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?

Ulrich Hegerl [b e], Simon Schreynemackers [a b], Milica Petrovic [a b], Sascha Ludwig [c], Johannes Leimhofer [a d], Andreas Dominik [f], Dominik Heider [g], Hanna Reich [a b], the MONDY Consortium

[a] Research Centre of the German Foundation for Depression and Suicide Prevention, Department of Psychiatry, Psychosomatic Medicine and Psychotherapy, University Hospital, Goethe University, Heinrich-Hoffmann-Straße 10, 60528, Frankfurt am Main, Germany
[b] German Foundation for Depression and Suicide Prevention, Goerdelerring 9, 04109, Leipzig, Germany
[c] Institute for Applied Informatics, University Leipzig, Goerdelerring 9, 04109, Leipzig, Germany
[d] Institute of Computer Science, Heinrich-Heine-University, Graf-Adolf-Straße 63, 40210, Düsseldorf, Germany
[e] Goethe Research Professorship, Department for Psychiatry, Psychosomatics and Psychotherapy, University Hospital, Goethe University, Heinrich-Hoffmann-Straße 10, 60528, Frankfurt am Main, Germany
[f] University of Applied Sciences, Department of Mathematics Natural Sciences and Informatics, Life Science Informatics Group, Giessen, Germany
[g] University of Münster, Institute of Medical Informatics, Münster, Germany

Introduction

This document outlines the predefined methodology for addressing the research questions presented in the Overview and Research Questions section. Upon approval, it will be uploaded to a Git repository, and released via Zenodo to generate a Digital Object Identifier (DOI), which serves as the official protocol for the forthcoming data analysis.

The data analyzed in this work originates from an exploratory study conducted in accordance with the study protocol described by (Reich et al., 2025). The protocol is based on the clinical trial registered under DRKS00032618 (https://drks.de/search/en/trial/DRKS00032618).

The complete analysis was performed using Python, particularly using the capabilities of the scikit-learning library (Pedregosa et al., 2011) for building machine learning pipelines.

Overview and Research Questions

This proof-of-concept study aims to analyze whether AI-based individual analysis of long-term time series comprising data collected via smartphone and smartwatch (i.e., voice, activity, smartphone usage, self-ratings) can provide at least some patients with objective information about changes in the severity of their depression. It will thereby address the research questions RQI.a and RQI.b as described in (Reich et al., 2025), by answering the following specific research questions:

1. In how many of the enrolled patients with clinically diagnosed depression can analysis of their individual, up-to-one-year passive time-series data (variables described in section Data Collection) identify objective markers that estimate the severity of depressive symptomatology as measured by daily PHQ-2 scores?

2. For patients in whom a significant linear prediction is achieved, which sensor or app-derived data source contributes most to the explanation of individual depression severity?

3. How does the application of non-linear ensemble methods influence the results of research questions 1 and 2? I.e., how many additional patients exhibit a reliable predictive pattern, and do the dominant data sources shift when non-linear relationships are allowed?

Study Design and Cohort

The present investigation draws on our data collected in the MONDY exploratory n-of-1 trial described by (Reich et al., 2025) (German Clinical Trials Register DRKS00032618). To summarize briefly, N = 15 adults with recurrent major depression were monitored for up to twelve months. Participants complete the Patient Health Questionnaire-2 (PHQ-2, (Arrieta et al., 2017)) every evening and a full Patient Health Questionnaire-9 (PHQ-9, (Kroenke & Spitzer, 2002)) once per week, providing repeated self-reported measures of depressive symptom severity. Mobile sensing was carried out with a Samsung Galaxy Watch 5 and the companion "iTrackDepression" smartphone application, which continuously recorded telephone activity, prompted voice segments for acoustic analysis, step counts, wrist accelerometry, app-usage logs, message traffic and optical heart-rate–derived heart-rate-variability (HRV), as detailed in the Mobile Sensing section of the Reich et al. protocol.

For the present models we extracted daily features exclusively from behavioral and physiological data recorded between 06:00 a.m. to 6:00 p.m. This daytime window precedes the evening PHQ-2 assessment (available from 6:00 p.m. to 0:00 a.m.). All features described in subsection Passive Sensors of the Data Collection section originate from these daytime sensor streams, while the corresponding PHQ-2 (daily) scores serve as the outcomes to be estimated.

Background

Passive Sensors

Phone and Communication Data

Smartphone call logs offer a low-burden, privacy-preserving lens on everyday social behavioral, an area that reliably deteriorates in major depressive episodes (Kupferberg et al., 2016). Early work by (Saeb et al., 2015) showed that shorter and fewer calls discriminated individuals with PHQ-9–defined depression with an AUC of 0.74. A systematic review of 39 smartphone-sensor studies identified call duration and call frequency as the most consistently replicated digital markers of mild depression (Choi et al., 2024). In a large (n = 1,013) longitudinal cohort, within-person drops in incoming and outgoing call counts forecasted higher subsequent PHQ-8 scores (Stamatis et al., 2024). Complementing these population findings, a recent multimodal phenotyping study of clinically diagnosed patients reported a positive association between outgoing call duration and depression severity ($\beta = 0.05$) (Aledavood et al., 2025).

