Journal of Affective Disorders 394 (2026) 120426

Contents lists available at ScienceDirect

Journal of Affective Disorders

journal homepage: www.elsevier.com/locate/jad

Research paper

Sleep disturbance recorded via wearable sensors predicts depression severity 9 years later

Nur Hani Zainal [a,*], Peter F. Hitchcock [b]

[a] National University of Singapore, Department of Psychology, Kent Ridge Campus, Singapore
[b] Emory University, Department of Psychology, Atlanta, GA, USA

ARTICLE INFO

Keywords:
Depression
Electrocardiogram
Longitudinal
Machine learning
Passive sensors
Sleep

ABSTRACT

Background: Major depressive disorder (MDD) is prevalent and poses major public health implications. Autonomic nervous system (ANS) dysregulation and sleep disturbances are theorized to be distal risk factors. However, previous research has depended on cross-sectional designs, small predictor sets, and suboptimal methods, limiting temporal inference and predictive accuracy. We thus capitalized on machine learning to identify physiology and sleep predictors of nine-year MDD symptoms.
Method: Community adults ($N = 1054$) participated in a study that included baseline physiological electrocardiogram (ECG) and sleep actigraphy wearable assessments. Clinical interviews were administered to assess for psychiatric symptoms at baseline and nine-year follow-up. Eight ML models were trained to predict MDD severity using 80 baseline variables via a 70–30 train-test split with 5-fold cross-validation with 81 baseline variables to predict MDD severity.
Results: The best model (gradient boosting machine) had 10 variables with strong predictive accuracy in the test set ($R^2 = 19.8\%$). Baseline MDD, generalized anxiety, and panic disorder symptoms strongly predicted nine-year MDD severity. Longer total sleep time, lower sleep efficiency, and higher average wake time during sleep phases were key correlates of higher nine-year MDD severity. Other correlates included fewer average sleep bouts and shorter wake times during active phases, as well as nonlinear patterns of wake time length and percentage during rest phases. Physiology ECG variables had limited incremental predictive value.
Conclusions: Wearable actigraphy-indexed sleep disturbances predicted long-term MDD symptoms beyond baseline severity and ANS dysregulation indices. Combining passive sleep sensors into routine assessments might optimize MDD prevention and treatment.

1. Introduction

Major depressive disorder (MDD) symptoms, such as concentration issues, fatigue, motivational deficits, and sleep disturbances, are common in the general population (American Psychiatric Association, 2022; Zainal and Newman, 2021). Meta-analyses of epidemiological reports indicated that 12-month prevalence estimates of elevated MDD symptoms ranged from 19% to 42% among adults (Moreno-Agostino et al., 2021) and youths (Shorey et al., 2022) globally in the general population. Persistent MDD symptoms interfere with various social, school, and work functions, escalating to poorer quality of life across all developmental stages (Hohls et al., 2021; Sivertsen et al., 2015). Thus, identifying distal risk factors for future MDD symptoms is critical for prevention and treatment.

One plausible set of risk factors is the presence of autonomic nervous system (ANS) dysregulation markers, such as low heart rate variability (HRV) and a high resting heart rate. ANS dysregulation refers to a lack of balance between strong stress-triggering sympathetic activity and weak calmness-inducing parasympathetic activity (Sameroff, 2020; Sgoifo et al., 2015). According to the transactional model of stress and depression, such ANS dysregulation might reduce adaptability to changing environmental demands, thereby precipitating and perpetuating MDD symptoms over long durations. Relatedly, ANS imbalance is intimately related to emotion dysregulation, such as low cognitive flexibility, which may also predict later MDD symptoms (Fantini-Hauwel et al., 2020). Indeed, early case-control, cross-sectional studies

[*] Corresponding author at: National University of Singapore (NUS), Department of Psychology, Faculty of Arts and Social Sciences, Block AS4, Office #03-25, 9 Arts Link, 117570, Singapore.
E-mail address: hanizainal@nus.edu.sg (N.H. Zainal).

https://doi.org/10.1016/j.jad.2025.120426
Received 31 May 2025; Received in revised form 10 September 2025; Accepted 10 October 2025
Available online 15 October 2025

found that patients with MDD consistently exhibited lower levels of diverse HRV indicators than healthy controls, with small to moderate effect sizes (Hedge's $g = -0.462$ to -0.096; cf. meta-analysis by Koch et al., 2019). Moreover, longitudinal reports showed that higher resting heart rate and lower HRV indices predicted higher future incident MDD rates (Jandackova et al., 2016), perseverative cognitions, and MDD symptoms in both community adults and psychiatric samples (Carnevali et al., 2018; Gentili et al., 2017), suggesting a potential etiological role of ANS imbalance in MDD.

Markers of sleep dysregulation may also serve as important distal risk factors for MDD symptoms through diverse plausible mechanisms. Sleep disturbances might adversely affect cognitive and emotional regulation, which could fuel MDD symptoms (Palmer and Alfano, 2017). Other possible key pathways implicate perturbations in the hypothalamic-pituitary-adrenal (HPA) axis and circadian rhythms, which govern sleep-wake cycles (Asarnow, 2020). Suboptimal changes in levels of catecholamine neurotransmitters in the suprachiasmatic nucleus of the hypothalamus and brain areas that regulate physical activity, mood, and sleep-wake patterns could trigger MDD symptoms over time (Grippo and Johnson, 2009; Thase, 2006).

Empirical data support the notion that sleep disturbances precipitate MDD symptoms. For instance, data from seven studies showed that myriad sleep-problem indices preceded MDD symptoms in adolescents (refer to meta-analysis by Lovato and Gradisar, 2014). Other meta-analyses showed that hypersomnia (excessive sleep), insomnia (inadequate sleep), and related sleep disturbances predicted greater future MDD severity (Zhai et al., 2015), including suicide ideation (Liu et al., 2020). Conversely, four empirical studies have indicated that treating insomnia substantially decreases subsequent MDD prevalence rates in adults (cf. systematic review by Boland et al., 2023).

The present study leveraged a suite of precision medicine approaches to investigate how physiological ECG and sleep wearable passive sensor variables predicted MDD severity after a nine-year follow-up (Iglesias et al., 2025). First, we applied machine learning (ML) algorithms to detect complex, nonlinear associations among high-dimensional baseline variables, expanding on previous research that depended on ordinary least squares (OLS) regression. Compared to OLS, ML methods model main effects and interactions more flexibly. ML approaches further optimally manage the bias-variance trade-off via internal validation processes, such as nested cross-validation (Lewis et al., 2023; Yarkoni and Westfall, 2017). Second, unlike most prior prognostic ML studies, which only provided discrimination metrics to predict the presence of a binary clinical outcome, the present study conducted calibration analysis to test the degree to which predicted scores matched actual continuous MDD severity outcomes (Huang et al., 2020). Furthermore, examining dimensional severity outcomes, rather than categorical clinical endpoints, is more consistent with the Research Domain Criteria framework (Morris et al., 2022), as it captures a broader range of functioning or lack thereof (Kelly et al., 2018). Third, in contrast to previous research limited by small sample sizes and predictor sets (Luedtke et al., 2019), we utilized a comprehensive set of physiological and sleep predictors, including actigraphy-derived measures of physical activity and sleep, as well as ECG, in a well-powered sample. Our approach thus enabled a data-driven, stringent evaluation of baseline variables with the strongest correlations with nine-year MDD severity, exploiting low-burden passive sleep sensors that overcome recall biases in subjective sleep reports (Massar et al., 2021). Fourth, most studies on this topic have been cross-sectional (e.g., Blood et al., 2015), which limits the ability to draw causal conclusions (Pearl, 2014). In contrast, here, we provide a rare test of how our predictors forecast MDD symptoms nine years later. Fifth, we tested the ECG and sleep predictors coupled with baseline comorbid depression and anxiety symptoms, thereby providing a strong test of the independent prognostic contribution of these objective measures over and above symptoms.

In summary, we employed ML approaches to identify the multivariable predictors of nine-year MDD symptoms using a large baseline predictor set with diverse physiological ECG and sleep actigraphy variables. Our hypotheses were twofold. First, we expected the optimal model to perform well, serving as a prerequisite for interpreting complex multivariable predictor patterns. Optimal performance was defined as an *R*-squared (R^2) value of ≥15.0% (Gupta et al., 2024), indicating that the predictors accounted for a meaningful proportion of variance in the nine-year MDD severity outcome. Second, we anticipated that theory-driven ECG physiology and sleep actigraphy variables at baseline would predict higher nine-year MDD severity.

2. Method

2.1. Participants

Participants ($N = 1054$) took part in the Midlife Development in the United States (MIDUS) Biomarker project (Love et al., 2010) as well as the MIDUS survey studies at baseline (Wave 1; W1; 2004 to 2006; Ryff et al., 2021) and follow-up (Wave 2; W2; 2013 to 2014; Ryff et al., 2019). They visited one of three data collection sites (Los Angeles, California; Madison, Wisconsin; Washington, D.C.). The mean age was 58.04 years ($SD = 11.62$, range = 35 to 86). Regarding sex, 45.3% (477/1054) were men, and the remaining 54.7% (577/1054) were women. With respect to racial identity, 91.2% (961/1054) identified as White, and the remaining 8.8% (93/1054) identified as Asian, African American, Pacific Islander, or Native American. Concerning education, 44.1% (465/1054) had a college education and above, 28.5% (300/1054) had some college education, 22.6% (238/1054) had a high school diploma, and the remaining 4.8% (51/1054) had no high school education. Table 1 details the sociodemographic and clinical variables separately for the training and test sets.

2.2. Procedures

At W1, participants completed a brief clinical interview and a series of surveys administered by MIDUS research staff to assess the presence and severity of past 12-month symptoms of MDD, generalized anxiety disorder (GAD), and panic disorder (PD). They also underwent ECG physiology and sleep actigraphy protocols at W1 (Laborde et al., 2017;

Table 1
Descriptive data of sociodemographic and clinical variables at W1 ($N = 1054$).

