceptualization, Funding acquisition, Methodology, Project administration, Writing – original draft, Writing – review & editing. Mondy consortium: acquisition of funding, project administration, methods, review and editing.

Acknowledgement

We sincerely thank all participants who generously contributed their time, effort, and personal experiences to the MONDY clinical trial. Their willingness to engage with the study procedures and share sensitive information made this research possible. We deeply appreciate their commitment and trust, without which this work could not have been conducted.

Data Availability

The data of individual participants underlying the results reported in this manuscript are not publicly available due to privacy considerations but are available upon reasonable request.

References

1. Cui L, Li S, Wang S, Wu X, Liu Y, Yu W, et al. Major depressive disorder: hypothesis, mechanism, prevention and treatment. *Signal Transduct Target Ther.* 2024;9:30. <https://doi.org/10.1038/s41392-024-01738-y>.

2. Liu M, Xie F, Dai Z, Dai Q, Chi X-L. Trend and predictive psychosocial factors of persistent depression/non-depression in Chinese adolescents: A three-year longitudinal study. *PloS One*. 2024;19:e0308303. <https://doi.org/10.1371/journal.pone.0308303>.
3. Monroe SM, Harkness KL. Major depression and its recurrences: Life course matters. *Annu Rev Clin Psychol*. 2022;18:329–57. <https://doi.org/10.1146/annurev-clinpsy-072220-021440>.
4. Strawbridge R, McCrone P, Ulrichsen A, Zahn R, Eberhard J, Wasserman D, et al. Care pathways for people with major depressive disorder: a European Brain Council Value of Treatment study. *Eur Psychiatry J Assoc Eur Psychiatr*. 2022;65:1–21. <https://doi.org/10.1192/j.eurpsy.2022.28>.
5. Thornicroft G, Chatterji S, Evans-Lacko S, Gruber M, Sampson N, Aguilar-Gaxiola S, et al. Undertreatment of people with major depressive disorder in 21 countries. *Br J Psychiatry*. 2017;210:119–24. <https://doi.org/10.1192/bjp.bp.116.188078>.
6. Moitra M, Santomauro D, Collins PY, Vos T, Whiteford H, Saxena S, et al. The global gap in treatment coverage for major depressive disorder in 84 countries from 2000–2019: A systematic review and Bayesian meta-regression analysis. *PLoS Med*. 2022;19:e1003901. <https://doi.org/10.1371/journal.pmed.1003901>.
7. Breuer L, Czaplicki A, Hegerl U, Reich H. Utilization of psychotherapy, pharmacotherapy, and their combination by individuals with current or residual depression: results from five annual nationally representative German surveys. *Sci Rep*. 2025;15:43876. <https://doi.org/10.1038/s41598-025-32371-6>.
8. World health statistics 2017: monitoring health for the SDGs, sustainable development goals. <https://www.who.int/publications/i/item/9789241565486>. Accessed 19 Feb 2026.
9. Ould Brahim L, Lambert SD, Feeley N, Coumoundouros C, Schaffler J, McCusker J, et al. The effects of self-management interventions on depressive symptoms in adults with chronic physical disease(s) experiencing depressive symptomatology: a systematic review and meta-analysis. *BMC Psychiatry*. 2021;21:584. <https://doi.org/10.1186/s12888-021-03504-8>.
10. Pearce M, Garcia L, Abbas A, Strain T, Schuch FB, Golubic R, et al. Association Between Physical Activity and Risk of Depression: A Systematic Review and Meta-analysis. *JAMA Psychiatry*. 2022;79:550–9. <https://doi.org/10.1001/jamapsychiatry.2022.0609>.
11. Conijn JM, Emons WHM, Page BF, Sijsma K, van der Does W, Carlier IV, et al. Response inconsistency of patient-reported symptoms as a predictor of discrepancy between patient and clinician reported depression severity. *Assessment*. 2018;25:917–28. <https://doi.org/10.1177/1073191116666949>.
12. Cuijpers P, Smit F, Bohlmeijer E, Hollon SD, Andersson G. Efficacy of cognitive–behavioural therapy and other psychological treatments for adult depression: meta-analytic study of publication bias. *Br J Psychiatry*. 2010;196:173–8. <https://doi.org/10.1192/bjp.bp.109.066001>.
13. Saeb S, Zhang M, Karr CJ, Schueller SM, Corden ME, Kording KP, et al. Mobile Phone Sensor Correlates of Depressive Symptom Severity in Daily-Life Behavior: An Exploratory Study. *J Med Internet Res*. 2015;17:e175. <https://doi.org/10.2196/jmir.4273>.