Speech & Acoustic Features

A growing body of work shows that deviations in vocal production carry reliable signatures of affective and cognitive state. Early evidence demonstrated that glottal-source measures improve the automatic detection of major depressive disorder (Moore et al., 2008), and subsequent studies linked source-related descriptors to psychomotor retardation in depression (Quatieri & Malyska, 2012). Comprehensive reviews have since catalogued intensity, prosodic variability,

perturbation (jitter, shimmer), and spectral markers such as mel-frequency cepstral coefficients (MFCCs) as salient for depression and suicidality screening (Cummins, Scherer, et al., 2015; Cummins, Sethu, et al., 2015). More recent investigations confirm that voice-quality shifts and disfluencies distinguish individuals with recent suicidal ideation or attempts (Stasak et al., 2021), while convergent evidence from multimodal artificial-intelligence reviews positions acoustic biomarkers as a cornerstone of technology-enabled suicide risk assessment (Dhelim et al., 2022). Against this backdrop, we extracted a compact set of clinically motivated acoustic features from each daily voice recording (see (Reich et al., 2025)) to support downstream mental-health modelling.

Step and Activity Data

A systematic review and meta-analysis (33 observational studies, N = 96 173) found a dose-response between daily steps and depressive symptoms: compared with <5 000 steps/day, 5 000–7 499 steps were linked to fewer symptoms (SMD = -0.17), 7 500–9 999 steps to a larger reduction (SMD = -0.27) and ≥10 000 steps to a similar benefit (SMD = -0.26). Prospective cohorts showed that achieving ≥7 000 steps/day lowered incident depression risk by 31 % (RR = 0.69), with each additional 1 000 steps/day conferring ~9 % extra protection (Bizzozero-Peroni et al., 2024). Passive sensing studies echo these findings. In the RADAR-CNS retrospective analysis (>4 200 14-day windows), step count was the top longitudinal predictor of PHQ-8 scores (repeated-measures r = -0.14) and ranked second cross-sectionally (r = -0.19), outperforming sleep, screen-time and geolocation features (Sun et al., 2023). In the Electronic Framingham Heart Study, participants on moderate depressive-symptom trajectories walked 823 fewer daily steps (95 % CI -1 421 to -226) than those with persistently low symptoms over a one-year follow-up (Wang et al., 2023). The step sum after the day is therefore an evident predictor for depression severity.

App Usage & Social Interaction

A recent systematic review by (Choi et al., 2024) identified digital behavioral markers such as app usage patterns, call logs, and SMS activity as significant correlates of stress, anxiety, and depression. App foreground time serves as a reliable proxy for on-screen engagement, reflecting the duration of direct user interaction with specific applications (Tian et al., 2022). Additionally, network traffic associated with communication and social media apps can provide insight into the volume of information exchange, offering an indirect measure of social interaction (Yue et al., 2020).

Heart-Rate Variability

Heart-rate-variability indices capture the neuro-cardiac imbalance that accompanies dysregulation of emotion-related autonomic circuits, and a growing body of evidence shows that lower or dysregulated HRV predicts current depressive symptom severity, future onset and treatment response across community, perinatal, occupational and cardiac cohorts (Carney & Freedland, 2009; da Estrela et al., 2021; de Vries et al., 2022; Kemp et al., 2010; Kumar et al., 2024; Lim et al., 2020; Singh Solorzano et al., 2022).

Sleep Data

Recent evidence shows that heart-rate-variability signals alone suffice for accurate, fully automated sleep staging. The latest systematic review by (Yu et al., 2025) synthesized 65 studies published between 2010 and 2024 and reported median epoch-level accuracies of ≈ 78 % for three-stage (Wake, NREM, REM) and ≈ 70 % for four-stage classification — approaching human inter-scorer agreement in polysomnography (Nie et al., 2024). Among the seminal contributions highlighted are the LSTM architecture of (Radha et al., 2019), which reached κ = 0.61 on 5 584 MESA nights and demonstrated the stage-specific value of temporal HRV patterns (Radha et al.,

2019). The fully convolutional model of (Sridhar et al., 2020)., was trained on > 10 000 nights and achieved 0.77 accuracy (κ = 0.66) on an external clinical cohort (Sridhar et al., 2020). The wearable-oriented network of (Fonseca et al., 2023), which maintained κ ≈ 0.64 while reducing inference time fifty-fold, underscoring the feasibility of on-device staging (Fonseca et al., 2023). These findings corroborate that robust and computationally economical HRV-based staging, particularly with combined CNN- and GRU-type architectures, is attainable.

