

# DETECTION OF NATURALISTIC DEPRESSION SYMPTOM VARIABILITY

## 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

George D. Price<sup>1,2</sup>, Michael V. Heinz<sup>1,3</sup>, Seo Ho Song<sup>4</sup>, Matthew Nemesure<sup>1,2</sup>, Nicholas C. Jacobson<sup>1,2,3,5</sup>

<sup>1</sup>Center for Technology and Behavioral Health, Geisel School of Medicine, Dartmouth College, Lebanon, NH, United States

<sup>2</sup>Quantitative Biomedical Sciences Program, Dartmouth College, Lebanon, NH, United States

<sup>3</sup>Department of Psychiatry, Geisel School of Medicine, Dartmouth College, Lebanon, NH, United States


<sup>4</sup>Department of Psychiatry, Beth Israel Deaconess Medical Center, Boston, MA, United States


<sup>5</sup>Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College, Lebanon, NH, United States


### Author Note

George D. Price  <https://orcid.org/0000-0002-9164-4973>

Michael V. Heinz  <https://orcid.org/0000-0003-0866-0508>

Seo Ho Song  <https://orcid.org/0000-0003-2970-2746>

Matthew Nemesure  <https://orcid.org/0000-0002-2369-600X>

Nicholas C. Jacobson  <https://orcid.org/0000-0002-8832-4741>

We have no known conflicts of interest to disclose.

Correspondence of this article should be addressed to George D. Price, Dartmouth College, 46 Centerra Parkway Suite 300, Office # 336S, Lebanon, NH 03766.

Email: [George.Price.GR@dartmouth.edu](mailto:George.Price.GR@dartmouth.edu)

30

## Abstract

31 Major Depressive Disorder (MDD) presents considerable challenges to diagnosis and management due to  
32 symptom variability across time. Only recent work has highlighted the clinical implications for  
33 interrogating depression symptom *variability*. Thus, the present work investigates how sociodemographic,  
34 comorbidity, movement, and sleep data is associated with long-term depression symptom *variability*.  
35 Participant information included ( $N = 939$ ) baseline sociodemographic and comorbidity data,  
36 longitudinal, passively-collected wearable data, and Patient Health Questionnaire-9 (PHQ-9) scores  
37 collected over 12 months. An ensemble machine learning approach was used to detect long-term  
38 depression symptom variability via: (i) a domain-driven feature selection approach, and (ii) an exhaustive  
39 feature inclusion approach. SHapley Additive exPlanations (SHAP) was used to interrogate variable  
40 importance and directionality. The composite domain-driven and exhaustive inclusion models were both  
41 capable of moderately detecting long-term depression symptom variability ( $r = 0.33$  and  $r = 0.39$ ,  
42 respectively). Our results indicate the incremental predictive validity of sociodemographic, comorbidity,  
43 and passively-collected wearable movement and sleep data in detecting long-term depression symptom  
44 variability.

45

46

## DETECTION OF NATURALISTIC DEPRESSION SYMPTOM VARIABILITY

### 47 **Using Digital Phenotyping to Capture Depression Symptom Variability: Detecting Naturalistic** 48 **Variability in Depression Symptoms Across One Year Using Passively-Collected Wearable** 49 **Movement and Sleep Data**

50

51 Major Depressive Disorder (MDD) is highly prevalent and burdensome, socially and  
52 economically. An estimated 8% of all U.S. adults (nearly 21 M) experienced a depressive episode in the  
53 last year [1], and an estimated 6% (15 M) experienced associated severe functional impairment [1].  
54 Depression is ranked in the top twenty leading causes of disability, globally [2] and is estimated to cost  
55 \$326 billion USD annually, an increase of 38% in the last decade [3]. Many people with MDD do not  
56 receive treatment, with one in three people with active symptoms failing to receive care [1]. Further,  
57 MDD is frequently misdiagnosed by primary care, which is often the first point of contact for those with  
58 clinical symptoms [4].

59 MDD presents considerable challenges to effective diagnosis and management, due, in part, to its  
60 dynamic nature and variable trajectory [5, p. 2]. The longitudinal course of MDD, as described by the  
61 DSM-5, allows for considerable variability across persons, such that some individuals may experience  
62 only discrete episodes separated by long periods of remission, while others experience chronic,  
63 unrelenting symptoms over years [6, p. 5]. Research to date has explored person-to-person differences in  
64 depression course and variability over time, with empirical evidence for heterogeneity in symptom  
65 trajectory [7]–[9], as well as difficulty in predicting longitudinal course [10]. These findings suggest that  
66 cross-sectional *severity* (“level of depression”) and *presence* (“depressed vs. not depressed”) outcomes  
67 alone, while providing informative “snapshots” in time, are insufficient for understanding the naturalistic  
68 course of MDD, and thus, the core nature of MDD.

69 We posit that depression symptom variability, *per se*, is an important outcome, which has  
70 meaningful basic science and translational implications. For the purpose of our study, we define  
71 *depression symptom variability* to mean the degree of within-person variation in reported depression  
72 symptom severity across time. Indeed, research to date examining depression temporal dynamics

## DETECTION OF NATURALISTIC DEPRESSION SYMPTOM VARIABILITY

73 (Nemesure et al., 2022), has revealed considerable within and between person symptom variability over  
74 time. We provide a theoretical and empirical basis for the importance of depression symptom variability  
75 as an outcome. First, variability is important to explore as a core metric of depression's naturalistic,  
76 longitudinal course. Together with other summative longitudinal metrics, such as mean severity,  
77 variability provides an important summary of depression's longitudinal course. Depression symptom  
78 variability is a necessary precondition for relapse and remission (i.e., major depressive episodes), which  
79 are important outcome and prognostic markers in MDD [6, p. 5]. Further, depression temporal variability  
80 may help to inform diagnostic distinctions, such as that between MDD and Persistent Depressive Disorder  
81 (PDD), with the latter theoretically showing less long term temporal variability than the former as well as  
82 more severe functional impairment [11]. Therefore, a nuanced understanding of depression's course,  
83 including an understanding of those factors associated with symptom variability, is fundamental to  
84 effective assessment and management. A highly variable course, for instance, would require more  
85 frequent assessments to accurately describe the disorder trajectory, and likely more temporally dynamic  
86 interventions.

87         Second, depression symptom variability has been associated with important clinical, prognostic  
88 and treatment outcomes. Specifically, higher depression symptom variability has been positively  
89 associated with (i) a higher risk of suicide attempts [12], (ii) lower family functioning (in maternal  
90 depression) [13], cognitive decline [14], and (iv) pathological narcissism [15] (an important prognostic  
91 marker for mental health treatment) [16]. Depressed mood variability has also been shown to interact with  
92 perceived self-esteem instability in predicting future depression at six-month follow-up [17], and a  
93 *variable*, chronic depression course has been associated with all-cause mortality in older adults [18].  
94 Additionally, rapid symptom fluctuation in depressed people has been associated with involvement in  
95 violence [19]. Given these impactful clinical and prognostic associations, it is of considerable importance

## DETECTION OF NATURALISTIC DEPRESSION SYMPTOM VARIABILITY

96 to understand naturalistic depression symptom *variability*, including the personalized features which may  
97 contribute to a fluctuating course.

98         Of important transdiagnostic consideration, there is face validity that depression variability may  
99 have a relation to affective instability, the latter of which has been studied in relation to depression  
100 utilizing repeat assessment of both high and low-arousal negative affect features [20]; low-arousal  
101 negative affect features (e.g., ‘tired’, ‘bored’, ‘droopy’) [21] have considerable overlap with the core  
102 neurovegetative depressive symptoms including low energy, depressed mood, and reduced interest [6, p.  
103 5]. Thus it may be a reasonable assumption that affective instability may be at least partially explained by  
104 temporal depression variability, and therefore understanding depression variability may help in  
105 understanding affective instability, which is also an important consideration in borderline personality and  
106 bipolar disorders [22].