14. Stamatis CA, Meyerhoff J, Meng Y, Lin ZCC, Cho YM, Liu T, 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;3:1–8. <https://doi.org/10.1038/s44184-023-00041-y>.
15. Choi A, Ooi A, Lottridge D. Digital Phenotyping for Stress, Anxiety, and Mild Depression: Systematic Literature Review. *JMIR MHealth UHealth.* 2024;12:e40689. <https://doi.org/10.2196/40689>.
16. Aledavood T, Luong N, Baryshnikov I, Darst R, Heikkilä R, Holmén J, 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:e63622. <https://doi.org/10.2196/63622>.
17. Moore E, Clements MA, Peifer JW, Weisser L. Critical analysis of the impact of glottal features in the classification of clinical depression in speech. *IEEE Trans Biomed Eng.* 2008;55:96–107. <https://doi.org/10.1109/TBME.2007.900562>.
18. Quatieri TF, Malyska N. Vocal-source biomarkers for depression: a link to psychomotor activity. In: *Interspeech 2012.* ISCA; 2012. p. 1059–62. <https://doi.org/10.21437/Interspeech.2012-311>.
19. Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri TF. A review of depression and suicide risk assessment using speech analysis. *Speech Commun.* 2015;71:10–49. <https://doi.org/10.1016/j.specom.2015.03.004>.
20. Stasak B, Epps J, Schatten HT, Miller IW, Provost EM, Arney MF. Read speech voice quality and disfluency in individuals with recent suicidal ideation or suicide attempt. *Speech Commun.* 2021;132:10–20. <https://doi.org/10.1016/j.specom.2021.05.004>.
21. Bizzozero-Peroni B, Díaz-Goñi V, Jiménez-López E, Rodríguez-Gutiérrez E, Sequí-Domínguez I, Núñez de Arenas-Arroyo S, et al. Daily Step Count and Depression in Adults: A Systematic Review and Meta-Analysis. *JAMA Netw Open.* 2024;7:e2451208. <https://doi.org/10.1001/jamanetworkopen.2024.51208>.
22. Smagula SF, Ancoli-Israel S, Blackwell T, Boudreau R, Stefanick ML, Paudel ML, et al. Circadian rest-activity rhythms predict future increases in depressive symptoms among community-dwelling older men. *Am J Geriatr Psychiatry Off J Am Assoc Geriatr Psychiatry.* 2015;23:495–505. <https://doi.org/10.1016/j.jagp.2014.06.007>.
23. Wang X, Pathiravasan CH, Zhang Y, Trinquart L, Borrelli B, Spartano NL, et al. Association of Depressive Symptom Trajectory With Physical Activity Collected by mHealth Devices in the Electronic Framingham Heart Study: Cohort Study. *JMIR Ment Health.* 2023;10:e44529. <https://doi.org/10.2196/44529>.
24. Sun S, Folarin AA, Zhang Y, Cummins N, Garcia-Dias R, Stewart C, 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. <https://doi.org/10.2196/45233>.
25. Ho FY-Y, Poon C-Y, Wong VW-H, Chan K-W, Law K-W, Yeung W-F, et al. Actigraphic monitoring of sleep and circadian rest-activity rhythm in individuals with major depressive disorder or depressive