Data Collection

Active Self-Ratings

In the evening assessments, daily self-reported depressive symptom severity is assessed using the Patient Health Questionnaire-2 (PHQ-2 (Arrieta et al., 2017)), a very short measure assessing two core symptoms of depression (loss of interest/joylessness and feeling down/depressed/hopeless). Following previously established procedures (Lorenz et al., 2020; Siepe, 2022), answers are given on a Visual Analogue Scale ranging from 0-10 with labelled endpoints (0 = never, 10 = all the time).

Passive Sensors

Phone and Communication Data

For the purpose of the current analyses, four daily log-derived metrics capture complementary facets of individual behavior: total duration calls (interaction intensity), total calling frequency (activity volume), number of contacts (network breadth), and number of missed calls (responsiveness). Collectively, the evidence (see section Background) supports using these four call-based features as objective, passively sensed proxies of social withdrawal, anhedonia, and diminished engagement central to depressive pathology.

Speech & Acoustic Features

For every 30- to 60-second Daily-Dialog utterance, we used openSMILE to compute key speech descriptors. Intensity (loudness_sma3_amean) reflects the perceived sound-pressure level, whereas its variability (loudness_sma3_stddevNorm) captures dynamism and emphasis over time. The mean fundamental frequency (f_0) expressed in semitones relative to 27.5 Hz (F0semitoneFrom27.5Hz_sma3nz_amean) characterizes average vocal pitch, and pitch sigma (F0semitoneFrom27.5Hz_sma3nz_stddevNorm) quantifies melodic fluctuation. Jitter (jitterLocal_sma3nz_amean) measures cycle-to-cycle period perturbations, indexing phonatory stability, while shimmer (shimmerLocaldB_sma3nz_amean) assesses amplitude consistency. The harmonics-to-noise ratio gauges the balance between periodic and aperiodic energy, serving as a roughness–clarity indicator. Finally, the first four MFCCs (mfcc1-4V_sma3nz_amean) condense the spectral envelope, capturing timbral and articulatory attributes. Collectively, these parameters yield a concise yet information-rich representation of prosodic, spectral, and voice-quality characteristics for each response, in line with vocal biomarkers previously implicated in mood and suicide-risk assessment (Cummins, Scherer, et al., 2015; Cummins, Sethu, et al., 2015; Dhelim et al., 2022; Moore et al., 2008; Quatieri & Malyska, 2012; Stasak et al., 2021).

Step and Activity Data

To retrieve activity data, raw tri-axial wrist-accelerometer signals were resampled, gravity-/orientation-corrected and collapsed to the Euclidean Norm minus One (ENMO), which the SciKit-Digital-Health (SKDH) pipeline averaged in non-overlapping 5-s epochs. To reduce noise yet preserve short, vigorous bursts, the algorithm assigns each clock minute the highest of these epoch means found in rolling 6-, 15- and 60-min windows (max_accel_lens = 6, 15, 60 min) (Lin et

al., 2023). Sleep and non-wear segments are masked. The remaining minute-wise ENMO values are then mapped to intensity bands validated against indirect calorimetry for the dominant wrist: < 50 mg (sedentary), 50–110 mg (light), 110–440 mg (moderate) and ≥ 440 mg (vigorous) (Migueles et al., 2019). ENMO itself exhibits a near-linear relationship with activity energy expenditure measured by doubly-labelled-water techniques in free-living adults, supporting its use as a metabolic proxy (White et al., 2019). Summing the classified minutes across 12 h yields daily totals of sedentary time, light activity and moderate-to-vigorous physical activity (MVPA) in minutes.

App Usage & Social Interaction

Daily communication- and social-media metrics were derived from the passive Android virtual smartphone sensor app traffic (per-package network bytes) and app usage (per-package foreground time). We first restricted the analysis to apps with ≥10 000 Google Play installs in the Communication or Social categories and that appeared on at least one participant's handset, yielding 43 package IDs. Each sensor record contained (1) start–end timestamps together with Wi-Fi and mobile byte counters returned by the Android functions NetworkStats.Bucket.getRxBytes()/getTxBytes() and (2) the cumulative foreground time per package provided by UsageStatsManager.getTotalTimeInForeground() at INTERVAL_DAILY resolution. All timestamps were converted to Central European Time (Europe/Berlin) and analyses were limited to waking hours (06:00–18:00).

For every patient and calendar day we then summed (Wi-Fi + mobile) bytes across communication apps to obtain total communication app data received and transmitted and across social-media apps to obtain total social media app data received and transmitted. Analogously, we computed total communication app usage and total social media app usage by differencing successive foreground-time counters, converting milliseconds to hours, and discarding physiologically implausible outliers (>24 h). These features quantify the volume of information flow and on-screen engagement in two behaviorally distinct domains and have been highlighted in recent systematic reviews as robust digital correlates of stress, anxiety and depression (Choi et al., 2024).