	Train set ($n = 738$)		Test set ($n = 316$)	
Continuous variables	*M*	(*SD*)	*M*	(*SD*)
Age	57.86	(11.44)	58.47	(12.06)
Baseline MDD severity	0.57	(1.53)	0.65	(1.64)
Baseline GAD severity	14.62	(8.67)	14.73	(8.41)
Baseline PD severity	0.46	(1.23)	0.36	(1.06)
9-Year MDD severity	0.49	(1.43)	0.55	(1.57)
Categorical variables	*n*	(%)	*n*	(%)
Sex				
Men	326	(44.17)	151	(47.78)
Women	412	(55.83)	165	(52.22)
Racial identity				
Declined to disclose	21	(2.85)	9	(2.85)
Multiracial	6	(0.81)	2	(0.63)
White	675	(91.46)	286	(90.51)
African American	21	(2.85)	8	(2.53)
Native American	1	(0.14)	4	(1.27)
Asian	1	(0.14)	1	(0.32)
Other	13	(1.76)	6	(1.90)
Education				
College education and above	318	(43.09)	147	(46.52)
High school	171	(23.17)	67	(21.20)
No high school degree	32	(4.34)	19	(6.01)
Some college education	217	(29.40)	83	(26.27)

Note. MDD, major depressive disorder; W1, wave 1 (2004–2006); GAD, generalized anxiety disorder; PD, panic disorder. The MDD symptom scales excluded sleep disturbance items.

Lee et al., 2025). At W2, they completed the same brief clinical interview. Given the research aims, only data from participants who fulfilled these study procedures were used.

2.3. Measures

2.3.1. W1 and W2 MDD symptoms

MIDUS researchers administered the Composite International Diagnostic Interview-Short Form (CIDI-SF), which was aligned with the Diagnostic and Statistical Manual of Mental Disorders, Third Edition-Revised (DSM-III-R; Kessler et al., 1998a; Kessler and Üstün, 2004; Kessler et al., 1998b). Participants responded if they experienced appetite problems, depressed mood, difficulties focusing, fatigue, motivational deficits, sleep disturbances, sense of worthlessness, and suicide ideation in the past 12 months (rated as the *presence* [1] or *absence* [0] of symptoms). A total sum of all item scores indicated the degree of MDD severity, ranging from 0 (*lowest*) to 7 (*highest*). Prior studies have shown that the continuous version of this scale had good internal consistency and strong construct validity (Zainal and Newman, 2022a, 2022b).

2.3.2. W1 GAD symptoms

The DSM-III-R-concordant CIDI-SF interview assessed 10 GAD symptoms related to worry over the past 12 months (Kessler et al., 1998a; Kessler and Üstün, 2004; Kessler et al., 1998b): concentration issues, fatigue, feeling keyed up, irritability, mind going blank, muscle tension, restlessness, and sleep difficulties. Symptoms were rated on a 4-point scale (1 = *never* to 4 = *a lot more*) and summed to yield a total score (range: 10 to 40). Prior work demonstrated the dimensional CIDI-SF GAD scale's good internal consistency and excellent construct validity (Ng et al., 2024).

2.3.3. W1 PD symptoms

The DSM-III-R-consistent CIDI-SF interview was administered to measure six PD symptoms (recorded as *present* [1] or *absent* [0]; Kessler et al., 1998a; Kessler and Üstün, 2004; Kessler et al., 1998b): chest or stomach discomfort, heart racing, hot flashes or chills, sense of unreality, sweating, and trembling or shaking. A total score was calculated by summing all the item scores (theoretical scores ranged from 0 to 6). Previous studies evidenced good internal consistency and excellent construct validity of the CIDI-SF PD severity scale (Zainal and Newman, 2022a, 2022b).

2.3.4. W1 sleep actigraphy

Participants were instructed to wear a sleep actigraphy smartwatch (Philips Corporation; Amsterdam, The Netherlands; Andover, MA, USA) while filling out a sleep diary for seven consecutive days and nights. Participants were asked to start the actigraphy data collection on Tuesday morning after returning from the data collection site and end this part of the protocol the following Tuesday morning. The Actiwatch identified movement counts at 30-s periods in wakeful, resting, and sleeping phases by contrasting the computed total activity counts to a wake threshold score of 40 (Aqua et al., 2024). During times when the total activity counts were equal to or less than the wake threshold score, the period was identified as the sleep phase. Rest phases were identified through sleep diaries first, or via event indices and adjacent data second, if diary data were missing. The actigraphy marked wake and sleep times based on self-reported sleep diary records. Missing data, which occurred for various reasons (e.g., premature removal, misremembering to wear the smartwatch, unexpected night shifts, and traveling across time zones), was reviewed. Such actigraphy periods were identified and removed per recommended practices (Brindle et al., 2019). Other passive sensing sleep indices captured by the actigraphy were sleep bouts (counts), sleep efficiency (%), sleep onset latency (SOL), sleep time, total sleep time (TST), wake after sleep onset (WASO), and wake time (Crowley et al., 2018; Yip et al., 2021). These values were aggregated by computing the mean across the days with valid data. Despite their high collinearity, we added these passive sleep wearable markers in the same predictor set to model their unique contributions in multivariable models. Previous studies have shown that specific predictors may distinctly predict MDD severity outcomes, even among correlated sleep indices, highlighting the need to test their independent predictive utility (Shrivastava et al., 2014; Yan et al., 2022).

2.3.5. Operational definitions of sleep indices

Rest period was defined as the self-reported or actigraphy-captured time during the evening or nighttime when a participant wound down from activities before going to bed. TST was indexed as the sum of sleep-scored epochs measured by Actiware during the rest period. WASO was defined as the sum of wake-scored epochs from sleep onset to final awakening. SOL was defined as the epoch-based interval between bedtime and the onset of sleep. Sleep efficiency was measured as sleep-period epoch length divided by the rest period duration, expressed as a percentage. Activity counts were marked as the number of actigraphy activity counts (arbitrary units) captured by the actigraphy. For all activity counts, their average, minimum, and maximum values were recorded during each 30-s epoch in active, rest, and sleep phases.

W1 physiology ECG

Following a caffeine-free light breakfast, participants had ECG electrodes placed beneath each clavicle (left and right shoulders) and their lower left abdomens. Further, respiratory effort bands were fastened encircling their abdomen and chest to track breathing patterns. While sitting, their dominant hand was placed on a keypad to facilitate computer-administered stressor tasks, a mental arithmetic task, a working memory test (Paced Auditory Serial Addition Test; PASAT; Spreen and Strauss, 1998), and an inhibitory control test (Stroop, 1935), presented in a counterbalanced manner to prevent order effects (Kimhy et al., 2013). After brief practice and calibration phases to gather high-quality signals (25 to 30 min), baseline functioning was assessed for 11 min. Subsequently, the first stressor test was administered, followed by a recovery interval, and then the second stressor test, with another recovery interval in between. This was followed by the final recovery phase, which lasted approximately 6 min.

HF-HRV was used to assess cardiac vagal control. Following recommendations, a National Instruments Analog-to-Digital (AD) board digitized analog signals at 500 Hz, and this data was transferred to and gathered by a microcomputer (Berntson et al., 1997). RR interval series were generated by processing the ECG waveforms using a registered software that identified physiological events, such as R-waves, on a routine basis (Allen et al., 2007). Mistakes in identifying R-waves were rectified in accordance with best practices (Laborde et al., 2017). Natural log transformation was applied to these physiological indices. Abdominal and chest respiratory rate signals were processed using another registered software, which generated average scores of respiratory rates on a minute-to-minute basis.

Reliable and stable HRV estimates were derived during each period of rest and recovery by following a highly recommended and well-established protocol in psychophysiology research (Quintana et al., 2016). Mean HF-HRV and LF-HRV values were calculated in 5 to 10-min intervals during the rest phase, both stressor phases, and their related recovery phases (Gruenewald et al., 2023). Subtracting the scores between the recovery and stressor task periods created a vagal recovery score. Given that cardiac vagal control reduces in response to stressors and increases in recovery phases (Shaffer and Ginsberg, 2017), a higher vagal recovery score indicated greater increase in the post-stressor phase HF-HRV.

Note that in addition to HF-HRV and LF-HRV, the MIDUS project also captured and computed the standard deviation of the RR intervals (SDRR) and root mean square of successive differences (RMSSD). The natural log of these variables was also computed. Both SDRR and RMSSD reflected the overall beat-to-beat variations and short-term parasympathetic activity, respectively (Shaffer and Ginsberg, 2017), supplementing the frequency-dimension ECG markers.

2.4. Data analyses

A 70–30 train-test split stratified by W2 MDD diagnostic status yielded 739 participants in the train set and 316 in the test set. The splitting approach was stratified to ensure equal percentages of MDD diagnoses across both datasets, thereby preserving concordance in multivariable data distribution for model building, training, and testing (James et al., 2013). Relatedly, our total sample size aligned with best practices that propose the number of data points should be a minimum of 10 to 15 times that of the predictor variables to facilitate stability in parameter estimation and minimize overfitting (problems pertaining to high variance; Goldenholz et al., 2023; Rajput et al., 2023). As we had 80 baseline predictors, our sample size aligned with this standard (Wisz et al., 2008).

All analyses were conducted in *R* (R Core Team, 2025). All initial preprocessing steps, including random forest imputation and data transformations, were conducted separately in the train and test sets to avoid data leakage. Missing data, initially 47% of the predictor data set and 0% of the outcome variable, were managed using random forest imputation with the *missRanger* package (Mayer, 2024). This approach outperforms standard multiple imputation by permitting nonlinear relations and interactions, as well as accommodating various data types (Shah et al., 2014), and can optimally handle large amounts of missing data (Lee and Shi, 2021). These factors render the use of random forest imputation robust even with high levels of missingness in the predictor set, including under the missing not at random (MNAR) assumption (Afkanpour et al., 2024; Tang and Ishwaran, 2017). Data normalization was done on continuous and integer predictors. Near-zero variance predictors were removed. One-hot encoding was conducted on categorical variables (James et al., 2013). Further, sensitivity analyses were conducted by assessing the performance metrics of each model described in the next section while combining both random forest imputation and inverse probability weights (IPWs) based on completer status. These measures were conducted to address possible biases arising from attrition, MNAR missing pattern, and confounders (Daza et al., 2017). Tables S1 to S8 in the online supplemental materials (OSM) summarize the descriptives of all baseline ECG and sleep actigraphy variables that served as predictors in the train and test sets post-preprocessing.