- symptoms: A meta-analysis. *J Affect Disord.* 2024;361:224–44.
<https://doi.org/10.1016/j.jad.2024.05.155>.
26. Tian Y, Zhou K, Pelleg D. What and How long: Prediction of Mobile App Engagement. *ACM Trans Inf Syst.* 2022;40:1–38. <https://doi.org/10.1145/3464301>.
 27. Yue C, Ware S, Morillo R, Lu J, Shang C, Bi J, et al. Automatic depression prediction using Internet traffic characteristics on smartphones. *Smart Health.* 2020;18:100137.
<https://doi.org/10.1016/j.smhl.2020.100137>.
 28. Carney RM, Freedland KE. Depression and heart rate variability in patients with coronary heart disease. *Cleve Clin J Med.* 2009;76 4 suppl 2:S13–7. <https://doi.org/10.3949/ccjm.76.s2.03>.
 29. Kemp AH, Quintana DS, Gray MA, Felmingham KL, Brown K, Gatt JM. Impact of Depression and Antidepressant Treatment on Heart Rate Variability: A Review and Meta-Analysis. *Biol Psychiatry.* 2010;67:1067–74. <https://doi.org/10.1016/j.biopsych.2009.12.012>.
 30. Lim MH, Eres R, Vasan S. Understanding loneliness in the twenty-first century: an update on correlates, risk factors, and potential solutions. *Soc Psychiatry Psychiatr Epidemiol.* 2020;55:793–810. <https://doi.org/10.1007/s00127-020-01889-7>.
 31. Rykov Y, Thach T-Q, Bojic I, Christopoulos G, Car J. Digital Biomarkers for Depression Screening With Wearable Devices: Cross-sectional Study With Machine Learning Modeling. *JMIR MHealth UHealth.* 2021;9:e24872. <https://doi.org/10.2196/24872>.
 32. Lamichhane B, Moukaddam N, Sabharwal A. Mobile sensing-based depression severity assessment in participants with heterogeneous mental health conditions. *Sci Rep.* 2024;14:18808.
<https://doi.org/10.1038/s41598-024-69739-z>.
 33. Zhuparris A, Maleki G, van Londen L, Koopmans I, Aalten V, Yocarini IE, et al. A smartphone- and wearable-based biomarker for the estimation of unipolar depression severity. *Sci Rep.* 2023;13:18844. <https://doi.org/10.1038/s41598-023-46075-2>.
 34. Wright AGC, Woods WC. Personalized Models of Psychopathology. *Annu Rev Clin Psychol.* 2020;16:49–74. <https://doi.org/10.1146/annurev-clinpsy-102419-125032>.
 35. 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;13:1–10. <https://doi.org/10.1038/s41398-023-02669-y>.
 36. Amin R, Schreynemackers S, Oppenheimer H, Petrovic M, Hegerl U, Reich H. Use of Mobile Sensing Data for Longitudinal Monitoring and Prediction of Depression Severity: Systematic Review. *J Med Internet Res.* 2025;27:e57418. <https://doi.org/10.2196/57418>.
 37. Minich M, Moreno M. Real-world adolescent smartphone use is associated with improvements in mood: An ecological momentary assessment study. *PloS One.* 2024;19:e0298422.
<https://doi.org/10.1371/journal.pone.0298422>.
 38. Xu P, Huang Y, Hou Q, Cheng J, Ren Z, Ye R, et al. Relationship between physical activity and mental health in a national representative cross-section study: Its variations according to obesity and