107         Machine learning methods, operating on highly dimensional datasets, have shown great promise  
108 in modeling important clinically-relevant outcomes in MDD [23]–[25]. Advances in computing power  
109 and passive data streaming have made possible the application of ecologically-valid, person-generated  
110 health data (e.g., sleep, movement) to personalized depression models [23], complementing more  
111 traditional demographic features. Price et al. for example, utilized actigraphy data to effectively detect  
112 MDD presence in a large cohort [26]. Naturalistic movement and sleep data are promising candidates for  
113 modeling MDD symptom variability, given their established relationship to major depressive episodes  
114 and their capacity for predicting depression severity [27], [28]. In particular, sleep and movement  
115 problems are core features of depression [6, p. 5], and sleep problems are a known risk factor for  
116 depression recurrence [29], a plausible driver of long-term symptom variability. Additionally, such  
117 passively-collected features have contributed to empirical support for MDD-associated (1) sleep and  
118 circadian rhythm irregularities [30], [31], (2) reduced locomotion [32], and (3) reduced daily activity [33].  
119 These efforts inform our understanding of features associated with depression *presence* and *severity*, and

## DETECTION OF NATURALISTIC DEPRESSION SYMPTOM VARIABILITY

120 thereby serve as a benchmark for identifying biodemographic and behavioral characteristics that may also  
121 have an association with long-term depression symptom *variability*.

122 To build upon efforts by Makhmutova et al. (2021, 2022) in the development of the Prediction of  
123 Severity-Change Depression (PSYCHE-D) model and data source [34], [35], the present work leveraged  
124 a stacked ensemble machine learning approach applied to baseline biodemographic (i.e.,  
125 sociodemographic and comorbidity) features and objective, wearable passively-collected movement and  
126 sleep data, to explore factors associated with long-term depression symptom *variability*.

127 Methodologically, our work is unique in our direct model comparisons on the basis of feature-selection  
128 and feature-type. First, we compared a model trained on theory-informed feature selection against a  
129 parallel model trained on an exhaustive feature set. Second, we compare a model trained on baseline  
130 demographic features to a parallel model trained on passively-derived sleep and activity features. Further  
131 we examine the incremental predictive gain when combining both types of features; for all models we  
132 utilize a robust stacked ensemble approach. We hypothesized that (1) features having known association  
133 with depression presence and severity would also associate with long-term symptom variability. Further,  
134 (2) we hypothesized that biodemographic and objective passively-collected movement and sleep data  
135 each contain complementary information and, thus, when combined would produce improved model  
136 prediction compared to either singular information modality, as accounting for complementarity during  
137 feature selection has been shown to increase model performance [36], [37]. To test our hypotheses, we  
138 used 12-month longitudinal data [38] comprising personal biodemographic data, movement and sleep  
139 metrics statistically derived from passively-collected wearable accelerometry data, and quarterly PHQ-9  
140 scores. A cross-validation framework, coupled with a stacked ensemble machine learning approach, was  
141 implemented to model depression symptom variability using features with empirical associations with  
142 depression. For model interpretability, we used an algorithmic approach to quantify the relative

## DETECTION OF NATURALISTIC DEPRESSION SYMPTOM VARIABILITY

143 importance and directionality of biodemographic features, statistical movement and sleep features, and  
144 both in concert for predicting depression symptom variability.

### 145 **Methods**

#### 146 **Study Sample**

147  
148 The present work used publicly-available biodemographic, wearable passively-collected  
149 movement and sleep, and depression symptom data originally collected over a 12-month period provided  
150 in the PSYCHE-D dataset [39], which was captured as part of the DiSCover Project developed by  
151 Evidation Health [38]. Participants were originally recruited via Achievement, a community of adults in  
152 the United States that can connect consumer-grade fitness applications and wearable (e.g., Fitbit, Garmin)  
153 to the study platform. Participant inclusion was limited in the present analyses to individuals with twelve  
154 consecutive months of objective accelerometer information, reflecting non-missing values for some or all  
155 of the related movement and sleep metrics for each month, and a reported Patient Health Questionnaire-9  
156 (PHQ-9) [40] composite score completed at baseline and every subsequent three-month time point for the  
157 12-month study period ( $N = 939$ , 70.61% female, 29.39% male,  $age_{mean} = 42.55 \pm 10.23$ , 91.37% White,  
158 4.69% Black, 4.05% Hispanic, 2.66% Asian, 2.23% Race not specified, 10.81% required financial  
159 assistance from the government) (See **Figure 1**). A full description of the original DiSCover Project study  
160 design, recruitment protocols, and participant baseline demographic information is provided by Lee et al  
161 [38].

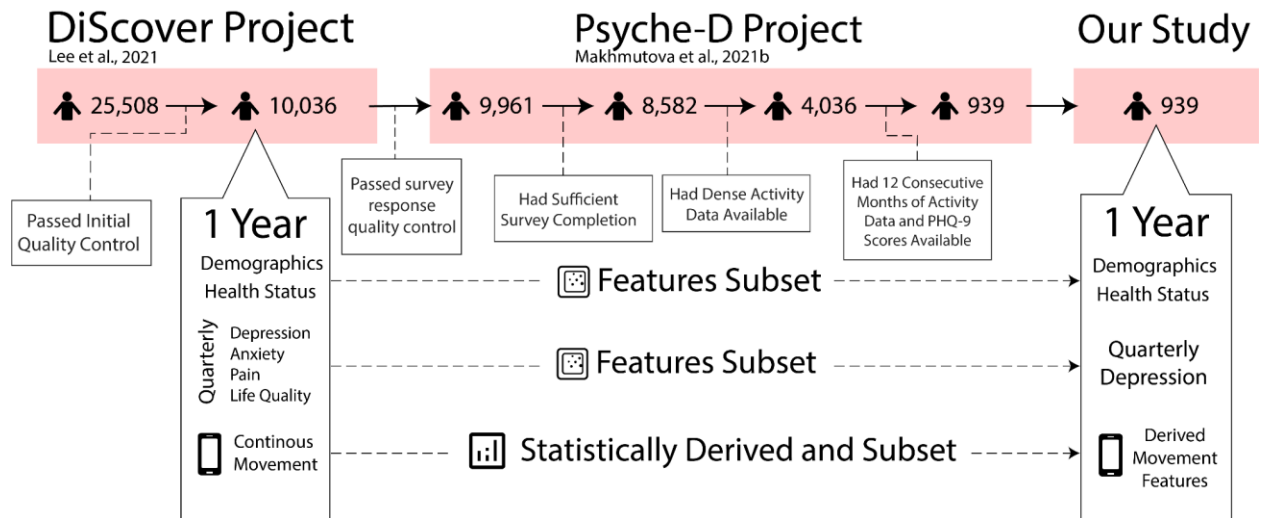
162

163

164

# DETECTION OF NATURALISTIC DEPRESSION SYMPTOM VARIABILITY

165 **Figure 1.**



166

167

168

169



## DETECTION OF NATURALISTIC DEPRESSION SYMPTOM VARIABILITY

### 170 **Study Measures**

171           The original PSYCHE-D dataset contains 150 person-generated health data (PGHD) features  
172 reflecting baseline biodemographic information, derived passively-collected movement and sleep  
173 information, and Patient Health Questionnaire-9 (PHQ-9) composite scores ( $PHQ-9_{mean} = 6.80 \pm 5.72$ ;  
174 42.79% No Depressive Symptoms, 28.78% Mild Depressive Symptoms, 17.61% Moderate Depressive  
175 Symptoms, 7.41% Moderately Severe Depressive Symptoms, 3.41% Severe Depressive Symptoms) [39];  
176 a common screening tool for MDD [41] consisting of nine items which reflect the degree to which each  
177 item was bothersome over the last two weeks (e.g., feeling down, depressed, or hopeless) [40].  
178 Makhmutova et al. describes the PGHD feature collection and processing in further detail [34]. The  
179 dataset was subset for the present analyses to 20 features consisting of a combination of eight baseline  
180 biodemographic (i.e., Sex, Race, BMI, Pregnancy Status, Money Assistance, Comorbid Diabetes Type I,  
181 Comorbid Diabetes Type II, Comorbid Migraines), and twelve derived passively-collected movement and  
182 sleep data (i.e., Average Awake Activity, Low Physical Activity Duration, Moderate-to-Vigorous  
183 Activity Duration, Active Day Count, Sedentary Day Count, Nighttime Sleep Variability, Average  
184 Weekday Sleep, Average Weekend Sleep, Sleep Start Time, Variability In Sleep Start Time, Weekly  
185 Hypersomnia Count, Weekly Hyposomnia Count). These features were chosen based on known direct or  
186 indirect associations with depression, outlined in **Supplementary Table 1**, as feature engineering and  
187 selection informed by domain knowledge has been shown to improve predictive performance and model  
188 interpretability [42].