Heart-Rate Variability

Day-time inter-beat-intervals (IBI) timeseries (06:00–18:00 UTC) were pre-processed with NeuroKit2 v0.2, which applied beat localization, ectopic-beat correction and manufacturer quality-flags. Only segments containing ≥ 300 clean IBIs were retained (Makowski et al., 2021). From each segment we extracted the full NeuroKit2 HRV panel (25 time-domain and 10 frequency-domain metrics; 35 variables in total). All variables were then scaled. Millisecond measures were z-standardized, spectral powers were $\log_1$p-transformed and the LF/HF ratio was min–max rescaled before multicollinearity pruning. First, predictors exhibiting an absolute pair-wise Spearman rank correlation > 0.90 were discarded. The survivors underwent an iterative variance-inflation analysis; any variable with VIF > 10 was removed until all remaining VIFs were ≤ 10, consistent with the guideline proposed (O'brien, 2007). Where statistical and physiological considerations conflicted, domain knowledge guided retention of the more informative metric. This two-stage procedure reduced the 35 candidates to 15 mutually non-redundant HRV indices that constitute the predictor block for all subsequent models. In addition, the number of used IBI datapoints was deliberately kept for every day-segment because the amount of analyzable data is a strong determinant of HRV estimate reliability and therefore acts as a wear-time/quality covariate that can be adjusted for in downstream models. All Heart-Rate features including the eliminated and final ones are listed at Table 1.

Table 1. Heart-Rate variability features.

Method	Features
Initial NeuroKit2 features	MeanNN, SDNN, SDANN1, SDNNI1, SDANN2, SDNNI2, SDANN5, SDNNI5, RMSSD, SDSD, CVNN, CVSD, MedianNN, MadNN, MCVNN, IQRNN, SDRMSSD, Prc20NN, Prc80NN, pNN50, pNN20, MinNN, MaxNN, HTI, , ULF, VLF, LF, HF, VHF, TP, LFHF, LFn, HFn, LnHF,
Pairwise Spearman filter (ρ > 0.9) + domain review	Variables dropped due to high correlation: SDNNI1, SDANN2, SDNNI2, SDNNI5, SDSD, HTI, CVNN, pNN20, HFn,TP, VHF
Iterative VIF > 10 +	Variables dropped due to multicollinearity: SDNN (VIF=1410.44) CVSD (VIF=1161.03) Prc20NN (VIF=466.26) Prc80NN (VIF=241.55) LnHF (VIF=155.36) MCVNN (VIF=89.45) LF (VIF=75.82) MaxNN (VIF=60.53) LFn (VIF=44.95) RMSSD (VIF=38.00) SDANN1 (VIF=25.42) pNN50 (VIF=18.29) VLF (VIF=14.70) IQRNN (VIF=13.58)
Final HRV features	MeanNN (VIF=5.214) SDANN5 (VIF=2.025) MedianNN (VIF=5.147) MadNN (VIF=2.689) SDRMSSD (VIF=2.976) MinNN (VIF=7.602) ULF (VIF=4.475) HF (VIF=3.248) LFHF (VIF=7.445)

Sleep Data

The pre-trained wrn-gru-mesa classifier implemented in SleepECG (Brunner & Hofer, 2023) was adopted for this work. Day-time (06:00–18:00 UTC) inter-beat-interval (IBI) series, identical to those used for HRV feature extraction, were beat localized, ectopic-beat corrected, and quality-masked with NeuroKit2 (Makowski et al., 2021). Clean beat times were segmented into 30-s epochs and fed to the wrn-gru-mesa model, a wide-residual CNN followed by a GRU sequence-aggregator trained on 1971 MESA nights and validated on 1000 SHHS nights, where it achieved 75 % epoch-level accuracy (κ = 0.54) for three-stage scoring (Brunner & Hofer, 2023). An epoch was labeled as sleep whenever the condition P_Wake < P_NREM + P_REM was met. Within the intraday period from 06:00–18:00 and the last night period from 18:00-6:00, successive beat times yielded new IBIs. Implausible gaps (> 10 s) were discarded, and the remaining sleep-flagged IBIs were summed to derive each participant's daily sleep duration in hours.

Data Preparation and Quality Control

Missing-Value Imputation

Missing values in the daily feature set were imputed using multivariate imputation by chained equations (MICE) to preserve inter-feature correlations and appropriately quantify uncertainty in the imputed values. MICE iteratively models each feature with missing entries as a function of the other features, refining estimates in a chained sequence until convergence, and is well-suited for longitudinal and psychiatric research settings (Azur et al., 2011). This approach allowed leveraging the multivariate structure of the daily feature matrix while maintaining consistency with the observed data distributions

Scaling

Prior to model training, all features were scaled using min-max scaling. This normalization method transforms each feature to a fixed range, typically [0, 1]. In the context of linear models, where feature importance can be interpreted based on the magnitude of model coefficients, consistent scaling across features is essential. By placing all variables on the same scale, Min-Max scaling allows for a more meaningful comparison of coefficient values, helping to identify which features contribute most to the model. This is particularly relevant for our work, where the primary aim is not the prediction of depression severity, but rather the identification of interpretable digital markers of depression.