Eight multivariable ML models predicting W2 MDD severity were tested (refer to page 14 of the OSM for details on each ML algorithm). The model with the most optimal performance was the gradient boosting machine (GBM; Schroeders et al., 2022). The GBM utilized a 5-fold cross-validation on the 80% training subset to choose hyperparameters (Kovač et al., 2024). The hyperparameters (embedded in *italics* in this paragraph) were specified as 500, 1000, or 1500 trees (*n.trees*), a learning rate of 0.01, 0.05, or 0.1 (*shrinkage*), and an interaction depth of 3, 5, or 7 (*interaction.depth*) (James et al., 2013). The stopping rule was specified to continue tree splitting until a maximum of 10 or 20 predictors per terminal node (*n.minobsinnode*). The subsampling fraction (*bag.fraction*), which represents the portion of the data in the train set randomly selected to produce the next tree in the expansion, was 0.5. While tuning, 80% of the training data were utilized to fit different ML algorithms, and the remaining 20% for internal validation (*train.fraction* = 0.8); following this, the final model was fit on the entire training set (*train.fraction* = 1.0). All available data in the training set were used for model training, with no independent training data segment set aside to estimate the out-of-sample loss function (*train.fraction* = 1). Essential performance metrics, namely the root mean squared error (RMSE), mean absolute error (MAE), and *R*-squared (R^2), were calculated (James et al., 2013). Lower RMSE and MAE values, as well as higher R^2 values, indicated better model performance. We used an elastic net regularization filter on the training set to choose the top 20 W1 variables (Zou and Hastie, 2005), and the held-out test set evaluated the final GBM trained on these chosen variables. To quantify uncertainty, we computed the 95% confidence intervals (CIs) of each model's performance metric, facilitating more robust comparisons.

Moreover, the relative importance of the top 10 W1 predictors was analyzed using Shapley additive explanations (SHAP) bee swarm plots, an interpretable ML method (Lundberg and Lee, 2017; Molnar, 2022), using the 316 held-out test set observations. The SHAP bee swarm plot has a rich, intuitive appeal that helps readers understand the contribution of each predictor to the model output at the participant level. Each dot in the density plot for each predictor indicates a participant-level SHAP value for a specific predictor, and the *x*-axis position signifies the direction and magnitude of that predictor's impact on the predicted W2 MDD severity score (Lundberg et al., 2020). Positive SHAP values indicate that the unique predictor increases the model's output (i.e., greater level predicted higher W2 MDD severity), whereas negative values suggest a decreasing impact. The color scheme relays the original predictor value (with red reflecting larger values and blue smaller ones), permitting readers to recognize how the raw predictor values are associated with their impact on the outcome (Kovač et al., 2024). The vertical distribution of dots for each predictor indicates heterogeneity in its effect across participants, possibly due to nonlinearities and interactions with other predictors. In addition, we generated the partial dependence plots, which constituted the bases of the SHAP values, to display the marginal effect of each W1 variable on the predicted W2 MDD severity score and to increase the intuitive appeal of SHAP outcomes (Kerrigan et al., 2025). Together, the SHAP bee swarm plot and partial dependence plots facilitate the interpretation of both overall predictor importance and participant-level variability in predictor influence.

Importantly, we prioritized predicting W2 MDD severity, excluding sleep items. This approach is sound in minimizing criterion contamination because the predictor variables and outcome endpoint share overlapping content (refer to Dahlke et al., 2018, for an example in the context of a longitudinal study). As sleep disturbances are a key symptom of MDD and a shared predictor, including sleep items in both the predictor set and the W2 MDD severity outcome measure might artificially inflate correlations due to common variance rather than true predictive associations (Kell, 2022). To this end, the sensitivity analyses, which included sleep items, functioned as a robustness check (VanderWeele and Ding, 2017), validating that inferences about the prognostic utility of the top 10 W1 variables stayed consistent. These sensitivity analyses are crucial for testing the generalizability and stability of multivariable predictive models, particularly when predictors are highly overlapping.

Diverse calibration metrics were computed to offer unique and complementary appraisals of the degree to which predicted values and observed scores matched. The mean calibration error indicated the overall average discrepancy from perfect calibration (Jiang et al., 2011). Comparatively, the root mean square calibration error assigns more weight to greater discrepancies by computing the square root of the mean squared errors, emphasizing the effect of large miscalibrations. The expected calibration errors are computed by calculating the mean predicted difference between actual and predicted values within the predictive distribution of the model (Huang et al., 2020). The maximum calibration error identifies the unique, most significant observed model calibration difference. By determining the degree to which models may over- or under-predict specific score ranges, their real-world clinical utility could be evaluated more rigorously (Riley and Collins, 2023). Together, smaller error indices reflect good calibration, lower expected error suggests good model generalizability, and lower maximum error indicates fewer anomalous miscalibrations.

3. Results

3.1. Multivariable ML model performance in the test dataset

Table 2 overviews the multivariable ML model performance metrics

Table 2
Model performance of multivariable ML models predicting W2 MDD symptom severity in the test data set, excluding sleep items in the MDD symptom scales, with the top 20 W1 variables (Random forest imputation without IPW).

Model	Metric	Estimate	LCI	UCI
LASSO	MAE	0.170	0.152	0.188
	RMSE	0.239	0.210	0.264
	R^2	0.172	−0.001	0.284
Ridge	MAE	0.166	0.147	0.184
	RMSE	0.238	0.207	0.265
	R^2	0.176	0.043	0.269
ENR	MAE	0.167	0.149	0.185
	RMSE	0.238	0.208	0.264
	R^2	0.179	0.029	0.280
CART	MAE	0.130	0.107	0.151
	RMSE	0.241	0.201	0.277
	R^2	0.154	0.062	0.221
RF	MAE	0.172	0.155	0.192
	RMSE	0.245	0.217	0.275
	R^2	0.132	−0.024	0.221
GBM	MAE	0.141	0.122	0.163
	RMSE	0.235	0.201	0.268
	R^2	0.198	0.091	0.286
SVM	MAE	0.144	0.122	0.167
	RMSE	0.258	0.217	0.295
	R^2	0.031	−0.013	0.068
SLR	MAE	0.153	0.133	0.174
	RMSE	0.241	0.203	0.272
	R^2	0.157	0.065	0.228

Note. ML, machine learning; W2, wave 2 (2013–2014); MDD, major depressive disorder; IPW, inverse probability weights based on completer status; LCI, lower bound of the 95% confidence intervals (CIs); UCI, upper bound of the 95% CIs; LASSO, least absolute shrinkage, and selection operator; MAE, mean absolute error; RMSE, root mean squared error; R^2, R-squared; ENR, elastic net regularization; CART, classification and regression trees; RF, random forest; GBM, gradient boosting machine; SVM, support vector machine; SLR, Super Learner.

that excluded sleep items from the MDD scales at both time points and selected the top 20 features with an elastic net regularization filter. GBM was the most optimal multivariable ML model for predicting W2 MDD severity, with the lowest RMSE (0.235 [95% CIs] [0.201 to 0.268]) and MAE (0.141 [0.122 to 0.163]) values and the highest R^2 value (19.8% [9.1% to 28.6%]). The GBM algorithm accounted for 19.8% of the variance of W2 MDD severity in the held-out test set.

3.2. *Top 10 multivariable predictors of nine-year W2 MDD severity*

Fig. 1 summarizes the top 10 W1 predictors of W2 MDD severity in descending order of importance. The SHAP values ranged between about −0.05 and + 0.17, suggesting that the unique W1 variables had between small negative and moderate-to-strong positive contributions to the model's output on the normalized scale. Fig. 2 presents the corresponding partial dependence plots. The top three W1 correlates of W2 MDD severity were higher levels of psychiatric symptoms (the number in parentheses denoted the relative importance): (1) MDD severity; (2) GAD severity; (3) PD severity. The remaining seven W1 correlates of greater W2 MDD severity were actigraphy-indexed sleep-wake variables. These correlates included three sleep phase W1 variables: (4) longer TST; (5) lower sleep efficiency; (7) higher average wake time during sleep phase. Other W1 correlates of higher W2 MDD severity were two active phase variables: (6) fewer average sleep bouts; (10) shorter wake times. Nonlinear patterns emerged for two rest phase W1 variables: (8) high and low (vs. moderate) wake time percentage; (9) moderate (vs. high and low) wake time.

3.3. *Model calibration*

Table 3 presents the point estimates of diverse model calibration metrics in the test dataset for the primary GBM model with the top 20 W1 correlates of W2 MDD severity that used random forest imputation without IPWs. Fig. 3 shows the corresponding model calibration plot. The values of the key calibration metrics were as follows: calibration slope (1.621 [1.057 to 2.073]); calibration intercept (−0.061 [−0.097 to −0.017]); calibration R^2 (0.232 [0.103 to 0.371]); mean calibration error (0.141 [0.121 to 0.162]); and expected calibration error (0.050 [0.038 to 0.081]). Collectively, these metrics suggest that the final model had moderate-to-good calibration, where predictions displayed systematic bias but overall acceptable-to-good levels of mean to maximum deviation from perfect calibration.

3.4. *Sensitivity analyses*

Sensitivity analyses determined the degree to which the observed patterns remained identical under different predictor set lengths and missing data management approaches. The GBM continued to perform well if all 80 W1 variables were included in the predictor set (Table S9), and if the top 20 W1 variables were identified through the elastic net regularization filter while using both random forest imputation and IPWs (Table S10). Similar patterns of model performance metrics and predictor-outcome associations were observed if the predictor set and outcome included the CIDI-SF sleep-related items in the analyses (Tables S11 to S14 and Figs. S1 to S3 in the OSM).

4. Discussion

The present study used physiological ECG and sleep actigraphy wearable data to test their prognostic value in predicting nine-year MDD severity. No physiological and sociodemographic variables emerged as significant incremental distal risk factors of nine-year MDD severity. However, the final best-performing GBM multivariable ML model, which outperformed several other models (e.g., linear LASSO), revealed that sleep, rest, and active wake actigraphy indices were critical in predicting nine-year MDD severity. Below, we consider several potential implications for advancing clinical theory and practice.

Our multivariable ML model accounted for 19.8% of the variance in nine-year MDD severity and attained low RMSE and MAE values in the held-out test. These outcomes implied that the model offered clinically meaningful yet moderate prognostic value. Due to the intrinsic complexity, subjectivity, and variability in assessing mental health outcomes, predictive accuracies with modest levels could still aid in early detection of high-risk individuals and targeted prevention (Meehan et al., 2022). Identical R^2 values have been construed as practically significant, particularly in contexts where unique baseline variables account for small additional variance in the outcome (Gao, 2023). Although models explaining more than two-fifths of the variance in clinical outcomes are regarded as robust, accounting for close to 20% out-of-sample variance with rigorous predictor selection approaches and algorithms, such as GBM, could offer real-world implications. Future external validation studies are also required to refine model generalizability and transportability (Debray et al., 2023; Guerreiro et al., 2024).