### 189 **Data Preprocessing**

190           All data pre-processing was performed in R (v 4.0.2) [43]. Baseline biodemographic feature data  
191 types were interrogated and converted according to their reporting structure (e.g., Migraine comorbidity  
192 was converted from numerical to categorical). To account for missingness of certain biodemographic and

## DETECTION OF NATURALISTIC DEPRESSION SYMPTOM VARIABILITY

193 movement and sleep-related metrics, multivariate imputation by chained equations (*mice*) with predictive  
194 mean matching was implemented using the *mice* package in R [44], as *mice* is well-suited to handling high  
195 proportions of missing data, and captures the uncertainty associated with approximating missing  
196 information [45]. Across all participants, 0.08% of the subsetting biodemographic information was missing,  
197 and 15.64% of the subsetting passively-collected movement and sleep-related metrics information was  
198 missing. Resultantly, five imputed datasets were generated, reflecting the plausible distribution of missing  
199 information, and used for subsequent analyses. Following imputation, summative metrics of the  
200 longitudinal passive-collected movement and sleep features were derived to represent the average and  
201 variability of each selected feature across the twelve-month data collection period. Average was calculated  
202 as the mean of the feature's values, and variability was calculated as the root mean square of successive  
203 differences (RMSSD) of the respective feature. The summative features were derived to reflect longitudinal  
204 movement and sleep behaviors, as well as avoid a nested data structure, such that each participant could be  
205 represented as a single row with their fixed baseline biodemographic features and their statistically-derived  
206 movement and sleep features. To interrogate the naturalistic fluctuation in *sequential* depressive symptoms  
207 across a twelve-month period, the RMSSD of depressive symptom change was calculated. As previously  
208 stated, individuals' composite PHQ-9 scores collected at months 0, 3, 6, 9, and 12 were used to calculate  
209 variability in depressive symptoms (RMSSD). Thus, an individual's  $PHQ-9_{RMSSD}$  represented a single  
210 metric of depressive symptom variability that captured fluctuation in symptom expression across the entire  
211 study. Additionally,  $PHQ-9_{RMSSD}$  was correlated with mean PHQ-9 score to establish that  $PHQ-9_{RMSSD}$  was  
212 not simply a proxy for depression symptom intensity ( $r = 0.54$ ,  $R^2 = 0.29$ ).

### 213 **Machine Learning Modeling Approach**

214 The present analyses was completed in Python (v 3.9) [46], and followed a 3-fold cross-validation  
215 framework (80%), allowing for a within-sample completely held-out test set (20%) to quantify predictive  
216 performance [47], and providing an efficacious approach in allowing for unbiased performance estimates

## DETECTION OF NATURALISTIC DEPRESSION SYMPTOM VARIABILITY

217 in machine learning modeling [48]. Specifically, a stacked ensemble machine learning approach was used  
218 across the five MICE-generated datasets to assess for predictive robustness across the plausible  
219 imputation distribution. Stacked ensemble machine learning approaches have shown the capacity to  
220 consistently outperform base algorithms in detecting depression [49], by leveraging algorithmically  
221 distinct machine learning models (e.g., linear models, tree based models) to individually train on the data.  
222 The individual model predictions are subsequently used as inputs to a final ‘meta’ model, which returns a  
223 consensus score. The stacked ensemble algorithms and hyperparameters implemented for the present  
224 analysis are provided in **Supplementary Table 2**. Additionally, the cross-validation architecture and  
225 random seed chosen for splitting the data was standardized across the three models (baseline  
226 biodemographic model; passively-collected movement and sleep model; composite model) to reflect  
227 consistency across the model progression. Further, an exhaustive feature-inclusion approach was  
228 implemented, where all originally collected features were incorporated or transformed for the three  
229 respective model types (See **Table 1A** and **Table 1B**) to evaluate performance with an increased feature  
230 space.

### 231 **Model Performance**

232 Model performance was reported for the validation and held-out test set for each of the machine  
233 learning models as the mean and standard deviation across the five MICE-imputed data sets for  
234 correlative strength ( $r$ ), and normalized mean absolute error ( $MAE_{norm}$ ). The  $MAE_{norm}$  reflects an  
235 outcome-agnostic representation of the model’s mean absolute error by dividing the mean absolute error  
236 by the range of the observed outcome, and thus represents the mean percentage error of the prediction.

### 237 **Model Introspection**

238 To assess the most influential features for model prediction across the three models, SHapley  
239 Additive exPlanations (SHAP) was implemented, and the top five most influential features were reported

## DETECTION OF NATURALISTIC DEPRESSION SYMPTOM VARIABILITY

240 for each model. SHAP provides a method for model introspection by iteratively perturbing the input  
241 features and assessing how this affects the model prediction [50]. Thus, SHAP provides a mechanism for  
242 determining feature importance, as well as the marginal contribution of each input variable to the model's  
243 prediction at the individual level, represented as the individual values positioning on the x-axis of **Figure**  
244 **2**. Specifically, an individual features SHAP values can be interpreted as the features' partial association  
245 with the outcome when controlling for all other input features in the model. Collectively, SHAP can  
246 estimate the relative magnitude of a feature's influence on a model's predictions, directional relationships  
247 between features and predicted outcomes, as well as different order interactions between features.

### 248 Code Availability

249 The data that support the findings of this study are available from the corresponding author, GP,  
250 upon reasonable request.

## 251 Results

### 252 Baseline Biodemographic Features

#### 253 *Baseline Biodemographic Modeling Results*

254  
255 Baseline biodemographic features were incorporated into a stacked ensemble machine learning  
256 approach to detect depression symptom variability (PHQ-9<sub>RMSD</sub>) (**Supplementary Table 1**). Averaged  
257 across the five MICE-imputed datasets, we found a weak, positive correlation ( $r = 0.27 \pm 0.00$ ,  $MAE_{norm}$   
258  $0.14 \pm 0.00$ ; see **Tables 1A & 1B**) between predicted long-term depression symptom variability outcomes  
259 and actual long-term depression symptom variability outcomes in the held-out test set (See **Figure 2A**).

## DETECTION OF NATURALISTIC DEPRESSION SYMPTOM VARIABILITY

### 260 *Relative Feature Importance and Directionality for the Baseline Biodemographic Model*

261           Using SHAP (see Methods section 2.5), we found comorbid migraines to be the most influential  
262 feature in the model’s prediction of higher depression symptom variability, followed by female sex, high  
263 body mass index (BMI), required financial assistance, and non-White race (See **Figure 2A**,  
264 **Supplementary Table 1**).

### 265 **Passively-Collected Movement and Sleep Features**

#### 266 *Passively-Collected Movement and Sleep Modeling Results*

267  
268           Statistically-derived features from wearable, passively-collected movement and sleep data  
269 (**Supplementary Table 1**) were incorporated into a stacked ensemble machine learning model to detect  
270 depression symptom variability (PHQ-9<sub>RMSSD</sub>). Similar to the biodemographic model, when averaged  
271 across the five MICE-imputed datasets, we found a weak, positive correlation ( $r = 0.27 \pm 0.01$ ,  $MAE_{norm}$   
272  $0.14 \pm 0.00$ ; see **Tables 1A & 1B**) between predicted long-term depression symptom variability outcomes  
273 and actual long-term depression symptom variability outcomes in the held-out test set (See **Figure 2B**).

### 274 *Relative Feature Importance and Directionality for the Passively-Collected Movement and Sleep Model*

275  
276           Using SHAP (see Methods section 2.5), we found (1) high weekday sleep duration, (2) high count  
277 of nights with less than five hours asleep (hyposomnia) in the last week, (3) lower recent step count, (4)  
278 high range of sleep duration, and (5) low weekend sleep duration to be the top five most influential  
279 features in the model’s prediction of high depression symptom variability (See **Figure 2B**,  
280 **Supplementary Table 1**). The top five features in the passively-collected movement and sleep reflect an  
281 average over twelve months.