- comorbidity. *J Affect Disord*. 2022;308:484–93. <https://doi.org/10.1016/j.jad.2022.04.037>.
39. Zhang Y, Folarin AA, Sun S, Cummins N, Ranjan Y, Rashid Z, et al. Predicting Depressive Symptom Severity through Individuals' Nearby Bluetooth Devices Count Data Collected by Mobile Phones: A Preliminary Longitudinal Study. *JMIR MHealth UHealth*. 2021;9:e29840. <https://doi.org/10.2196/29840>.
40. Ben-Zeev D, Scherer EA, Wang R, Xie H, Campbell AT. Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health. *Psychiatr Rehabil J*. 2015;38:218–26. <https://doi.org/10.1037/prj0000130>.
41. Jacobson NC, Weingarden H, Wilhelm S. Digital biomarkers of mood disorders and symptom change. *Npj Digit Med*. 2019;2:3. <https://doi.org/10.1038/s41746-019-0078-0>.
42. Reich H, Schreynemackers S, Amin R, Ludwig S, Zippelius J, Leimhofer J, et al. Links between self-monitoring data collected through smartphones and smartwatches and the individual disease trajectories of adult patients with depressive disorders: Study protocol of a one-year observational trial. *Contemp Clin Trials Commun*. 2025;45:101492. <https://doi.org/10.1016/j.conctc.2025.101492>.
43. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA*. 2013;310:2191–4. <https://doi.org/10.1001/jama.2013.281053>.
44. Arrieta J, Aguerrebere M, Raviola G, Flores H, Elliott P, Espinosa A, et al. Validity and Utility of the Patient Health Questionnaire (PHQ)-2 and PHQ-9 for Screening and Diagnosis of Depression in Rural Chiapas, Mexico: A Cross-Sectional Study. *J Clin Psychol*. 2017;73:1076–90. <https://doi.org/10.1002/jclp.22390>.
45. Lorenz N, Sander C, Ivanova G, Hegerl U. Temporal Associations of Daily Changes in Sleep and Depression Core Symptoms in Patients Suffering From Major Depressive Disorder: Idiographic Time-Series Analysis. *JMIR Ment Health*. 2020;7:e17071. <https://doi.org/10.2196/17071>.
46. Siepe BS, Sander C, Schultze M, Kliem A, Ludwig S, Hegerl U, et al. Time-Varying Network Models for the Temporal Dynamics of Depressive Symptomatology in Patients With Depressive Disorders: Secondary Analysis of Longitudinal Observational Data. *JMIR Ment Health*. 2024;11:e50136. <https://doi.org/10.2196/50136>.
47. Kroenke K, Spitzer RL. The PHQ-9: A new depression diagnostic and severity measure. *Psychiatr Ann*. 2002;32:509–15. <https://doi.org/10.3928/0048-5713-20020901-06>.
48. Dormann CF, Elith J, Bacher S, Buchmann C, Carl G, Carré G, et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*. 2013;36:27–46. <https://doi.org/10.1111/j.1600-0587.2012.07348.x>.
49. O'brien RM. A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Qual Quant*. 2007;41:673–90. <https://doi.org/10.1007/s11135-006-9018-6>.
50. Adamowicz L, Christakis Y, Czech MD, Adamusiak T. SciKit Digital Health: Python Package for Streamlined Wearable Inertial Sensor Data Processing. *JMIR MHealth UHealth*. 2022;10:e36762. <https://doi.org/10.2196/36762>.