Data Partitioning

For each individual's daily time series, data were partitioned into 70 % training and 30 % test data with random shuffling. The training data set was used for model fitting including hyperparameter tuning via five-fold cross-validation, while the test data was used for model evaluation.

Depression Variability & Drop-Out

Consistent prediction requires sufficient day-to-day fluctuation in the target variable. (Price et al., 2023) showed that patient-specific models lose almost all predictive value when weekly mood scores vary little, and (Balliu et al., 2024) likewise reported near-zero performance in individuals with flat Computerized Adaptive Test–Depression Inventory (CAT-DI) trajectories. Following these observations, we quantified within-person symptom variability using the root mean square of successive differences (RMSSD) computed from each participant's PHQ scores (Price et al., 2023). We computed each participant's PHQ-2 and PHQ-9 RMSSD over the entire observation window and provisionally flagged those whose value lay below the cohort's 25th percentile as "low-variance candidates". Additionally after model fitting, any case whose predictions failed to outperform a constant baseline ($R^2 \leq 0$ or MAPE lower equal to the mean-only model), indicating effectively invariant symptoms despite regular assessment, was flagged as a "flat-trajectory" profile. These participants are reported separately but are not counted in the primary responder analysis, thereby preserving full sample inclusion during training while avoiding over-interpretation of trajectories that contain no learnable signal. This procedure adapts the performance-based filtering strategy advocated by (Balliu et al., 2024; Price et al., 2023), and eliminates arbitrary variance cut-offs while retaining maximally transparent reporting.

Modeling

Training

To estimate the daily depression severity scores, two complementary regression approaches were employed: elastic net and random forest regression.

For the linear modeling approach, elastic net regression was applied, combining L1 (Lasso) and L2 (Ridge) regularization with an equal mixing parameter of $\alpha = 0.5$. The regularization strength (λ) was selected via 5-fold cross-validation on the training data segment.

For the nonlinear modeling approach, random forest regression was utilized. The hyperparameters, including the number of trees and maximum tree depth, were tuned using 5-fold cross-validation on the training segment.

Performance Evaluation

After hyperparameter tuning on the training set, we quantified predictive accuracy on the test set that had remained completely unseen during model development. For both the elastic net and random forest models, performance was summarized by the mean absolute error (MAE) and, specifically for the elastic net, by the coefficient of determination (R^2) on the test set.

Feature Importance

Feature importance was assessed by ranking elastic net model coefficients for the linear case in accordance with (Balliu et al., 2024) and by ranking random forest predictors through SHAP values on the test set for the non-linear case in accordance with (Price et al., 2023).

Evaluation

To answer RQ1 (“In how many of the enrolled patients with clinically diagnosed depression can analysis of their individual, up to one year passive time series data (variables described in section Data Collection) identify objective markers that estimate the severity of depressive symptomatology as measured by daily PHQ-2 scores?”), the proportion of patients whose linear models exceeded predefined clinical thresholds (for example R^2 greater than 0.30 or MAE below a symptom-scale point) was reported.

To answer RQ2 (“For patients in whom a significant linear prediction is achieved, which sensor or app derived data source contributes most to the explanation of individual depression severity?”), feature importance was assessed by aggregating results across all elastic net models. For each model, features with non-zero coefficients were identified, and the frequency of occurrence across participants was used to determine the most consistently contributing data sources.

To answer RQ3 (“How does the application of non-linear ensemble methods influence the results of research questions 1 and 2? I.e., how many additional patients exhibit a reliable predictive pattern, and do the dominant data sources shift when non-linear relationships are allowed?”), the proportion of patients whose non-linear models (random forests) exceeded predefined clinical thresholds was computed and compared to the results from RQ1. Feature importance was again assessed by aggregating non-zero feature contributions across all models. The most consistently contributing features from the non-linear models were then compared to those identified in RQ2 to examine shifts in dominant data sources when non-linear modeling is applied.

Study Protocol Deviations

Despite the study protocol's emphasis on capturing multi-level associations (Reich et al., 2025), we employed linear models in form of elastic nets separately for each participant to construct truly idiographic models. While multilevel models are well-suited for analyzing data with nested structures and can estimate both within- and between-person effects, our analytical focus lies exclusively on within-person relationships. As such, we opted not to implement multilevel approaches that inherently integrate across individuals in this work. Moreover, when applied to data from a single participant, multilevel models effectively reduce to standard linear regression, offering no advantage over simpler or more robust alternatives. Instead, elastic net models allow to estimate associations uniquely for each participant without borrowing strength from group-level patterns. The combination of L1 (Lasso) and L2 (Ridge) regularization in the elastic net enables effective feature selection while retaining correlated covariates. This makes the method particularly suitable for our purposes, as the models are not used for prediction, but rather to identify objective markers of depression based on feature importance.