**282 Combined Biodemographic and Passively-Collected Movement and Sleep Features***283 Biodemographic and Passively-Collected Movement and Sleep Modeling Results*

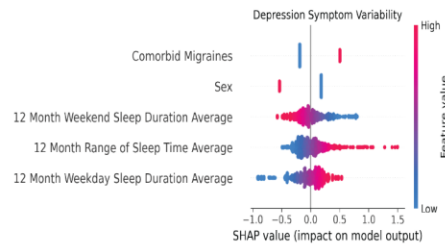
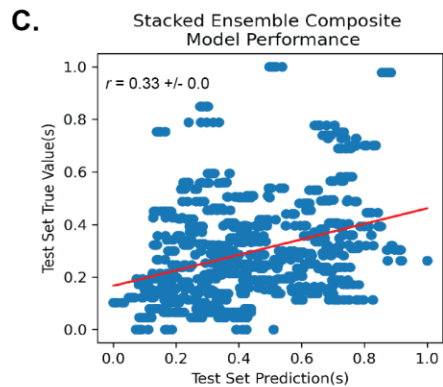
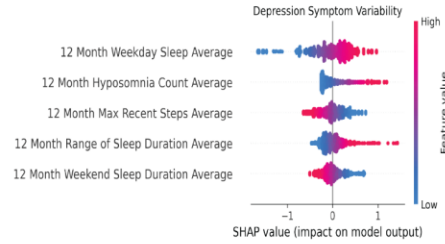
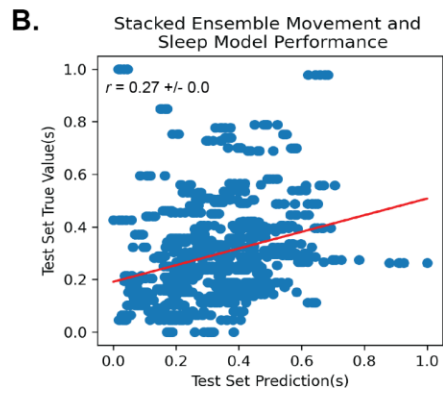
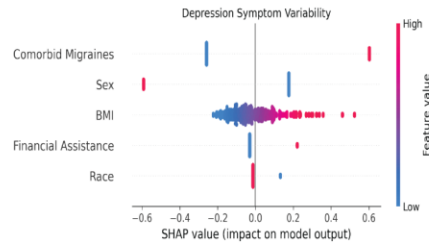
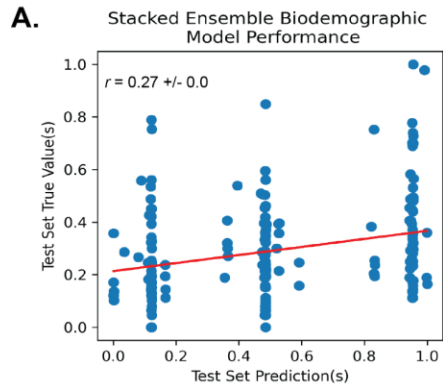
284  
285           Using a composite model of baseline biodemographic features (see Results section 3.1) and  
286 statistically-derived features from wearable passively-collected movement and sleep data (see Results  
287 section 3.2) we found a moderate, positive correlation ( $r = 0.33 \pm 0.01$ ,  $MAE_{norm} 0.14 \pm 0.00$ ; see **Tables**  
288 **1A & 1B**) between predicted depression score variability outcomes and actual depression score variability  
289 outcomes in the held-out test set (See **Figure 2C**).

*290 Relative Feature Importance for the Combined Biodemographic and Passively-Collected Movement and*  
*291 Sleep Model*

292  
293           Using SHAP (see Methods section 2.5), we identified (1) comorbid migraines to be most  
294 influential in the models prediction of high depression symptom variability ( $PHQ-9_{RMSSD}$ ), followed by (2)  
295 female sex, (3) lower duration of weekend sleep, averaged over 12-months, (4) higher range of time  
296 asleep, averaged over 12-months. and (5) higher duration of weekday sleep, averaged over 12-months  
297 (See **Figure 2C, Supplementary Table 1**).

# DETECTION OF NATURALISTIC DEPRESSION SYMPTOM VARIABILITY

298 **Figure 2.**



299

300

301 **Exhaustive Feature-Inclusion Modeling Results**

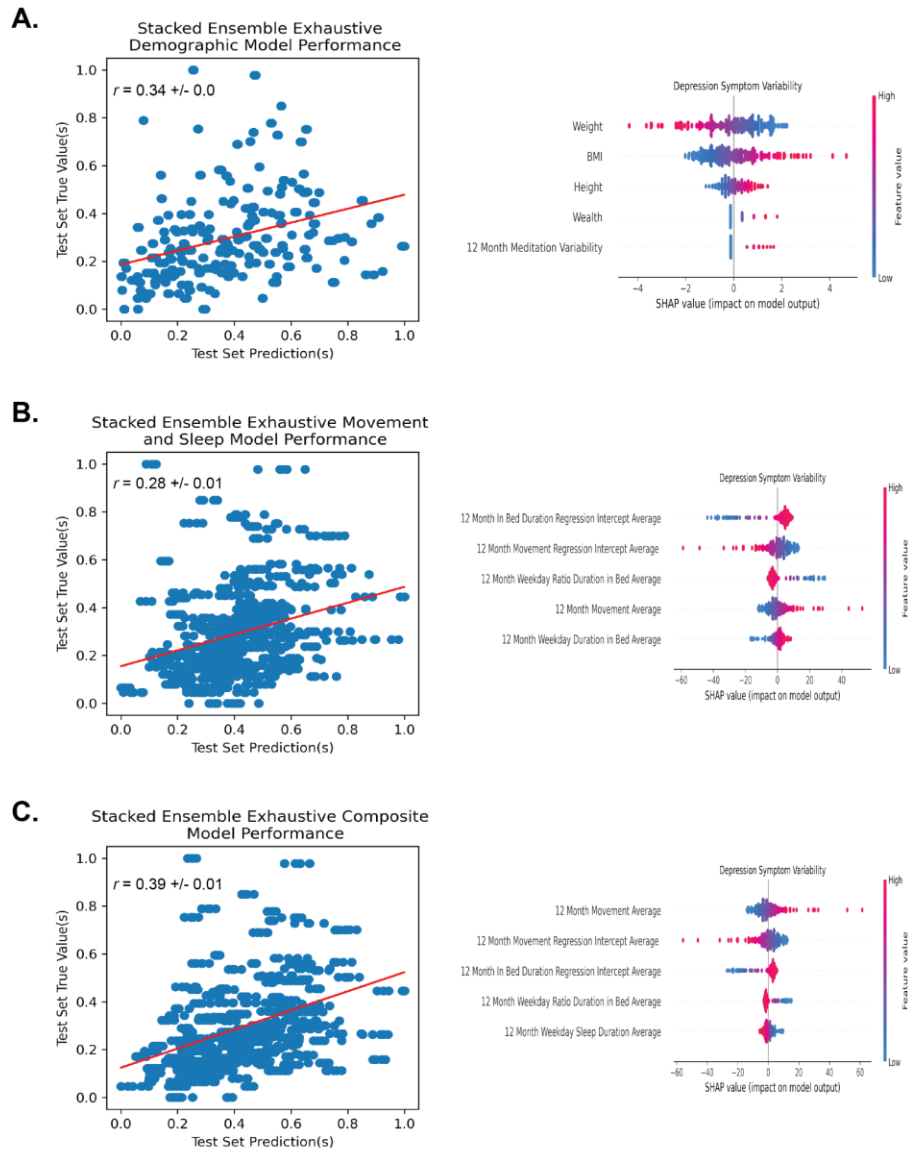
302  
303           Complementing the decision to subset biodemographic and passively-collected movement and  
304 sleep features using theoretical and empirical domain knowledge, we also constructed three parallel  
305 stacked ensemble machine learning models operating on the non-subsetted PSYCHE-D [39] feature set,  
306 including 49 original and statistically-derived biodemographic features, and 222 statistically-derived  
307 movement and sleep features. The exhaustive feature-inclusion approach showed marginal performance  
308 improvement compared to the theory-driven variable selection approach across the three model types (see  
309 **Tables 1A & 1B**, and **Figure 3**). Nevertheless, the exhaustive inclusion of all previously collected  
310 features introduced increased model complexity and reduced featured interpretability.