Unsurprisingly, strong relative importance rankings were observed for more baseline MDD, GAD, and PD symptoms in predicting stronger nine-year MDD severity. Beyond replicating the persistence of MDD symptoms across long periods (Garcia-Toro et al., 2013), these findings emphasize how comorbid GAD and PD symptoms can aggravate their course (Hung et al., 2019). Potential mechanisms, such as emotion regulation repertoires (Barber et al., 2023a) and social relationships (Barber et al., 2023b), of MDD symptom chronicity and prospective comorbidity thus deserve further attention to explain these patterns.

Simultaneously, both longer TST and wake time, as well as lower sleep efficiency during the sleep phase, were correlated with higher nine-year MDD severity. These actigraphy markers likely reflected circadian rhythm dysregulations at the hormonal (Riemann et al., 2020) and neural levels (Wolf et al., 2016), which interfered with sleep homeostasis and restoration processes over time. Ironically, longer TST has

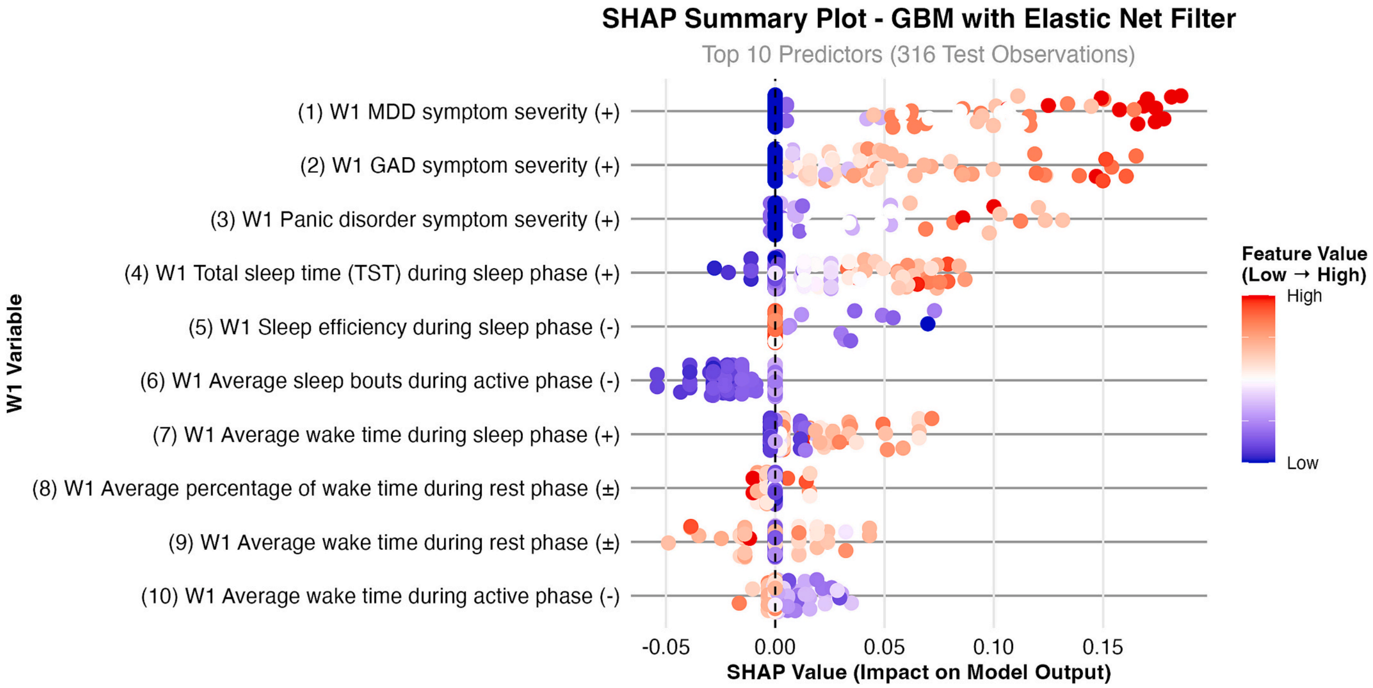


Fig. 1. SHAP Bee swarm plot of the multivariable GBM model of the top 10 W1 variables predicting W2 MDD symptom severity, excluding sleep items in the MDD symptom scales based on the testing sample

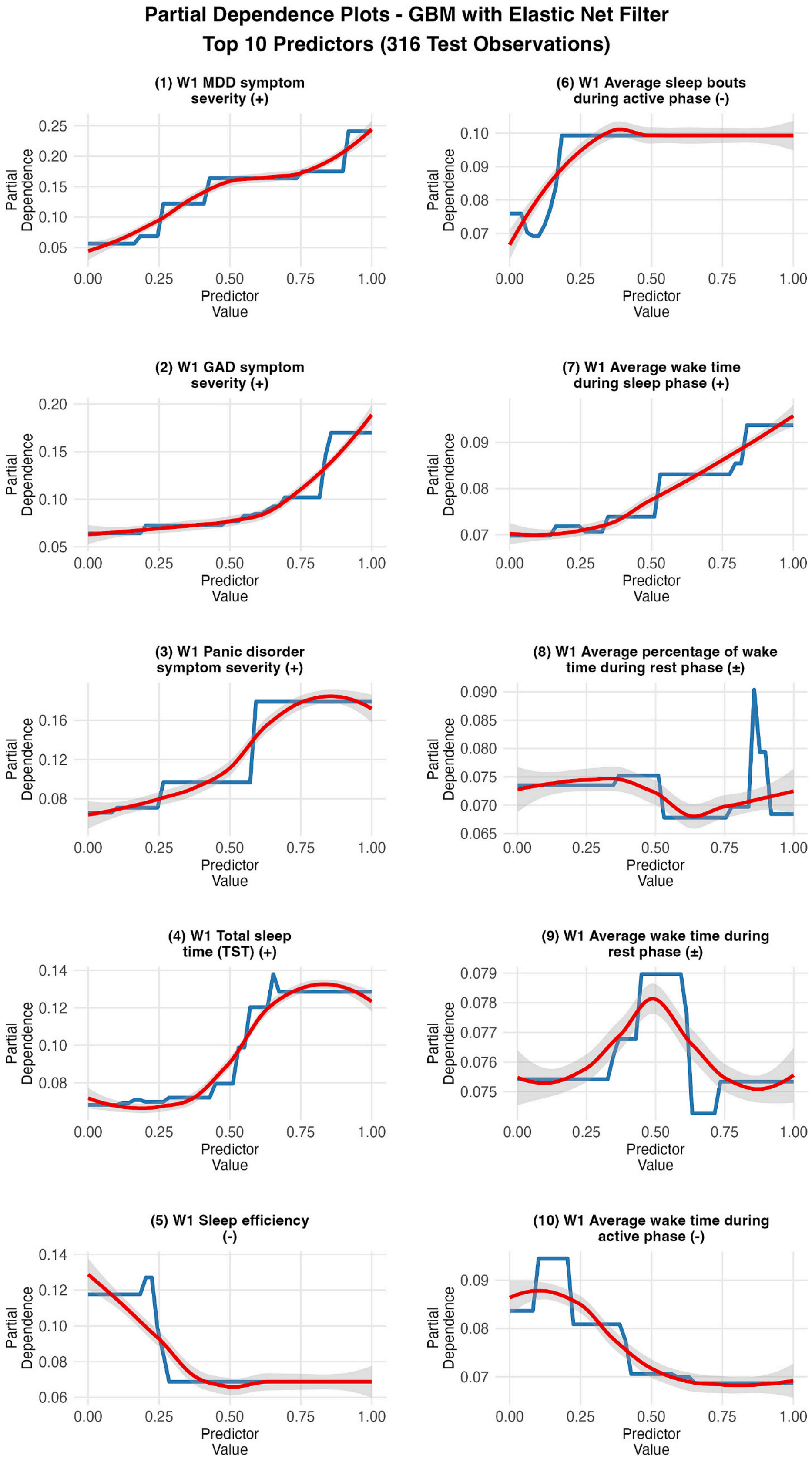
Note. SHAP, Shapley additive explanations; GBM, gradient boosting model; W1, wave 1 (2004–2006); W2, wave 2 (2013–2014); MDD, major depressive disorder; GAD, generalized anxiety disorder; TST, total sleep time. Each dot indicates a participant's data point. The *x*-axis indicates the SHAP value (i.e., the predictor's marginal effect on the model's predicted W2 MDD severity outcome). Positive SHAP values suggest that higher W1 variable values are correlated with higher predicted W2 MDD severity, whereas negative SHAP values imply a decreasing impact. The color gradient indicates the strength of the raw W1 variable value, such that blue points reflect lower values and red points indicate higher values. W1 variables are organized vertically by their global relative importance, i.e., W1 variables at the top are presented with the strongest mean effect on the model's output across all individuals. The (+) and (−) symbols indicate the overall sign of correlation between the W1 variable and the predicted W2 MDD severity in the fitted multivariable model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

been linked to atypical depression profiles (Ohayon and Roberts, 2015) and compromised neurocognition (Sen and Tai, 2023), probably indicating hypersomnia and sleep-wake fragmentation. Poorer sleep efficiency (i.e., lower ratio of sleep time to total time spent in bed) has been reliably identified as a depression risk factor (Yan et al., 2022), perhaps functioning as a proxy of nighttime somatic arousal and physiological hypersensitivity. Together, these findings could be situated in the context of the two-process theory of sleep (Borbely et al., 2016; Nutt et al., 2008). This framework proposes, as our findings suggested, that sleep is jointly managed by a homeostatic process, which regulates sleep need based on previous wakefulness, and a circadian process, which drives alertness and sleep timing across the 24-h everyday cycle.

Moreover, consistent with the two-process model, fewer sleep bouts and shorter wake times in the active phase were predictive of greater nine-year MDD severity, emphasizing their significance in mood regulation. Relatedly, the nonlinear rest phase patterns (high and low [vs. moderate] wake time percentage and moderate [vs. high and low] wake time) imply an intricate, non-monotonic association between sleep continuity patterns and nine-year MDD severity. These findings aligned with and extended prior observations of nonlinearities between sleep components and MDD severity (Shimizu et al., 2020; Yin et al., 2023). Collectively, the outcomes could be explained by how sleep aberrations, whether manifested as deficiencies or excesses, in the rest-wake phases might chronically disrupt optimal emotion regulation (Tsui and Chan, 2025), conferring distal risk for higher MDD severity.