Additionally, deviating from the use of Kernel SHAP as described in the study protocol, we employed Tree SHAP, which is specifically optimized for tree-based models such as Random Forests. Unlike the model-agnostic Kernel SHAP, which can be computationally intensive and less accurate for complex models, Tree SHAP leverages the internal structure of tree ensembles to provide more efficient and precise feature attributions.

Acknowledgements

The authors would like to express gratitude and acknowledge the contributions of all members of the MONDY (Secure and open platform for AI-based healthcare apps) Consortium: Yvonne Weber, Elisabeth Schriewer, Mohamed Alhaskir, Stefan Wolking, Florian Fischer, Rebeka Amin, Jil Zippelius, Tobias Dunker, Kismet Ekinci, Angela Carell, Enrico Lohmann.

References

Aledavood, T., Luong, N., Baryshnikov, I., Darst, R., Heikkilä, R., Holmén, J., Ikäheimonen, A., Martikkala, A., Riihimäki, K., Saleva, O., Triana, A. M., & Isometsä, E. (2025). Multimodal Digital Phenotyping Study in Patients With Major Depressive Episodes and Healthy Controls (Mobile Monitoring of Mood): Observational Longitudinal Study. *JMIR Mental Health*, *12*(1), e63622. https://doi.org/10.2196/63622

Arrieta, J., Aguerrebere, M., Raviola, G., Flores, H., Elliott, P., Espinosa, A., Reyes, A., Ortiz-Panozo, E., Rodriguez-Gutierrez, E. G., Mukherjee, J., Palazuelos, D., & Franke, M. F. (2017). 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. *Journal of Clinical Psychology*, *73*(9), Article 9. https://doi.org/10.1002/jclp.22390

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research*, *20*(1), 40–49. https://doi.org/10.1002/mpr.329

Balliu, B., Douglas, C., Seok, D., Shenhav, L., Wu, Y., Chatzopoulou, D., Kaiser, W., Chen, V., Kim, J., Deverasetty, S., Arnaudova, I., Gibbons, R., Congdon, E., Craske, M. G., Freimer, N., Halperin, E., Sankararaman, S., & Flint, J. (2024). Personalized mood prediction from patterns of behavior collected with smartphones. *Npj Digital Medicine*, *7*(1), 1–14. https://doi.org/10.1038/s41746-024-01035-6

Bizzozero-Peroni, B., Díaz-Goñi, V., Jiménez-López, E., Rodríguez-Gutiérrez, E., Sequí-Domínguez, I., Núñez de Arenas-Arroyo, S., López-Gil, J. F., Martínez-Vizcaíno, V., & Mesas, A. E. (2024). Daily Step Count and Depression in Adults: A Systematic Review and Meta-Analysis. *JAMA Network Open*, *7*(12), e2451208. https://doi.org/10.1001/jamanetworkopen.2024.51208

Brunner, C., & Hofer, F. (2023). SleepECG: A Python package for sleep staging based on heart rate. *Journal of Open Source Software*, *8*(86), 5411. https://doi.org/10.21105/joss.05411

Carney, R. M., & Freedland, K. E. (2009). Depression and heart rate variability in patients with coronary heart disease. *Cleveland Clinic Journal of Medicine*, *76*(4 suppl 2), S13–S17. https://doi.org/10.3949/ccjm.76.s2.03

Choi, A., Ooi, A., & Lottridge, D. (2024). Digital Phenotyping for Stress, Anxiety, and Mild Depression: Systematic Literature Review. *JMIR mHealth and uHealth*, *12*, e40689. https://doi.org/10.2196/40689

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, *71*, 10–49. https://doi.org/10.1016/j.specom.2015.03.004

Cummins, N., Sethu, V., Epps, J., Schnieder, S., & Krajewski, J. (2015). Analysis of acoustic space variability in speech affected by depression. *Speech Communication*, *75*, 27–49. https://doi.org/10.1016/j.specom.2015.09.003

da Estrela, C., McGrath, J., Booij, L., & Gouin, J.-P. (2021). Heart Rate Variability, Sleep Quality, and Depression in the Context of Chronic Stress. *Annals of Behavioral Medicine: A Publication of the Society of Behavioral Medicine*, *55*(2), 155–164. https://doi.org/10.1093/abm/kaaa039

de Vries, H., Kamphuis, W., van der Schans, C., Sanderman, R., & Oldenhuis, H. (2022). Trends in Daily Heart Rate Variability Fluctuations Are Associated with Longitudinal Changes in Stress and Somatisation in Police Officers. *Healthcare*, *10*(1), 144. https://doi.org/10.3390/healthcare10010144

Dhelim, S., Chen, L., Ning, H., & Nugent, C. (2022). Artificial intelligence for suicide assessment using Audiovisual Cues: A review. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-022-10290-6