311



# DETECTION OF NATURALISTIC DEPRESSION SYMPTOM VARIABILITY

312 **Figure 3**



313  
314  
315  
316

DETECTION OF NATURALISTIC DEPRESSION SYMPTOM VARIABILITY

317 **Table 1A.**  
318

Modeling Approach	Model 1: Demographic Information Model			Model 2: Movement and Sleep Information Model			Model 3: Composite Model		
	Variable Number	$r \pm SD$ (Test Set)	$r \pm SD$ (Validation Set)	Variable Number	$r \pm SD$ (Test Set)	$r \pm SD$ (Validation Set)	Variable Number	$r \pm SD$ (Test Set)	$r \pm SD$ (Validation Set)
Theory-Organized Stacked Ensemble	8	0.27 ± 0.00	0.27 ± 0.03	24	0.27 ± 0.01	0.22 ± 0.08	32	0.33 ± 0.01	0.31 ± 0.04
Full Variable Set Stacked Ensemble	49	0.34 ± 0.00	0.35 ± 0.09	222	0.28 ± 0.01	0.25 ± 0.07	271	0.39 ± 0.01	0.35 ± 0.08

319  
320  
321 **Table 1B.**  
322

Modeling Approach	Model 1: Demographic Information Model			Model 2: Movement and Sleep Information Model			Model 3: Composite Model		
	Variable Number	$MAE_{norm} \pm SD$ (Test Set)	$MAE_{norm} \pm SD$ (Validation Set)	Variable Number	$MAE_{norm} \pm SD$ (Test Set)	$MAE_{norm} \pm SD$ (Validation Set)	Variable Number	$MAE_{norm} \pm SD$ (Test Set)	$MAE_{norm} \pm SD$ (Validation Set)
Theory-Organized Stacked Ensemble	8	0.14 ± 0.00	0.11 ± 0.01	24	0.14 ± 0.00	0.11 ± 0.00	32	0.14 ± 0.00	0.11 ± 0.00
Full Variable Set Stacked Ensemble	49	0.13 ± 0.00	0.11 ± 0.00	222	0.14 ± 0.00	0.11 ± 0.00	271	0.13 ± 0.00	0.11 ± 0.00

323  
  
324  
  
325  
  
326

327

## Discussion

### 328 General Overview

329 The present results demonstrate the successful application of both biodemographic and passively-  
330 collected movement and sleep features for modeling the novel outcome, long-term depression symptom  
331 variability. We found moderate predictive capacity of the biodemographic and passively-collected  
332 movement and sleep features for long-term depression symptom variability detection when used in  
333 concert. This validates our hypothesis (1) of features indicative of depression severity also indicative of  
334 depression symptom variability and (2) the predictive utility of complementarity (i.e., unique information)  
335 between feature types. Regarding our theory-guided subsetting approach, we found modest improvements  
336 in predictive performance using a non-subset feature set with an increase in model complexity (see  
337 **Tables 1A & 1B**, and **Figure 3**).

### 338 Implications and Importance

339 The successful application of the biodemographic and passively-collected movement features  
340 used in the present analysis to detect depression symptom variability has promising mental health clinical  
341 implications, strengthening evidence for more objective and naturalistic assessments, with less burden to  
342 patients [51]. The work also validates our hypothesis of variables empirically correlated with major  
343 depressive disorder (e.g., sex, migraines, sleep disturbances) also having association with depression  
344 symptom variability. While biomarkers of depression *severity* have been studied more extensively, factors  
345 associated with depression symptom variability have had relatively less attention.

346 In this work, we make the case for (1) variability, *per se*, as an outcome of high importance, as  
347 well as (2) the importance and utility of predicting who is likely to have high variability. First, variability  
348 has been linked to important outcomes, including suicide attempts in high risk individuals [12], as well as  
349 family functioning in the case of maternal depression [13]. Thus, symptoms variability, itself, may be a

## DETECTION OF NATURALISTIC DEPRESSION SYMPTOM VARIABILITY

350 risk factor for important clinical outcomes. Second, long-term symptom variability is a necessary  
351 precondition for episodic depression relapse and remission. Relapse and remission counts have obvious  
352 importance as clinical outcomes by themselves, and have been associated with poorer long-term  
353 prognosis in MDD [52], [53]. Third, predicting person-level variability has implications for personalized  
354 medicine [54] approaches to mental healthcare. Identifying *who* is likely to have higher symptom  
355 variability over time, would allow for person-tailored assessment frequencies. For instance, a person with  
356 high depression symptom variability would require more frequent depression assessments compared to  
357 someone with lower depression symptom variability to adequately capture the disorder course over time.

### 358 **Model Introspection and Depression Symptom Variability Theory**

359         The presence of migraines was the most influential of the biodemographic features for predicting  
360 depression symptom variability and remained so even when combined with statistically derived passively-  
361 collected movement and sleep features (See **Figure 2**). Migraines have been established as highly  
362 comorbid with depression [55], [56]; additionally, research has demonstrated that migraines may perturb  
363 the naturalistic course of depression, prolonging the time to depression remission [57]. However, the  
364 direct relationship of migraines to depression symptom variability is not well understood. A plausible  
365 explanation stems from research demonstrating depression exacerbation in concurrence with migraine  
366 headache onset (a phenomenon reported in nearly one third of a depressed sample) [58]. Given the  
367 discrete and episodic nature of migraine headaches, [59] as well as the empirical support for simultaneity  
368 in migraine onset and depression exacerbation, it would follow that such patients would show heightened  
369 variability in their depression over time.

370         Following migraines, the next most influential features for modeling depression symptom  
371 variability in the biodemographic model included: (i) female sex, (ii) high BMI, (iii) required financial  
372 assistance, and (iv) non-White race. These findings may be contextualized in research to date, which  
373 demonstrated females had a considerably higher rate of depressive episodes [60], with higher frequency,

## DETECTION OF NATURALISTIC DEPRESSION SYMPTOM VARIABILITY

374 theoretically, serving as a proxy for variability. Further, required financial assistance may be a proxy for  
375 lower socioeconomic status, a known correlate of depression [61]; specific to *variability*, a large  
376 longitudinal cohort study ( $N = 12,650$ ) showed socioeconomic status predicted long-term patterns of  
377 change in intra-individual depression symptom variability [62]. However, it is also important to consider  
378 that markers of variability in depression, such as race and sex, could also be markers for events such as  
379 racism and discrimination, which may, themselves, have an episodic course [63]. While racism and  
380 discrimination have been shown to predict depressive symptoms, longitudinally [64], discriminatory  
381 events have also been shown to cause acute exacerbations in depression [65]. Such depression “spikes”  
382 over time may appear to be of a more variable course.

383 Movement and sleep features derived from passively-collected actigraphic data demonstrated  
384 capacity for modeling depression symptom variability. Sleep behaviors were highly represented among  
385 the most influential features in the movement and sleep model, as well as the composite model (See  
386 **Figure 2B, 2C**). Specifically, sleep duration (for both weekends and weekdays), range of sleep duration,  
387 and nights spent with hyposomnia were the most influential sleep-related features. These findings are  
388 generally consistent with well-established knowledge of the close relationship between sleep, activity, and  
389 depression [6], [66], validated with passively collected, objective data [32]. Notably, sleep quality and  
390 duration have bidirectional associations with psychosocial functioning amongst young adults [67].  
391 Moreover, short sleep duration and poor sleep quality are associated with a higher prevalence of  
392 depressive symptoms among university students [68]. This suggests a complex relationship between sleep  
393 and depression that is not merely unidirectional, but rather complicated by biopsychosocial variables.