51. Makowski D, Pham T, Lau ZJ, Brammer JC, Lespinasse F, Pham H, et al. NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behav Res Methods*. 2021;53:1689–96. <https://doi.org/10.3758/s13428-020-01516-y>.
52. Eyben F, Wöllmer M, Schuller B. openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor. *Proc. ACM Multimedia (MM)*; 2010. p. 1459–62.
53. Cummins N, Sethu V, Epps J, Schnieder S, Krajewski J. Analysis of acoustic space variability in speech affected by depression. *Speech Commun*. 2015;75:27–49. <https://doi.org/10.1016/j.specom.2015.09.003>.
54. Dhelim S, Chen L, Ning H, Nugent C. Artificial intelligence for suicide assessment using Audiovisual Cues: a review. *Artif Intell Rev*. 2022. <https://doi.org/10.1007/s10462-022-10290-6>.
55. Stasak B, Epps J, Schatten HT, Miller IW, Provost EM, Arney MF. Read speech voice quality and disfluency in individuals with recent suicidal ideation or suicide attempt. *Speech Commun*. 2021;132:10–20. <https://doi.org/10.1016/j.specom.2021.05.004>.
56. Luik AI, Zuurbier LA, Hofman A, Van Someren EJW, Tiemeier H. Stability and fragmentation of the activity rhythm across the sleep-wake cycle: the importance of age, lifestyle, and mental health. *Chronobiol Int*. 2013;30:1223–30. <https://doi.org/10.3109/07420528.2013.813528>.
57. Xavier NB, Abreu ACVO, Amando GR, Steibel EG, Pilz LK, Freitas JJ, et al. Chronobiological parameters as predictors of early treatment response in major depression. *J Affect Disord*. 2023;323:679–88. <https://doi.org/10.1016/j.jad.2022.12.002>.
58. Brunner C, Hofer F. SleepECG: a Python package for sleep staging based on heart rate. *J Open Source Softw*. 2023;8:5411. <https://doi.org/10.21105/joss.05411>.
59. Wilcox R. Introduction to robust estimation and hypothesis testing. 3rd ed. 2012. <https://doi.org/10.2307/2669876>.
60. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? *Int J Methods Psychiatr Res*. 2011;20:40–9. <https://doi.org/10.1002/mpr.329>.
61. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12:2825–30.
62. Madley-Dowd P, Hughes R, Tilling K, Heron J. The proportion of missing data should not be used to guide decisions on multiple imputation. *J Clin Epidemiol*. 2019;110:63–73. <https://doi.org/10.1016/j.jclinepi.2019.02.016>.
63. Zou H, Hastie T. Regularization and Variable Selection via the Elastic Net. *J R Stat Soc Ser B Stat Methodol*. 2005;67:301–20.
64. Balliu B, Douglas C, Seok D, Shenhav L, Wu Y, Chatzopoulou D, et al. Personalized mood prediction from patterns of behavior collected with smartphones. *Npj Digit Med*. 2024;7:1–14. <https://doi.org/10.1038/s41746-024-01035-6>.
65. 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;13:1–10. <https://doi.org/10.1038/s41398-023-02669-y>.
66. Breiman L. Random Forests. *Mach Learn*. 2001;45:5–32. <https://doi.org/10.1023/A:1010933404324>.
67. Hegerl U, Schreynemackers S, Petrovic M, Ludwig S, Leimhofer J, Dominik A, et al. Methods Protocol: Can AI-enabled analyses of smartphones and wearable data collected over one year provide patients suffering from depression with an objective marker of disease severity? 2025. <https://doi.org/10.17605/OSF.IO/XDU3P>.
68. Siepe BS, Sander C, Schultze M, Kliem A, Ludwig S, Hegerl U, et al. Time-Varying Network Models for the Temporal Dynamics of Depressive Symptomatology in Patients With Depressive Disorders: Secondary Analysis of Longitudinal Observational Data. *JMIR Ment Health*. 2024;11:e50136. <https://doi.org/10.2196/50136>. the MONDY Consortium
69. Mohamed Alhaskir ⁱ, Rebeka Amin ^{b c}, Angela Carell ^j, Tobias Dunker ^j, Kismet Ekinci ^j, Florian Fischer ⁱ, Enrico Lohmann ^f, Elisabeth Schriewer ⁱ, Yvonne Weber ⁱ, Stefan Wolking ⁱ, Jil Zippelius ^{bc}
70. ⁱ Section of Epileptology, Department of Neurology, and Institute for Medical Informatics, RWTH University of Aachen, Aachen, Germany^j adesso SE, Dortmund, Germany

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [summarystatisticsandmissingnessacrossfeatures.docx.docx](#)
- [featureconfig.docx](#)
- [hyperparametergrids.docx](#)