Neurobiologically, the actigraphy patterns observed in the present study might be accounted for by various brain substrates intimately linked to the association between sleep disturbances and MDD severity. Low sleep efficiency as a precursor to higher MDD severity might be tied to compromised white matter in the internal capsule and corona radiata, decreased activity in the lingual and postcentral gyri, and greater angular gyrus connectivity (Yang et al., 2020). The sleep disturbance patterns we observed might also be attributed to deficits in cuneus-temporal lobe connectivity (Zhu et al., 2020) and suboptimal rapid eye movement (REM) activity connected with implicated brain regions (Zhang et al., 2024). On that note, deviations in REM duration and latency that are entwined with connectivity between motor and parietal cortices might also contribute to the current findings (Liu et al., 2025). Longitudinal studies that combine multimodal measures, including actigraphy, neuroimaging, and polysomnography, are necessary to build on existing work and to test these conjectures.

Notably, ECG or ANS indices did not emerge as key incremental predictors of nine-year MDD severity, which we predicted based on the transactional models that connect ANS dysregulations with depression and stress reactivity (Eberhart and Hammen, 2010). The primary GBM algorithm prioritizes baseline variables by their incremental value conditioned on other factors in the predictor set (Yarkoni and Westfall, 2017). MDD and comorbid anxiety severity, coupled with key actigraphy indices, concurs with notions that ANS activity is integrated within broader circadian and emotion regulation processes instead of being independently predictive across long durations (Kinoshita et al., 2024; Stange et al., 2023). The nine-year time horizon may also weaken predictive power, as vagal processes reflect state-level shifts to proximal stressors, aging, cardiac, and metabolic processes that were not measured repeatedly. Further, the ECG protocol might have had lower ecological validity than actigraphy wearables that captured everyday physical activity and sleep patterns. The ECG approach also focused on frequency-dimensions and recovery indices instead of a broader range of temporal-dimensions and nonlinear assessments. On the whole, these accounts might explain the limited incremental contribution of ECG and



(caption on next page)

Fig. 2. Partial dependence plots (PDPs) of the multivariable GBM model of the top 10 W1 variables predicting W2 MDD symptom severity, excluding sleep items in the MDD symptom scales, based on the testing sample

Note. PDPs, partial dependence plots; GBM, gradient boosting model; W1, wave 1 (2004–2006); W2, wave 2 (2013–2014); MDD, major depressive disorder; GAD, generalized anxiety disorder; TST, total sleep time. These PDPs were based on the best-performing GBM model that used an elastic net regularization filter, displaying the association between W1 variables and W2 MDD severity. Each panel illustrates the marginal effect of a specific W1 variable on predicted W2 MDD severity. The *y*-axis indicates the model's computed W1 MDD severity, and the *x*-axis depicts the normalized range of predictor values. Panels that show (+) reflect W1 variables correlated with greater W2 MDD severity as their values increase. Conversely, panels that display (−) indicate W1 variables correlated with lower W2 MDD severity as their values increase. Non-linear associations, such as logarithmic and plateau patterns, represent inflection points where the W1 variable's effect on W2 MDD symptom severity changes. Collectively, this figure relays how both clinical variables and objective sleep actigraphy markers provide both additive and differential contributions to variability in W2 MDD severity.

Table 3
Model calibration metrics of multivariable ML models predicting W2 MDD symptom severity in the test data set, excluding sleep items in the MDD symptom scales, with the top 20 W1 variables (Random forest imputation without IPW).

Description	Estimate	LCI	UCI	Meaning
Calibration Slope	1.621	1.057	2.073	Agreement between predicted and observed
Calibration Intercept	−0.061	−0.097	−0.017	Systematic bias in predictions
Calibration R-squared	0.232	0.103	0.371	Explained variance of calibration
Mean Calibration Error	0.141	0.121	0.162	Average deviation from perfect calibration
Root Mean Square Calibration Error	0.235	0.199	0.267	Square-root mean deviation from calibration
Expected Calibration Error	0.050	0.038	0.081	Expected difference between predicted and true
Maximum Calibration Error	0.140	0.064	0.277	Largest observed calibration deviation

Note. GBM, gradient boosting machine; LCI, lower bound of the 95% confidence intervals (CIs); UCI, upper bound of the 95% CIs.

ANS baseline variables as correlates of nine-year MDD severity.

The present study had some limitations. First, replication efforts to test the external validity of these multivariable clinical prediction models using ML techniques are needed with larger sample sizes (Luedtke et al., 2019). Second, most participants self-reportedly identified as White (91.2%), were recruited from three U.S. universities, and were in midlife to older adulthood. Because sleep and stressor exposure patterns may differ across diverse cultural contexts and subgroups, the transportability of our outcomes is probably limited. Readers should, thus, construe our findings as exploratory for other culturally diverse groups. Future well-powered studies are encouraged to conduct external validation (Gallitto et al., 2025), model updating, or recalibration efforts (Fehr et al., 2023) by examining how the sign and strength of parameter estimates might vary by racial groups or structural factors. Third, based on prior research, future similar multivariable ML studies should investigate whether similar findings are observed when ecological wearable or sensor physiology measures are administered instead of laboratory-based ECG recordings (Ettore et al., 2023; Sato et al., 2023). Fourth, unmeasured third variables, such as genetic factors implicated in the etiology of MDD symptoms (Bunney et al., 2015), should be included in future research. Fifth, the bidirectional relations among physiology, sleep actigraphy, and MDD symptoms require more exploration. Sixth, although an approach robust to high missingness (random forest imputation) was used, future replication attempts should be conducted in a dataset with a lower proportion of missingness.

However, several notable strengths of the study were evident. The multivariable ML models demonstrated excellent performance and calibration outcomes in the primary model, which excluded sleep items, as well as in sensitivity analyses that included the sleep items. Advanced multivariable ML modeling with interpretable ML methods was also conducted to examine the predictors of nine-year MDD symptoms. These approaches detect possibly complex and nonlinear associations and offer intuitive and nuanced insights into the results (Molnar, 2022). Finally, the long nine-year timeframe enhanced the prognostic value of the present prospective analyses.

Several clinical implications merit attention if future studies externally validate the pattern of results. If externally validated in future replication attempts, the multivariable ML model suggests that an actionable ‘prognostic calculator’ of increased future MDD severity can be developed (Collins et al., 2024). Using passive sensors with sleep actigraphy in this calculator is a strength that reduces the assessment burden. In other words, findings highlight the potential value of early prevention and treatment efforts to target both anxiety and depression symptoms, given the observed chronic nature of MDD symptoms. Treatments or universal prevention programs that address sleep consolidation, continuity, and routine may be essential for reducing the risk of chronic MDD symptoms or incidence (Fang et al., 2019). Cognitive behavioral therapy for insomnia (CBT-I) has consistently shown strong evidence in simultaneously targeting MDD symptoms and sleep disturbances (cf. meta-analysis by Furukawa et al., 2024). However, clinical scientists and healthcare policymakers might benefit from increased resources for prevention science, particularly through the evaluation of more universal prevention programs. Ultimately, these efforts should enhance sleep characteristics, improve quality of life, and mitigate the risk of exacerbating or emerging long-term symptoms of MDD.

Preprints

This manuscript has not been posted on any preprint server or elsewhere.

CRediT authorship contribution statement

Nur Hani Zainal: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Peter F. Hitchcock:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Investigation, Conceptualization.

Ethical approval

The study was approved by the institutional review boards (IRBs) of Harvard University, Georgetown University, the University of California at Los Angeles, and the University of Wisconsin at Madison. The use of publicly available data exempted it from additional IRB approval.

Declaration of Generative AI and AI-assisted technologies in the writing process

The authors confirm that no generative AI tools were used in the writing process of this manuscript.

Funding

The present manuscript received funding from the National University of Singapore (NUS) Presidential Young Professorship (PYP) Start-

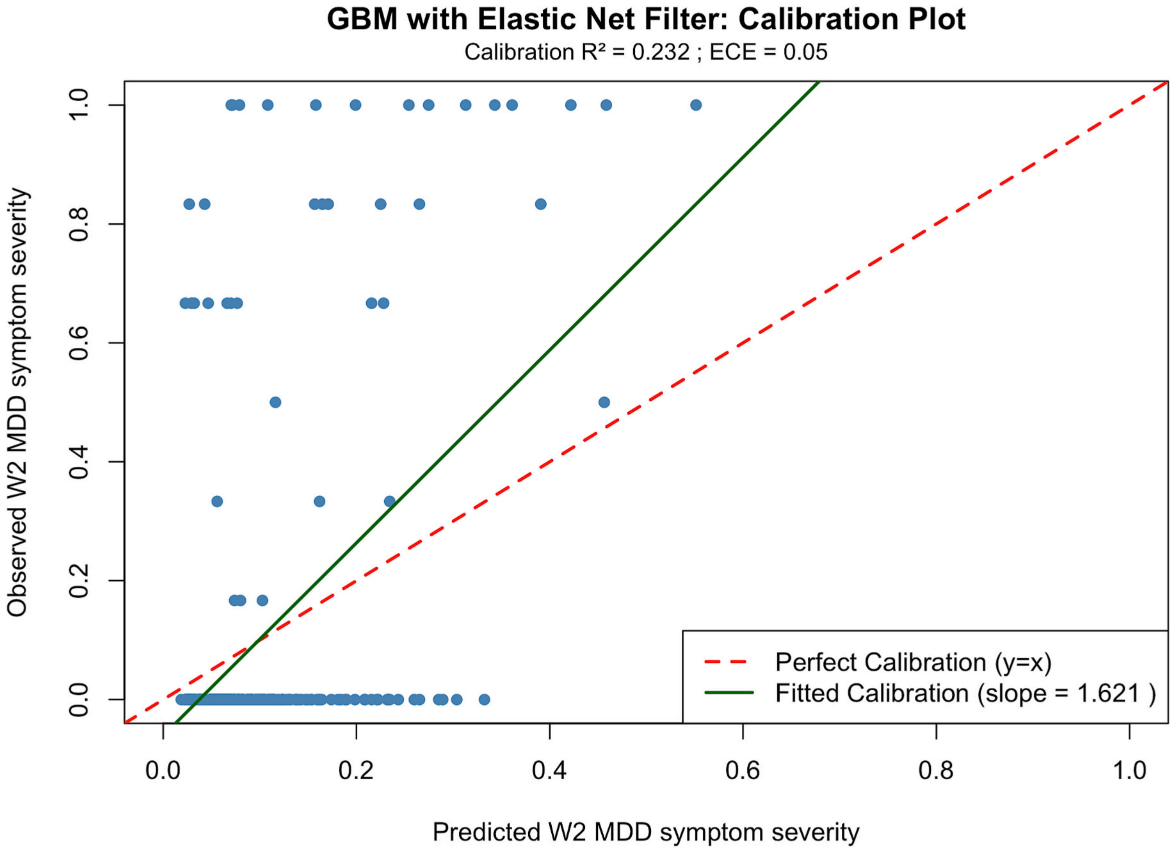


Fig. 3. Model calibration plot of best-performing multivariable ML model (GBM model), excluding sleep items in the MDD symptom scales based on the testing sample