394 Further, specific sleep profiles have been empirically correlated with longitudinal depression  
395 symptom variability [69], perhaps suggesting the existence of sleep markers for MDD variation.  
396 Curiously, sleep quality correlates more strongly with psychosocial functioning than sleep duration  
397 among young adults [67]. Our findings, *range of sleep duration*, *nights with hyposomnia*, and *sleep*

## DETECTION OF NATURALISTIC DEPRESSION SYMPTOM VARIABILITY

398 *duration*, may be further contextualized in research linking similar features (i.e., total sleep time and day-  
399 to-day variability in total sleep time) to next-day mood and depressive symptoms [70]. It follows that  
400 changes in mood may track with changes in sleep; thus, a higher range of nightly sleep duration would  
401 imply a wider range of depression severity. Recognizing the multifactorial nature of sleep, optimizing  
402 sleep architecture, quality, and duration collectively, yet intricately, influences depression outcomes. Both  
403 insufficient and excessive sleep durations have been shown to elevate depression risk [71], [72], with the  
404 latter being particularly pertinent when coupled with sustained poor sleep quality. Factors such as  
405 emotional exhaustion and stress, whether stemming from academic demands [73] or shift work [74],  
406 further complicate the intricate relationship between sleep and depression.

407         Recall that, in addition to a feature subsetting approach, guided by a priori domain knowledge, we  
408 comparatively tested an exhaustive feature set approach, using *all* biodemographic and *all* movement and  
409 sleep features (See **Figure 3C**). Despite the reduced interpretability of such a model, conferred by the  
410 inclusion of statistical features which are more convoluted, there is a modest increase in performance ( $r =$   
411  $0.39$ , compared to  $r = 0.33$  with reduced feature model), highlighting the utility and application of such an  
412 approach for a performance-driven task. In contrast to the domain-driven approach, the top five most  
413 influential features in the exhaustive-feature model were all derived from passively-collected movement  
414 and sleep data – none from biodemographic information or self-report. Notably, a subset of these features  
415 were generated from regression-based statistics on the passively-collected movement and sleep data [34],  
416 which have not been established in the literature on long-term depression symptom variability, but do  
417 seem to offer a substantive increase in information for the model's predictions, allowing for increased  
418 model performance. These findings suggest further consideration into the utility of feature engineering as  
419 it pertains to passively-collected movement and sleep data, as it offers clear advantages for tasks strictly  
420 concerned with improving predictive performance relating to long-term depression symptom variability.

### 421 **Strengths, Limitations, and Future Directions**

422           **The current study uniquely utilized long-term depression variability as an outcome**  
423 **measure. Additionally, our methods allow for a direct comparison between feature selection**  
424 **strategies, specifically theory-informed versus exhaustive, and between feature types, specifically**  
425 **passive sensing-derived features and baseline demographic features. A significant strength of our**  
426 **work lies in our application of a robust stacked ensemble approach, accommodating the potentially**  
427 **complex relationships among features.** Despite the strengths and novelty of our work, the study results  
428 must be considered in the context of several important limitations, described here. (1) The study  
429 population was limited in demographic diversity, and future research would benefit from analyzing a  
430 more nationally-representative sample when detecting depression symptom variability. Further, a  
431 consideration for depression symptom variability within demographic groups (e.g., gender, race) should  
432 be assessed, as influential biodemographic and passively-collected movement and sleep features are likely  
433 differentially expressed between populations, which would allow for more effective personalized  
434 treatment. (2) Recall that the outcome ( $PHQ-9_{RMSSD}$ ) is derived from self-reported PHQ-9 scores at three  
435 month intervals over the course of one year. As such, the temporal resolution of depression symptom  
436 variability is limited. A related but distinct limitation inherent in the original study design is the mismatch  
437 between the 2-week look-back period of the PHQ-9 and the 3 month interval at which the measurements  
438 were collected. In future research investigating depression symptom variability, ecological momentary  
439 assessments for depressive symptoms would be preferable. (4) Finally, the choice of one year over which  
440 to measure variability has important implications in the applicability and interpretation of results. While  
441 one year is likely sufficient to capture a single depressive episode [75], it may be insufficient to capture  
442 the temporal dynamics across multiple depressive episodes. Furthermore, while the present investigation  
443 of factors associated with depression symptom variability is appropriately conducted on a community  
444 sample, given that over one third of participants (38.8%) reported PHQ-9 scores both below and above

## DETECTION OF NATURALISTIC DEPRESSION SYMPTOM VARIABILITY

445 the clinical threshold for depression ( $\text{PHQ-9} \geq 10$ ), generalizability to a clinical sample remains uncertain.  
446 Thus, a future extension of this work would be validation and comparison on a clinical sample to assess  
447 both model performance as well as features most associated with the model's predictions.

### 448 **Conclusion**

449 In the present work, we emphasize depression symptom variability as an important clinical and  
450 research variable in mental health. *Variability* represents an important attribute of the depression's  
451 longitudinal course, as well as a dimension of heterogeneity between depressed persons. In addition,  
452 depression symptom variability has been linked to important clinical outcomes, such as suicide. Though  
453 much is known of factors associated with point-in-time depression severity, relatively little is known of  
454 long-term, naturalistic variability in depression, as well as person-specific factors which associate with  
455 variability. In the present work, we explore the capacity of biodemographic and passively-collected  
456 movement and sleep information to model depression symptom variability. We find positive results to  
457 suggest association between both biodemographic and passively-collected data types, independently, as  
458 well as evidence of complementarity in predictive capacity. Our work provides an early step toward the  
459 complementary, personalized use of unobtrusive data types in addressing the question of depression's  
460 temporal variability.

461

462

463

464



465

**Author Contributions**

466

GP, MH, SS, MN, and NJ contributed to conceptualization, methodology, and writing of the

467

original draft. GP and MH contributed to the validation and visualization of the analysis. GP contributed

468

to the formal analysis. MH and NJ provided supervision to the present work.

469

470

**Conflicts of Interest**

471

The author(s) declare that there were no conflicts of interest with respect to the authorship or the

472

publication of this article.

473

**Funding**

474

Funding Statement: This work was supported by the National Institute of Mental Health (NIMH)

475

and the National Institute of General Medical Sciences (NIGMS) (grant number 1 R01 MH123482-01).

476

477

478

## References

- 479 [1] NSDUH, “2020 National Survey of Drug Use and Health (NSDUH) Releases | CBHSQ  
480 Data.” Accessed: Mar. 23, 2022. [Online]. Available:  
481 [https://www.samhsa.gov/data/release/2020-national-survey-drug-use-and-health-nsduh-](https://www.samhsa.gov/data/release/2020-national-survey-drug-use-and-health-nsduh-releases)  
482 [releases](https://www.samhsa.gov/data/release/2020-national-survey-drug-use-and-health-nsduh-releases)
- 483 [2] T. Vos *et al.*, “Global burden of 369 diseases and injuries in 204 countries and territories,  
484 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019,” *The*  
485 *Lancet*, vol. 396, no. 10258, pp. 1204–1222, Oct. 2020, doi: 10.1016/S0140-  
486 6736(20)30925-9.
- 487 [3] P. E. Greenberg *et al.*, “The Economic Burden of Adults with Major Depressive Disorder in  
488 the United States (2010 and 2018),” *PharmacoEconomics*, vol. 39, no. 6, pp. 653–665,  
489 Jun. 2021, doi: 10.1007/s40273-021-01019-4.
- 490 [4] M. Vermani, M. Marcus, and M. A. Katzman, “Rates of Detection of Mood and Anxiety  
491 Disorders in Primary Care: A Descriptive, Cross-Sectional Study,” *Prim. Care Companion*  
492 *CNS Disord.*, Apr. 2011, doi: 10.4088/PCC.10m01013.
- 493 [5] L.-S. Chen, W. W. Eaton, J. J. Gallo, and G. Nestadt, “Understanding the heterogeneity of  
494 depression through the triad of symptoms, course and risk factors: a longitudinal,  
495 population-based study,” *J. Affect. Disord.*, vol. 59, no. 1, pp. 1–11, Jul. 2000, doi:  
496 10.1016/S0165-0327(99)00132-9.
- 497 [6] A. P. A. American Psychiatric Association, *Diagnostic and Statistical Manual of Mental*  
498 *Disorders (DSM-5)*. Arlington, Virginia: American Psychiatric Association, 2013.
- 499 [7] N. Kennedy, R. Abbott, and E. S. Paykel, “Longitudinal syndromal and sub-syndromal  
500 symptoms after severe depression: 10-year follow-up study,” *Br. J. Psychiatry*, vol. 184,  
501 no. 4, pp. 330–336, Apr. 2004, doi: 10.1192/bjp.184.4.330.
- 502 [8] K. L. Musliner, T. Munk-Olsen, W. W. Eaton, and P. P. Zandi, “Heterogeneity in long-term  
503 trajectories of depressive symptoms: Patterns, predictors and outcomes,” *J. Affect. Disord.*,  
504 vol. 192, pp. 199–211, Mar. 2016, doi: 10.1016/j.jad.2015.12.030.
- 505 [9] W. A. van Eeden, A. M. van Hemert, I. V. E. Carlier, B. W. Penninx, and E. J. Giltay,  
506 “Severity, course trajectory, and within-person variability of individual symptoms in patients  
507 with major depressive disorder,” *Acta Psychiatr. Scand.*, vol. 139, no. 2, pp. 194–205,  
508 2019, doi: 10.1111/acps.12987.
- 509 [10] J. L. Rushton, M. Forcier, and R. M. Schectman, “Epidemiology of depressive symptoms in  
510 the National Longitudinal Study of Adolescent Health,” *J. Am. Acad. Child Adolesc.*  
511 *Psychiatry*, vol. 41, no. 2, pp. 199–205, Feb. 2002, doi: 10.1097/00004583-200202000-  
512 00014.
- 513 [11] E. Schramm, D. N. Klein, M. Elsaesser, T. A. Furukawa, and K. Domschke, “Review of  
514 dysthymia and persistent depressive disorder: history, correlates, and clinical implications,”  
515 *Lancet Psychiatry*, vol. 7, no. 9, pp. 801–812, Sep. 2020, doi: 10.1016/S2215-  
516 0366(20)30099-7.
- 517 [12] N. M. Melhem *et al.*, “Severity and Variability of Depression Symptoms Predicting Suicide  
518 Attempt in High-Risk Individuals,” *JAMA Psychiatry*, vol. 76, no. 6, pp. 603–613, Jun. 2019,  
519 doi: 10.1001/jamapsychiatry.2018.4513.
- 520 [13] R. Seifer, S. Dickstein, A. J. Sameroff, K. D. Magee, and L. C. Hayden, “Infant Mental  
521 Health and Variability of Parental Depression Symptoms,” *J. Am. Acad. Child Adolesc.*