Fonseca, P., Ross, M., Cerny, A., Anderer, P., van Meulen, F., Janssen, H., Pijpers, A., Dujardin, S., van Hirtum, P., van Gilst, M., & Overeem, S. (2023). A computationally efficient algorithm for wearable sleep staging in clinical populations. *Scientific Reports*, *13*(1), 9182. https://doi.org/10.1038/s41598-023-36444-2

Kemp, A. H., Quintana, D. S., Gray, M. A., Felmingham, K. L., Brown, K., & Gatt, J. M. (2010). Impact of Depression and Antidepressant Treatment on Heart Rate Variability: A Review and Meta-Analysis. *Biological Psychiatry*, *67*(11), 1067–1074. https://doi.org/10.1016/j.biopsych.2009.12.012

Kroenke, K., & Spitzer, R. L. (2002). The PHQ-9: A new depression diagnostic and severity measure. *Psychiatric Annals*, *32*(9), Article 9. https://doi.org/10.3928/0048-5713-20020901-06

Kumar, C., Sakshi, P., Sinha, N., Sunita, & Kumar, T. (2024). HRV changes in young adults with depression. *Journal of Family Medicine and Primary Care*, *13*(7), 2585–2588. https://doi.org/10.4103/jfmpc.jfmpc_926_23

Kupferberg, A., Bicks, L., & Hasler, G. (2016). Social functioning in major depressive disorder. *Neuroscience & Biobehavioral Reviews*, *69*, 313–332. https://doi.org/10.1016/j.neubiorev.2016.07.002

Lim, J.-A., Yun, J.-Y., Choi, Y., Choi, S.-H., Kwon, Y., Lee, H. Y., & Jang, J. H. (2020). Sex-Specific Differences in Severity of Depressive Symptoms, Heart Rate Variability, and Neurocognitive Profiles of Depressed Young Adults: Exploring Characteristics for Mild Depression. *Frontiers in Psychiatry*, *11*, 217. https://doi.org/10.3389/fpsyt.2020.00217

Lin, W., Karahanoglu, F. I., Demanuele, C., Khan, S., Cai, X., Santamaria, M., Di, J., & Adamowicz, L. (2023). SciKit digital health package for accelerometry-measured physical activity: Comparisons to existing solutions and investigations of age effects in healthy adults. *Frontiers in Digital Health*, *5*, 1321086. https://doi.org/10.3389/fdgth.2023.1321086

Lorenz, N., Sander, C., Ivanova, G., & Hegerl, U. (2020). Temporal Associations of Daily Changes in Sleep and Depression Core Symptoms in Patients Suffering From Major Depressive Disorder: Idiographic Time-Series Analysis. *JMIR Mental Health*, *7*(4), e17071. https://doi.org/10.2196/17071

Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lespinasse, F., Pham, H., Schölzel, C., & Chen, S. H. A. (2021). NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behavior Research Methods*, *53*(4), 1689–1696. https://doi.org/10.3758/s13428-020-01516-y

Migueles, J. H., Cadenas-Sanchez, C., Rowlands, A. V., Henriksson, P., Shiroma, E. J., Acosta, F. M., Rodriguez-Ayllon, M., Esteban-Cornejo, I., Plaza-Florido, A., Gil-Cosano, J. J., Ekelund, U., van Hees, V. T., & Ortega, F. B. (2019). Comparability of accelerometer signal aggregation metrics across placements and dominant wrist cut points for the assessment of physical activity in adults. *Scientific Reports*, *9*(1), 18235. https://doi.org/10.1038/s41598-019-54267-y

Moore, E., Clements, M. A., Peifer, J. W., & Weisser, L. (2008). Critical analysis of the impact of glottal features in the classification of clinical depression in speech. *IEEE Transactions on Bio-Medical Engineering*, *55*(1), 96–107. https://doi.org/10.1109/TBME.2007.900562

Nie, Y., Ren, R., Yang, L., Shi, Y., Sanford, L. D., Zhang, Y., & Tang, X. (2024). Polysomnographic changes of obsessive-compulsive disorder: Evidence from case-control studies. *Sleep & Breathing = Schlaf & Atmung*, *29*(1), 21. https://doi.org/10.1007/s11325-024-03209-8

O'brien, R. M. (2007). A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Quality & Quantity*, *41*(5), 673–690. https://doi.org/10.1007/s11135-006-9018-6

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, *12*(null), 2825–2830.