Note. ML, machine learning; GBM, gradient boosting machine; MDD, major depressive disorder; R^2, R-squared; ECE, expected calibration error; W2, wave 2 (2013–2014). The x-axis presents the average predicted scores of MDD severity. Relatedly, the y-axis displays the matching average observed scores of MDD severity along deciles of predictions. The calibration plot contrasts the predicted W2 MDD symptom severity values with their observed scores on a normalized 0–1 scale. Each point represents a participant's data. The dashed red diagonal line denotes perfect calibration, i.e., predicted values totally corresponding with the observed scores. Points above the red line suggest underprediction, whereas points below the red line reflect overprediction. The solid green fitted line (slope = 1.621) indicates the model's predicted scores are restricted in range and consistently lower at higher W2 MDD symptom severity levels, i.e., small overestimation close to 0. The concentration of data points close to $y = 0$ indicates that most of the community adult participants have low MDD symptom severity. The reported calibration R^2 of 23.2% signals variability in observed W2 MDD symptom severity accounted by the calibration fit and indicates modest agreement. The ECE value of 0.05 represents the mean absolute difference between the predicted values and the observed scores of W2 MDD symptom severity, such that smaller values suggest better calibration. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Up Grant and White Space Fund. All funding agencies are not responsible for the analyses or interpretations presented here.

Declaration of competing interest

All authors report no conflicts of interest.

Acknowledgments

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jad.2025.120426.

References

Afkanpour, M., Hosseinzadeh, E., Tabesh, H., 2024. Identify the most appropriate imputation method for handling missing values in clinical structured datasets: a systematic review. BMC Med. Res. Methodol. 24 (1), 188. https://doi.org/10.1186/s12874-024-02310-6.

Allen, J.J., Chambers, A.S., Towers, D.N., 2007. The many metrics of cardiac chronotropy: a pragmatic primer and a brief comparison of metrics. Biol. Psychol. 74 (2), 243–262. https://doi.org/10.1016/j.biopsycho.2006.08.005.

American Psychiatric Association, 2022. Diagnostic and Statistical Manual of Mental Disorders (DSM-5), DSM-5-TR ed. American Psychiatric Association Publishing https://psychiatryonline.org/doi/book/10.1176/appi.books.9780890425787.

Aqua, J.K., Barnum, O., Johnson, D.A., 2024. Sleep as a contributor to socioeconomic disparities in hypertension: the midlife in the United States (MIDUS II). Study. Sleep 47 (9). https://doi.org/10.1093/sleep/zsae142.

Asarnow, L.D., 2020. Depression and sleep: what has the treatment research revealed and could the HPA axis be a potential mechanism? Curr. Opin. Psychol. 34, 112–116. https://doi.org/10.1016/j.copsyc.2019.12.002.

Barber, K.E., Zainal, N.H., Newman, M.G., 2023a. The mediating effect of stress reactivity in the 18-year bidirectional relationship between generalized anxiety and depression severity. J. Affect. Disord. 325, 502–512. https://doi.org/10.1016/j.jad.2023.01.041.

Barber, K.E., Zainal, N.H., Newman, M.G., 2023b. Positive relations mediate the bidirectional connections between depression and anxiety symptoms. J. Affect. Disord. 324, 387–394. https://doi.org/10.1016/j.jad.2022.12.082.

Berntson, G.G., Thomas Bigger Jr., J., Eckberg, D.L., Grossman, P., Kaufmann, P.G., Malik, M., Van Der Molen, M.W., 1997. Heart rate variability: origins, methods, and interpretive caveats. Psychophysiology 34 (6), 623–648. https://doi.org/10.1111/j.1469-8986.1997.tb02140.x.

Blood, J.D., Wu, J., Chaplin, T.M., Hommer, R., Vazquez, L., Rutherford, H.J., Crowley, M.J., 2015. The variable heart: high frequency and very low frequency correlates of depressive symptoms in children and adolescents. J. Affect. Disord. 186, 119–126. https://doi.org/10.1016/j.jad.2015.06.057.

Boland, E.M., Goldschmied, J.R., Gehrman, P.R., 2023. Does insomnia treatment prevent depression? Sleep 46 (6). https://doi.org/10.1093/sleep/zsad104.

Borbely, A.A., Daan, S., Wirz-Justice, A., Deboer, T., 2016. The two-process model of sleep regulation: a reappraisal. J. Sleep Res. 25 (2), 131–143. https://doi.org/10.1111/jsr.12371.

Brindle, R.C., Yu, L., Buysse, D.J., Hall, M.H., 2019. Empirical derivation of cutoff values for the sleep health metric and its relationship to cardiometabolic morbidity: results from the midlife in the United States (MIDUS) study. Sleep 42 (9). https://doi.org/10.1093/sleep/zsz116.

Bunney, B.G., Li, J.Z., Walsh, D.M., Stein, R., Vawter, M.P., Cartagena, P., Bunney, W.E., 2015. Circadian dysregulation of clock genes: clues to rapid treatments in major depressive disorder. Mol. Psychiatry 20 (1), 48–55. https://doi.org/10.1038/mp.2014.138.

Carnevali, L., Thayer, J.F., Brosschot, J.F., Ottaviani, C., 2018. Heart rate variability mediates the link between rumination and depressive symptoms: a longitudinal study. Int. J. Psychophysiol. 131, 131–138. https://doi.org/10.1016/j.ijpsycho.2017.11.002.

Collins, G.S., Dhiman, P., Ma, J., Schlussel, M.M., Archer, L., Van Calster, B., Riley, R.D., 2024. Evaluation of clinical prediction models (part 1): from development to external validation. BMJ 384, e074819. https://doi.org/10.1136/bmj-2023-074819.

Crowley, S.J., Wolfson, A.R., Tarokh, L., Carskadon, M.A., 2018. An update on adolescent sleep: new evidence informing the perfect storm model. J. Adolesc. 67, 55–65. https://doi.org/10.1016/j.adolescence.2018.06.001.

Dahlke, J.A., Kostal, J.W., Sackett, P.R., Kuncel, N.R., 2018. Changing abilities vs. changing tasks: examining validity degradation with test scores and college performance criteria both assessed longitudinally. J. Appl. Psychol. 103 (9), 980–1000. https://doi.org/10.1037/apl0000316.

Daza, E.J., Hudgens, M.G., Herring, A.H., 2017. Estimating inverse-probability weights for longitudinal data with dropout or truncation: the xtrccipw command. Stata J. 17 (2), 253–278.

Debray, T.P.A., Collins, G.S., Riley, R.D., Snell, K.I.E., Van Calster, B., Reitsma, J.B., Moons, K.G.M., 2023. Transparent reporting of multivariable prediction models developed or validated using clustered data (TRIPOD-cluster): explanation and elaboration. BMJ 380, e071058. https://doi.org/10.1136/bmj-2022-071058.

Eberhart, N.K., Hammen, C.L., 2010. Interpersonal style, stress, and depression: an examination of transactional and diathesis-stress models. J. Soc. Clin. Psychol. 29 (1), 23–38. https://doi.org/10.1521/jscp.2010.29.1.23.

Ettore, E., Muller, P., Hinze, J., Riemenschneider, M., Benoit, M., Giordana, B., Konig, A., 2023. Digital phenotyping for differential diagnosis of major depressive episode: narrative review. JMIR Mental Health 10, e37225. https://doi.org/10.2196/37225.

Fang, H., Tu, S., Sheng, J., Shao, A., 2019. Depression in sleep disturbance: a review on a bidirectional relationship, mechanisms and treatment. J. Cell. Mol. Med. 23 (4), 2324–2332. https://doi.org/10.1111/jcmm.14170.

Fantini-Hauwel, C., Batsele, E., Gois, C., Noel, X., 2020. Emotion regulation difficulties are not always associated with negative outcomes on women: the buffer effect of HRV. Front. Psychol. 11, 697. https://doi.org/10.3389/fpsyg.2020.00697.

Fehr, J., Piccininni, M., Kurth, T., Konigorski, S., 2023. Assessing the transportability of clinical prediction models for cognitive impairment using causal models. BMC Med. Res. Methodol. 23 (1), 187. https://doi.org/10.1186/s12874-023-02003-6.

Furukawa, Y., Nagaoka, D., Sato, S., Toyomoto, R., Takashina, H.N., Kobayashi, K., Kasai, K., 2024. Cognitive behavioral therapy for insomnia to treat major depressive disorder with comorbid insomnia: a systematic review and meta-analysis. J. Affect. Disord. 367, 359–366. https://doi.org/10.1016/j.jad.2024.09.017.

Gallitto, G., Englert, R., Kincses, B., Kotikalapudi, R., Li, J., Hoffschlag, K., Spisak, T., 2025. External validation of machine learning models-registered models and adaptive sample splitting. Gigascience 14. https://doi.org/10.1093/gigascience/giaf036.

Gao, J., 2023. R-squared (R2) – how much variation is explained? Research Methods in Medicine & Health Sciences 5 (4), 104–109. https://doi.org/10.1177/26320843231186398.

Garcia-Toro, M., Rubio, J.M., Gili, M., Roca, M., Jin, C.J., Liu, S.M., Blanco, C., 2013. Persistence of chronic major depression: a national prospective study. J. Affect. Disord. 151 (1), 306–312. https://doi.org/10.1016/j.jad.2013.06.013.

Gentili, C., Valenza, G., Nardelli, M., Lanata, A., Bertschy, G., Weiner, L., Pietrini, P., 2017. Longitudinal monitoring of heartbeat dynamics predicts mood changes in bipolar patients: a pilot study. J. Affect. Disord. 209, 30–38. https://doi.org/10.1016/j.jad.2016.11.008.

Goldenholz, D.M., Sun, H., Ganglberger, W., Westover, M.B., 2023. Sample size analysis for machine learning clinical validation studies. Biomedicines 11 (3). https://doi.org/10.3390/biomedicines11030685.