## DETECTION OF NATURALISTIC DEPRESSION SYMPTOM VARIABILITY

- 522        *Psychiatry*, vol. 40, no. 12, pp. 1375–1382, Dec. 2001, doi: 10.1097/00004583-200112000-  
523        00007.
- 524        [14] B. W. Rovner, R. J. Casten, and B. E. Leiby, “Variability in Depressive Symptoms Predicts  
525        Cognitive Decline in Age-Related Macular Degeneration,” *Am. J. Geriatr. Psychiatry*, vol.  
526        17, no. 7, pp. 574–581, Jul. 2009, doi: 10.1097/JGP.0b013e31819a7f46.
- 527        [15] S. Dawood and A. Pincus, “Pathological Narcissism and the Severity, Variability, and  
528        Instability of Depressive Symptoms,” *Personal. Disord. Theory Res. Treat.*, vol. 9, pp. 144–  
529        154, Apr. 2018, doi: 10.1037/per0000239.
- 530        [16] W. D. Ellison, K. N. Levy, N. M. Cain, E. B. Ansell, and A. L. Pincus, “The Impact of  
531        Pathological Narcissism on Psychotherapy Utilization, Initial Symptom Severity, and Early-  
532        Treatment Symptom Change: A Naturalistic Investigation,” *J. Pers. Assess.*, vol. 95, no. 3,  
533        pp. 291–300, May 2013, doi: 10.1080/00223891.2012.742904.
- 534        [17] E. Franck and R. De Raedt, “Self-esteem reconsidered: Unstable self-esteem outperforms  
535        level of self-esteem as vulnerability marker for depression,” *Behav. Res. Ther.*, vol. 45, no.  
536        7, pp. 1531–1541, Jul. 2007, doi: 10.1016/j.brat.2007.01.003.
- 537        [18] S. W. Geerlings, A. T. F. Beekman, D. J. H. Deeg, J. W. R. Twisk, and W. V. Tilburg,  
538        “Duration and severity of depression predict mortality in older adults in the community,”  
539        *Psychol. Med.*, vol. 32, no. 4, pp. 609–618, May 2002, doi: 10.1017/S0033291702005585.
- 540        [19] C. L. Odgers, E. P. Mulvey, J. L. Skeem, W. Gardner, C. W. Lidz, and C. Schubert,  
541        “Capturing the Ebb and Flow of Psychiatric Symptoms With Dynamical Systems Models,”  
542        *Am. J. Psychiatry*, vol. 166, no. 5, pp. 575–582, May 2009, doi:  
543        10.1176/appi.ajp.2008.08091398.
- 544        [20] E. H. Bos, P. de Jonge, and R. F. A. Cox, “Affective variability in depression: Revisiting the  
545        inertia–instability paradox,” *Br. J. Psychol.*, vol. 110, no. 4, pp. 814–827, Nov. 2019, doi:  
546        10.1111/bjop.12372.
- 547        [21] L. Feldman Barrett and J. A. Russell, “Independence and bipolarity in the structure of  
548        current affect,” *J. Pers. Soc. Psychol.*, vol. 74, no. 4, pp. 967–984, 1998, doi:  
549        10.1037/0022-3514.74.4.967.
- 550        [22] C. Henry, V. Mitropoulou, A. S. New, H. W. Koenigsberg, J. Silverman, and L. J. Siever,  
551        “Affective instability and impulsivity in borderline personality and bipolar II disorders:  
552        similarities and differences,” *J. Psychiatr. Res.*, vol. 35, no. 6, pp. 307–312, Nov. 2001, doi:  
553        10.1016/S0022-3956(01)00038-3.
- 554        [23] M. V. Heinz, N. X. Thomas, N. D. Nguyen, T. Z. Griffin, and N. C. Jacobson,  
555        “Technological Advances in Clinical Assessment,” in *Reference Module in Neuroscience*  
556        *and Biobehavioral Psychology*, Elsevier, 2021. doi: 10.1016/B978-0-12-818697-8.00171-0.
- 557        [24] M. D. Nemesure, M. V. Heinz, R. Huang, and N. C. Jacobson, “Predictive modeling of  
558        depression and anxiety using electronic health records and a novel machine learning  
559        approach with artificial intelligence,” *Sci. Rep.*, vol. 11, no. 1, Art. no. 1, Jan. 2021, doi:  
560        10.1038/s41598-021-81368-4.
- 561        [25] A. B. R. Shatte, D. M. Hutchinson, and S. J. Teague, “Machine learning in mental health: a  
562        scoping review of methods and applications,” *Psychol. Med.*, vol. 49, no. 9, pp. 1426–  
563        1448, Jul. 2019, doi: 10.1017/S0033291719000151.
- 564        [26] G. Price, M. V. Heinz, A. C. Collins, and N. C. Jacobson, “Detecting Major Depressive  
565        Disorder Presence Using Passively-Collected Wearable Movement Data in a Nationally-  
566        Representative Sample.” *PsyArXiv*, Jul. 03, 2023. doi: 10.31234/osf.io/9p4xr.
- 567        [27] N. C. Jacobson, H. Weingarden, and S. Wilhelm, “Digital biomarkers of mood disorders  
568        and symptom change,” *NPJ Digit. Med.*, vol. 2, p. 3, 2019, doi: 10.1038/s41746-019-0078-  
569        0.