Price, G. D., Heinz, M. V., Song, S. H., Nemesure, M. D., & Jacobson, N. C. (2023). 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. *Translational Psychiatry*, *13*(1), 1–10. https://doi.org/10.1038/s41398-023-02669-y

Quatieri, T. F., & Malyska, N. (2012). Vocal-source biomarkers for depression: A link to psychomotor activity. *Interspeech 2012*, 1059–1062. https://doi.org/10.21437/Interspeech.2012-311

Radha, M., Fonseca, P., Moreau, A., Ross, M., Cerny, A., Anderer, P., Long, X., & Aarts, R. M. (2019). Sleep stage classification from heart-rate variability using long short-term memory neural networks. *Scientific Reports*, *9*(1), Article 1. https://doi.org/10.1038/s41598-019-49703-y

Reich, H., Schreynemackers, S., Amin, R., Ludwig, S., Zippelius, J., Leimhofer, J., Dunker, T., Schriewer, E., Carell, A., Weber, Y., & Hegerl, U. (2025). 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. *Contemporary Clinical Trials Communications*, *45*, 101492. https://doi.org/10.1016/j.conctc.2025.101492

Saeb, S., Zhang, M., Karr, C. J., Schueller, S. M., Corden, M. E., Kording, K. P., & Mohr, D. C. (2015). Mobile Phone Sensor Correlates of Depressive Symptom Severity in Daily-Life Behavior: An Exploratory Study. *Journal of Medical Internet Research*, *17*(7), e175. https://doi.org/10.2196/jmir.4273

Siepe, B. (2022). *Temporal dynamics of depressive symptomatology: An idiographic time series analysis* [Master thesis]. Johann Wolfgang Goethe-Universität Frankfurt.

Singh Solorzano, C., Violani, C., & Grano, C. (2022). Pre-partum HRV as a predictor of postpartum depression: The potential use of a smartphone application for physiological recordings. *Journal of Affective Disorders*, *319*, 172–180. https://doi.org/10.1016/j.jad.2022.09.056

Sridhar, N., Shoeb, A., Stephens, P., Kharbouch, A., Shimol, D. B., Burkart, J., Ghoreyshi, A., & Myers, L. (2020). Deep learning for automated sleep staging using instantaneous heart rate. *Npj Digital Medicine*, *3*(1), 1–10. https://doi.org/10.1038/s41746-020-0291-x

Stamatis, C. A., Meyerhoff, J., Meng, Y., Lin, Z. C. C., Cho, Y. M., Liu, T., Karr, C. J., Liu, T., Curtis, B. L., Ungar, L. H., & Mohr, D. C. (2024). Differential temporal utility of passively sensed smartphone features for depression and anxiety symptom prediction: A longitudinal cohort study. *Npj Mental Health Research*, *3*(1), 1–8. https://doi.org/10.1038/s44184-023-00041-y

Stasak, B., Epps, J., Schatten, H. T., Miller, I. W., Provost, E. M., & Armey, M. F. (2021). Read speech voice quality and disfluency in individuals with recent suicidal ideation or suicide attempt. *Speech Communication*, *132*, 10–20. https://doi.org/10.1016/j.specom.2021.05.004

Sun, S., Folarin, A. A., Zhang, Y., Cummins, N., Garcia-Dias, R., Stewart, C., Ranjan, Y., Rashid, Z., Conde, P., Laiou, P., Sankesara, H., Matcham, F., Leightley, D., White, K. M.,

Oetzmann, C., Ivan, A., Lamers, F., Siddi, S., Simblett, S., … Consortium, R.-C. (2023). Challenges in Using mHealth Data From Smartphones and Wearable Devices to Predict Depression Symptom Severity: Retrospective Analysis. *Journal of Medical Internet Research*, *25*(1), e45233. https://doi.org/10.2196/45233

Tian, Y., Zhou, K., & Pelleg, D. (2022). What and How long: Prediction of Mobile App Engagement. *ACM Transactions on Information Systems*, *40*(1), 1–38. https://doi.org/10.1145/3464301

Wang, X., Pathiravasan, C. H., Zhang, Y., Trinquart, L., Borrelli, B., Spartano, N. L., Lin, H., Nowak, C., Kheterpal, V., Benjamin, E. J., McManus, D. D., Murabito, J. M., & Liu, C. (2023). Association of Depressive Symptom Trajectory With Physical Activity Collected by mHealth Devices in the Electronic Framingham Heart Study: Cohort Study. *JMIR Mental Health*, *10*(1), e44529. https://doi.org/10.2196/44529

White, T., Westgate, K., Hollidge, S., Venables, M., Olivier, P., Wareham, N., & Brage, S. (2019). Estimating energy expenditure from wrist and thigh accelerometry in free-living adults: A doubly labelled water study. *International Journal of Obesity*, *43*(11), 2333–2342. https://doi.org/10.1038/s41366-019-0352-x

Yu, R., Li, Y., Zhao, K., & Fan, F. (2025). A review of automatic sleep stage classification using machine learning algorithms based on heart rate variability. *Sleep and Biological Rhythms*, *23*(2), 113–125. https://doi.org/10.1007/s41105-024-00563-8

Yue, C., Ware, S., Morillo, R., Lu, J., Shang, C., Bi, J., Kamath, J., Russell, A., Bamis, A., & Wang, B. (2020). Automatic depression prediction using Internet traffic characteristics on smartphones. *Smart Health*, *18*, 100137. https://doi.org/10.1016/j.smhl.2020.100137