Grippo, A.J., Johnson, A.K., 2009. Stress, depression and cardiovascular dysregulation: a review of neurobiological mechanisms and the integration of research from preclinical disease models. Stress 12 (1), 1–21. https://doi.org/10.1080/10253890802046281.

Gruenewald, T., Seeman, T.E., Choo, T.H., Scodes, J., Snyder, C., Pavlicova, M., Sloan, R.P., 2023. Cardiovascular variability, sociodemographics, and biomarkers of disease: the MIDUS study. Front. Physiol. 14, 1234427. https://doi.org/10.3389/fphys.2023.1234427.

Guerreiro, J., Garriga, R., Lozano Bagen, T., Sharma, B., Karnik, N.S., Matic, A., 2024. Transatlantic transferability and replicability of machine-learning algorithms to predict mental health crises. npj Digital Medicine 7 (1), 227. https://doi.org/10.1038/s41746-024-01203-8.

Gupta, A., Stead, T.S., Ganti, L., 2024. Determining a meaningful R-squared value in clinical medicine. Academic Medicine & Surgery. https://doi.org/10.62186/001c.125154.

Hohls, J.K., Konig, H.H., Quirke, E., Hajek, A., 2021. Anxiety, depression and quality of life - a systematic review of evidence from longitudinal observational studies. Int. J. Environ. Res. Public Health 18 (22). https://doi.org/10.3390/ijerph182212022.

Huang, Y., Li, W., Macheret, F., Gabriel, R.A., Ohno-Machado, L., 2020. A tutorial on calibration measurements and calibration models for clinical prediction models. J. Am. Med. Inform. Assoc. 27 (4), 621–633. https://doi.org/10.1093/jamia/ocz228.

Hung, C.I., Liu, C.Y., Yang, C.H., 2019. Persistent depressive disorder has long-term negative impacts on depression, anxiety, and somatic symptoms at 10-year follow-up among patients with major depressive disorder. J. Affect. Disord. 243, 255–261. https://doi.org/10.1016/j.jad.2018.09.068.

Iglesias, D., Sorrel, M.A., Olmos, R., 2025. Cross-validation and predictive metrics in psychological research: Do not leave out the leave-one-out. Behav. Res. Methods 57 (3), 85. https://doi.org/10.3758/s13428-024-02588-w.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning, 103. Springer New York.

Jandackova, V.K., Britton, A., Malik, M., Steptoe, A., 2016. Heart rate variability and depressive symptoms: a cross-lagged analysis over a 10-year period in the Whitehall II study. Psychol. Med. 46 (10), 2121–2131. https://doi.org/10.1017/S003329171600060X.

Jiang, X., Osl, M., Kim, J., Ohno-Machado, L., 2011. Smooth isotonic regression: a new method to calibrate predictive models. AMIA Summits on Translational Science Proceedings 2011, 16–20.

Kell, H.J., 2022. The criterion problem in cross-cultural performance research. Int. J. Cross-cult. Manag. 22 (3), 389–411. https://doi.org/10.1177/14705958221100669.

Kelly, J.R., Clarke, G., Cryan, J.F., Dinan, T.G., 2018. Dimensional thinking in psychiatry in the era of the research domain criteria (RDoC). Ir. J. Psychol. Med. 35 (2), 89–94. https://doi.org/10.1017/ipm.2017.7.

Kerrigan, D., Barr, B., Bertini, E., 2025. PDPilot: exploring partial dependence plots through ranking, filtering, and clustering. IEEE Trans. Vis. Comput. Graph. 31 (10), 7377–7390. https://doi.org/10.1109/TVCG.2025.3545025.

Kessler, R.C., Üstün, T.B., 2004. The world mental health (WMH) survey initiative version of the World Health Organization (WHO) composite international diagnostic interview (CIDI). Int. J. Methods Psychiatr. Res. 13 (2), 93–121. https://doi.org/10.1002/mpr.168.

Kessler, R.C., Andrews, G., Mroczek, D., Ustun, B., Wittchen, H.-U., 1998a. The World Health Organization composite international diagnostic interview short-form (CIDI-SF). Int. J. Methods Psychiatr. Res. 7 (4), 171–185. https://doi.org/10.1002/mpr.47.

Kessler, R.C., Wittchen, H.-U., Abelson, J.M., Mcgonagle, K., Schwarz, N., Kendler, K.S., Zhao, S., 1998b. Methodological studies of the composite international diagnostic interview (CIDI) in the US national comorbidity survey (NCS). Int. J. Methods Psychiatr. Res. 7 (1), 33–55. https://doi.org/10.1002/mpr.33.

Kimhy, D., Crowley, O.V., McKinley, P.S., Burg, M.M., Lachman, M.E., Tun, P.A., Sloan, R.P., 2013. The association of cardiac vagal control and executive functioning–findings from the MIDUS study. J. Psychiatr. Res. 47 (5), 628–635. https://doi.org/10.1016/j.jpsychires.2013.01.018.

Kinoshita, S., Hanashiro, S., Tsutsumi, S., Shiga, K., Kitazawa, M., Wada, Y., Kishimoto, T., 2024. Assessment of stress and well-being of Japanese employees using wearable devices for sleep monitoring combined with ecological momentary assessment: pilot observational study. JMIR Form. Res. 8, e49396. https://doi.org/10.2196/49396.

Koch, C., Wilhelm, M., Salzmann, S., Rief, W., Euteneuer, F., 2019. A meta-analysis of heart rate variability in major depression. Psychol. Med. 49 (12), 1948–1957. https://doi.org/10.1017/S0033291719001351.

Kovač, N., Ratković, K., Farahani, H., Watson, P., 2024. A practical applications guide to machine learning regression models in psychology with Python. Methods in Psychology 11. https://doi.org/10.1016/j.metip.2024.100156.

Laborde, S., Mosley, E., Thayer, J.F., 2017. Heart rate variability and cardiac vagal tone in psychophysiological research - recommendations for experiment planning, data analysis, and data reporting. Front. Psychol. 8, 213. https://doi.org/10.3389/fpsyg.2017.00213.

Lee, S.A., Fisher, Z., Almeida, D.M., 2025. Daily reciprocal relationships between affect, physical activity, and sleep in middle and later life. Ann. Behav. Med. 59 (1). https://doi.org/10.1093/abm/kaae072.

Lee, T., Shi, D., 2021. A comparison of full information maximum likelihood and multiple imputation in structural equation modeling with missing data. Psychol. Methods 26 (4), 466–485. https://doi.org/10.1037/met0000381.

Lewis, M.J., Spiliopoulou, A., Goldmann, K., Pitzalis, C., McKeigue, P., Barnes, M.R., 2023. nestedcv: An R package for fast implementation of nested cross-validation with embedded feature selection designed for transcriptomics and high-dimensional data. Bioinformatics Adv. 3 (1), vbad048. https://doi.org/10.1093/bioadv/vbad048.

Liu, R.T., Steele, S.J., Hamilton, J.L., Do, Q.B.P., Furbish, K., Burke, T.A., Gerlus, N., 2020. Sleep and suicide: a systematic review and meta-analysis of longitudinal studies. Clin. Psychol. Rev. 81, 101895. https://doi.org/10.1016/j.cpr.2020.101895.

Liu, S., Chen, J., Guan, L., Xu, L., Cai, H., Wang, J., Yu, Y., 2025. The brain, rapid eye movement sleep, and major depressive disorder: a multimodal neuroimaging study. Prog. Neuro-Psychopharmacol. Biol. Psychiatry 136, 111151. https://doi.org/10.1016/j.pnpbp.2024.111151.

Lovato, N., Gradisar, M., 2014. A meta-analysis and model of the relationship between sleep and depression in adolescents: recommendations for future research and clinical practice. Sleep Med. Rev. 18 (6), 521–529. https://doi.org/10.1016/j.smrv.2014.03.006.

Love, G.D., Seeman, T.E., Weinstein, M., Ryff, C.D., 2010. Bioindicators in the MIDUS national study: protocol, measures, sample, and comparative context. J. Aging Health 22 (8), 1059–1080. https://doi.org/10.1177/0898264310374355.

Luedtke, A., Sadikova, E., Kessler, R.C., 2019. Sample size requirements for multivariate models to predict between-patient differences in best treatments of major depressive disorder. Clin. Psychol. Sci. 7 (3), 445–461. https://doi.org/10.1177/2167702618815466.

Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Proces. Syst. 30, 4768–4777. https://doi.org/10.5555/3295222.3295230.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Lee, S.I., 2020. From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. 2 (1), 56–67. https://doi.org/10.1038/s42256-019-0138-9.

Massar, S.A.A., Chua, X.Y., Soon, C.S., Ng, A.S.C., Ong, J.L., Chee, N., Chee, M.W.L., 2021. Trait-like nocturnal sleep behavior identified by combining wearable, phone-use, and self-report data. NPJ Digit. Med. 4 (1), 90. https://doi.org/10.1038/s41746-021-00466-9.

Mayer, M., 2024. missRanger: fast imputation of missing values. In (Version 2.5.0). https://CRAN.R-project.org/package=missRanger.

Meehan, A.J., Lewis, S.J., Fazel, S., Fusar-Poli, P., Steyerberg, E.W., Stahl, D., Danese, A., 2022. Clinical prediction models in psychiatry: a systematic review of two decades of progress and challenges. Mol. Psychiatry 27 (6), 2700–2708. https://doi.org/10.1038/s41380-022-01528-4.

Molnar, C., 2022. Interpretable machine learning: a guide for making black-box models explainable, 2nd ed. https://christophm.github.io/interpretable-ml-book/cite.html.

Moreno-Agostino, D., Wu, Y.T., Daskalopoulou, C., Hasan, M.T., Huisman, M., Prina, M., 2021. Global trends in the prevalence and incidence of depression: a systematic review and meta-analysis. J. Affect. Disord. 281, 235–243. https://doi.org/10.1016/j.jad.2020.12.035.

Morris, S.E., Sanislow, C.A., Pacheco, J., Vaidyanathan, U., Gordon, J.A., Cuthbert, B.N., 2022. Revisiting the seven pillars of RDoC. BMC Med. 20 (1), 220. https://doi.org/10.1186/s12916-022-02414-0.

Ng, M.H.S., Zainal, N.H., Newman, M.G., 2024. Positive reappraisal coping mediates the relationship between parental abuse and lack of affection on adulthood generalized anxiety severity. J. Anxiety Disord. 102, 102826. https://doi.org/10.1016/j.janxdis.2024.102826.

Nutt, D., Wilson, S., Paterson, L., 2008. Sleep disorders as core symptoms of depression. Dialogues Clin. Neurosci. 10 (3), 329–336. https://doi.org/10.31887/DCNS.2008.10.3/dnutt.

Ohayon, M.M., Roberts, L.W., 2015. Challenging the validity of the association between oversleeping and overeating in atypical depression. J. Psychosom. Res. 78 (1), 52–57. https://doi.org/10.1016/j.jpsychores.2014.09.018.

Palmer, C.A., Alfano, C.A., 2017. Sleep and emotion regulation: an organizing, integrative review. Sleep Med. Rev. 31, 6–16. https://doi.org/10.1016/j.smrv.2015.12.006.

Pearl, J., 2014. Interpretation and identification of causal mediation. Psychol. Methods 19 (4), 459–481. https://doi.org/10.1037/a0036434.

Quintana, D.S., Alvares, G.A., Heathers, J.A., 2016. Guidelines for reporting articles on psychiatry and heart rate variability (GRAPH): recommendations to advance research communication. Transl. Psychiatry 6 (5), e803. https://doi.org/10.1038/tp.2016.73.

R Core Team, 2025. R: A language and environment for statistical computing. (Version 4.4.1) R Foundation for Statistical Computing. https://www.R-project.org/.

Rajput, D., Wang, W.J., Chen, C.C., 2023. Evaluation of a decided sample size in machine learning applications. BMC Bioinformatics 24 (1), 48. https://doi.org/10.1186/s12859-023-05156-9.

Riemann, D., Krone, L.B., Wulff, K., Nissen, C., 2020. Sleep, insomnia, and depression. Neuropsychopharmacology 45 (1), 74–89. https://doi.org/10.1038/s41386-019-0411-y.

Riley, R.D., Collins, G.S., 2023. Stability of clinical prediction models developed using statistical or machine learning methods. Biom. J. 65 (8), e2200302. https://doi.org/10.1002/bimj.202200302.

Ryff, C., Almeida, D., Ayanian, J., Binkley, N., Carr, D.S., Coe, C., Williams, D., 2019. *Midlife in the United States (MIDUS 3), 2013–2014* Inter-University Consortium for Political and Social Research [Distributor]. https://doi.org/10.3886/ICPSR36346.v7.

Ryff, C.D., Almeida, D.M., Ayanian, J.Z., Carr, D.S., Cleary, P.D., Coe, C., Williams, D.R., 2021. *Midlife in the United States (MIDUS 2), 2004–2006* Inter-University Consortium for Political and Social Research [Distributor]. https://doi.org/10.3886/ICPSR04652.v8.

Sameroff, A.J., 2020. It's more complicated. Annu. Rev. Dev. Psychol. 2 (Volume 2, 2020), 1–26. https://doi.org/10.1146/annurev-devpsych-061520-120738.

Sato, S., Hiratsuka, T., Hasegawa, K., Watanabe, K., Obara, Y., Kariya, N., Matsui, T., 2023. Screening for major depressive disorder using a wearable ultra-short-term HRV monitor and signal quality indices. Sensors 23 (8). https://doi.org/10.3390/s23083867.

Schroeders, U., Schmidt, C., Gnambs, T., 2022. Detecting careless responding in survey data using stochastic gradient boosting. Educ. Psychol. Meas. 82 (1), 29–56. https://doi.org/10.1177/00131644211004708.

Sen, A., Tai, X.Y., 2023. Sleep duration and executive function in adults. Curr. Neurol. Neurosci. Rep. 23 (11), 801–813. https://doi.org/10.1007/s11910-023-01309-8.

Sgoifo, A., Carnevali, L., Alfonso Mde, L., Amore, M., 2015. Autonomic dysfunction and heart rate variability in depression. Stress 18 (3), 343–352. https://doi.org/10.3109/10253890.2015.1045868.

Shaffer, F., Ginsberg, J.P., 2017. An overview of heart rate variability metrics and norms. Front. Public Health 5, 258. https://doi.org/10.3389/fpubh.2017.00258.

Shah, A.D., Bartlett, J.W., Carpenter, J., Nicholas, O., Hemingway, H., 2014. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. Am. J. Epidemiol. 179 (6), 764–774. https://doi.org/10.1093/aje/kwt312.

Shimizu, M., Gillis, B.T., Buckhalt, J.A., El-Sheikh, M., 2020. Linear and nonlinear associations between sleep and adjustment in adolescence. Behav. Sleep Med. 18 (5), 690–704. https://doi.org/10.1080/15402002.2019.1665049.

Shorey, S., Ng, E.D., Wong, C.H.J., 2022. Global prevalence of depression and elevated depressive symptoms among adolescents: a systematic review and meta-analysis. Br. J. Clin. Psychol. 61 (2), 287–305. https://doi.org/10.1111/bjc.12333.

Shrivastava, D., Jung, S., Saadat, M., Sirohi, R., Crewson, K., 2014. How to interpret the results of a sleep study. J. Community Hosp. Intern. Med. Perspect. 4 (5), 24983. https://doi.org/10.3402/jchimp.v4.24983.

Sivertsen, H., Bjorklof, G.H., Engedal, K., Selbaek, G., Helvik, A.S., 2015. Depression and quality of life in older persons: a review. Dement. Geriatr. Cogn. Disord. 40 (5–6), 311–339. https://doi.org/10.1159/000437299.

Spreen, O., Strauss, E., 1998. A Compendium of Neuropsychological Tests. Oxford University Press.

Stange, J.P., Li, J., Xu, E.P., Ye, Z., Zapetis, S.L., Phanord, C.S., Langenecker, S.A., 2023. Autonomic complexity dynamically indexes affect regulation in everyday life. J. Psychopathol. Clin. Sci. 132 (7), 847–866. https://doi.org/10.1037/abn0000849.

Stroop, J.R., 1935. Studies of interference in serial verbal reactions. J. Exp. Psychol. 18 (6), 643–662. https://doi.org/10.1037/h0054651.

Tang, F., Ishwaran, H., 2017. Random forest missing data algorithms. Stat. Anal. Data Min. 10 (6), 363–377. https://doi.org/10.1002/sam.11348.

Thase, M.E., 2006. Depression and sleep: pathophysiology and treatment. Dialogues Clin. Neurosci. 8 (2), 217–226. https://doi.org/10.31887/DCNS.2006.8.2/mthase.

Tsui, H.T.C., Chan, W.S., 2025. Daily associations between sleep parameters and depressive symptoms in individuals with insomnia: investigating emotional reactivity and regulation as mediators. Behav. Sleep Med. 23 (1), 1–16. https://doi.org/10.1080/15402002.2024.2399620.

VanderWeele, T.J., Ding, P., 2017. Sensitivity analysis in observational research: introducing the E-value. Ann. Intern. Med. 167 (4), 268–274. https://doi.org/10.7326/M16-2607.

Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H., Guisan, A., 2008. Effects of sample size on the performance of species distribution models. Divers. Distrib. 14 (5), 763–773. https://doi.org/10.1111/j.1472-4642.2008.00482.x.

Wolf, E., Kuhn, M., Normann, C., Mainberger, F., Maier, J.G., Maywald, S., Nissen, C., 2016. Synaptic plasticity model of therapeutic sleep deprivation in major depression. Sleep Med. Rev. 30, 53–62. https://doi.org/10.1016/j.smrv.2015.11.003.

Yan, B., Zhao, B., Jin, X., Xi, W., Yang, J., Yang, L., Ma, X., 2022. Sleep efficiency may predict depression in a large population-based study. Front. Psychol. 13, 838907. https://doi.org/10.3389/fpsyt.2022.838907.

Yang, Y., Zhu, D.M., Zhang, C., Zhang, Y., Wang, C., Zhang, B., Yu, Y., 2020. Brain structural and functional alterations specific to low sleep efficiency in major depressive disorder. Front. Neurosci. 14, 50. https://doi.org/10.3389/fnins.2020.00050.

Yarkoni, T., Westfall, J., 2017. Choosing prediction over explanation in psychology: lessons from machine learning. Perspect. Psychol. Sci. 12 (6), 1100–1122. https://doi.org/10.1177/1745691617693393.

Yin, J., Wang, H., Li, S., Zhao, L., You, Y., Yang, J., Liu, Y., 2023. Nonlinear relationship between sleep midpoint and depression symptoms: a cross-sectional study of US adults. BMC Psychiatry 23 (1), 671. https://doi.org/10.1186/s12888-023-05130-y.

Yip, T., Chen, M., Wang, Y., Slopen, N., Chae, D., Priest, N., Williams, D., 2021. Linking discrimination and sleep with biomarker profiles: an investigation in the MIDUS study. Compr Psychoneuroendocrinol 5. https://doi.org/10.1016/j.cpnec.2020.100021.

Zainal, N.H., Newman, M.G., 2021. Depression and executive functioning bidirectionally impair one another across 9 years: evidence from within-person latent change and cross-lagged models. Eur. Psychiatry 64 (1), e43, 41–14. https://doi.org/10.1192/j.eurpsy.2021.2217.

Zainal, N.H., Newman, M.G., 2022a. Depression and worry symptoms predict future executive functioning impairment via inflammation. Psychol. Med. 1–11. https://doi.org/10.1017/S0033291721000398.

Zainal, N.H., Newman, M.G., 2022b. Inflammation mediates depression and generalized anxiety symptoms predicting executive function impairment after 18 years. J. Affect. Disord. 296, 465–475. https://doi.org/10.1016/j.jad.2021.08.077.

Zhai, L., Zhang, H., Zhang, D., 2015. Sleep duration and depression among adults: a meta-analysis of prospective studies. Depress. Anxiety 32 (9), 664–670. https://doi.org/10.1002/da.22386.

Zhang, C., Zhu, D.M., Zhang, Y., Chen, T., Liu, S., Chen, J., Yu, Y., 2024. Neural substrates underlying REM sleep duration in patients with major depressive disorder: a longitudinal study combining multimodal MRI data. J. Affect. Disord. 344, 546–553. https://doi.org/10.1016/j.jad.2023.10.090.

Zhu, D.M., Zhang, C., Yang, Y., Zhang, Y., Zhao, W., Zhang, B., Yu, Y., 2020. The relationship between sleep efficiency and clinical symptoms is mediated by brain function in major depressive disorder. J. Affect. Disord. 266, 327–337. https://doi.org/10.1016/j.jad.2020.01.155.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat Methodol. 67 (2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.