## DETECTION OF NATURALISTIC DEPRESSION SYMPTOM VARIABILITY

- 570 [28] I. Moshe *et al.*, “Predicting Symptoms of Depression and Anxiety Using Smartphone and  
571 Wearable Data,” *Front. Psychiatry*, vol. 12, 2021, Accessed: Jun. 08, 2022. [Online].  
572 Available: <https://www.frontiersin.org/article/10.3389/fpsy.2021.625247>
- 573 [29] M. J. Peterson and R. M. Benca, “Sleep in Mood Disorders,” *Sleep Med. Clin.*, vol. 3, no. 2,  
574 pp. 231–249, Jun. 2008, doi: 10.1016/j.jsmc.2008.01.009.
- 575 [30] A. Korszun, E. A. Young, N. C. Engleberg, C. B. Brucksch, J. F. Greden, and L. A.  
576 Crofford, “Use of actigraphy for monitoring sleep and activity levels in patients with  
577 fibromyalgia and depression,” *J. Psychosom. Res.*, vol. 52, no. 6, pp. 439–443, Jun. 2002,  
578 doi: 10.1016/S0022-3999(01)00237-9.
- 579 [31] Y. Rykov, T.-Q. Thach, I. Bojic, G. Christopoulos, and J. Car, “Digital Biomarkers for  
580 Depression Screening With Wearable Devices: Cross-sectional Study With Machine  
581 Learning Modeling,” *JMIR MHealth UHealth*, vol. 9, no. 10, p. e24872, Oct. 2021, doi:  
582 10.2196/24872.
- 583 [32] C. Burton, B. McKinstry, A. Szentagotai Tătar, A. Serrano-Blanco, C. Pagliari, and M.  
584 Wolters, “Activity monitoring in patients with depression: A systematic review,” *J. Affect.*  
585 *Disord.*, vol. 145, no. 1, pp. 21–28, Feb. 2013, doi: 10.1016/j.jad.2012.07.001.
- 586 [33] R. Wang *et al.*, “StudentLife: assessing mental health, academic performance and  
587 behavioral trends of college students using smartphones,” in *Proceedings of the 2014*  
588 *ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp*  
589 *'14 Adjunct*, Seattle, Washington: ACM Press, 2014, pp. 3–14. doi:  
590 10.1145/2632048.2632054.
- 591 [34] M. Makhmutova, R. Kainkaryam, M. Ferreira, J. Min, M. Jaggi, and I. Clay, “Prediction of  
592 self-reported depression scores using person-generated health data from a virtual 1-year  
593 mental health observational study,” in *Proceedings of the 2021 Workshop on Future of*  
594 *Digital Biomarkers*, Virtual Event Wisconsin: ACM, Jun. 2021, pp. 4–11. doi:  
595 10.1145/3469266.3469878.
- 596 [35] M. Makhmutova, R. Kainkaryam, M. Ferreira, J. Min, M. Jaggi, and I. Clay, “Predicting  
597 Changes in Depression Severity Using the PSYCHE-D (Prediction of Severity Change-  
598 Depression) Model Involving Person-Generated Health Data: Longitudinal Case-Control  
599 Observational Study,” *JMIR MHealth UHealth*, vol. 10, no. 3, p. e34148, Mar. 2022, doi:  
600 10.2196/34148.
- 601 [36] S. Singha and P. P. Shenoy, “An adaptive heuristic for feature selection based on  
602 complementarity,” *Mach. Learn.*, vol. 107, no. 12, pp. 2027–2071, Dec. 2018, doi:  
603 10.1007/s10994-018-5728-y.
- 604 [37] Y. Zhang, H. Lyu, Y. Liu, X. Zhang, Y. Wang, and J. Luo, “Monitoring Depression Trend on  
605 Twitter during the COVID-19 Pandemic.” arXiv, Jul. 01, 2020. Accessed: Jun. 09, 2022.  
606 [Online]. Available: <http://arxiv.org/abs/2007.00228>
- 607 [38] J. L. Lee *et al.*, “The DiSCover Project: Protocol and Baseline Characteristics of a  
608 Decentralized Digital Study Assessing Chronic Pain Outcomes and Behavioral Data,” *Pain*  
609 *Medicine*, preprint, Jul. 2021. doi: 10.1101/2021.07.14.21260523.
- 610 [39] M. Makhmutova, R. Kainkaryam, M. Ferreira, J. Min, M. Jaggi, and I. Clay, “PSYCHE-D:  
611 predicting change in depression severity using person-generated health data (DATASET).”  
612 Zenodo, Jul. 09, 2021. doi: 10.5281/ZENODO.5085146.
- 613 [40] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, “The PHQ-9: Validity of a brief depression  
614 severity measure,” *J. Gen. Intern. Med.*, vol. 16, no. 9, pp. 606–613, Sep. 2001, doi:  
615 10.1046/j.1525-1497.2001.016009606.x.
- 616 [41] B. Arroll *et al.*, “Validation of PHQ-2 and PHQ-9 to Screen for Major Depression in the  
617 Primary Care Population,” *Ann. Fam. Med.*, vol. 8, no. 4, pp. 348–353, Jul. 2010, doi:

## DETECTION OF NATURALISTIC DEPRESSION SYMPTOM VARIABILITY

- 618 10.1370/afm.1139.
- 619 [42] J. Heaton, "An empirical analysis of feature engineering for predictive modeling," in  
620 *SoutheastCon 2016*, Norfolk, VA, USA: IEEE, Mar. 2016, pp. 1–6. doi:  
621 10.1109/SECON.2016.7506650.
- 622 [43] R Core Team, "R: A Language and Environment for Statistical Computing." R Foundation  
623 for Statistical Computing, 2021. [Online]. Available: <https://www.R-project.org/>
- 624 [44] S. van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate Imputation by Chained  
625 Equations in R," *J. Stat. Softw.*, vol. 45, no. 3, pp. 1–67, 2011.
- 626 [45] H. Kang, "The prevention and handling of the missing data," *Korean J. Anesthesiol.*, vol.  
627 64, no. 5, p. 402, 2013, doi: 10.4097/kjae.2013.64.5.402.
- 628 [46] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA:  
629 CreateSpace, 2009.
- 630 [47] D. Berrar, "Cross-Validation," in *Encyclopedia of Bioinformatics and Computational*  
631 *Biology*, Elsevier, 2019, pp. 542–545. doi: 10.1016/B978-0-12-809633-8.20349-X.
- 632 [48] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model  
633 selection," *BMC Bioinformatics*, vol. 7, p. 91, Feb. 2006, doi: 10.1186/1471-2105-7-91.
- 634 [49] X. Tao, O. Chi, P. J. Delaney, L. Li, and J. Huang, "Detecting depression using an  
635 ensemble classifier based on Quality of Life scales," *Brain Inform.*, vol. 8, no. 1, p. 2, Dec.  
636 2021, doi: 10.1186/s40708-021-00125-5.
- 637 [50] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in  
638 *Advances in Neural Information Processing Systems*, 2017, pp. 4765–4774.
- 639 [51] M. V. Heinz *et al.*, "Association of Selective Serotonin Reuptake Inhibitor Use With  
640 Abnormal Physical Movement Patterns as Detected Using a Piezoelectric Accelerometer  
641 and Deep Learning in a Nationally Representative Sample of Noninstitutionalized Persons  
642 in the US," *JAMA Netw. Open*, vol. 5, no. 4, p. e225403, Apr. 2022, doi:  
643 10.1001/jamanetworkopen.2022.5403.
- 644 [52] N. S. Klein, G. A. Holtman, C. L. H. Bockting, M. W. Heymans, and H. Burger,  
645 "Development and validation of a clinical prediction tool to estimate the individual risk of  
646 depressive relapse or recurrence in individuals with recurrent depression," *J. Psychiatr.*  
647 *Res.*, vol. 104, pp. 1–7, Sep. 2018, doi: 10.1016/j.jpsychires.2018.06.006.
- 648 [53] H. G. Ruhe *et al.*, "Emotional Biases and Recurrence in Major Depressive Disorder.  
649 Results of 2.5 Years Follow-Up of Drug-Free Cohort Vulnerable for Recurrence," *Front.*  
650 *Psychiatry*, vol. 10, 2019, Accessed: Jun. 09, 2022. [Online]. Available:  
651 <https://www.frontiersin.org/article/10.3389/fpsy.2019.00145>
- 652 [54] S. Berrouguet, M. M. Perez-Rodriguez, M. Larsen, E. Baca-García, P. Courtet, and M.  
653 Oquendo, "From eHealth to iHealth: Transition to Participatory and Personalized Medicine  
654 in Mental Health," *J. Med. Internet Res.*, vol. 20, no. 1, p. e7412, Jan. 2018, doi:  
655 10.2196/jmir.7412.
- 656 [55] S. Jahangir, D. Adjepong, H. A. Al-Shami, and B. H. Malik, "Is There an Association  
657 Between Migraine and Major Depressive Disorder? A Narrative Review," *Cureus*, vol. 12,  
658 no. 6, p. e8551, 2020, doi: 10.7759/cureus.8551.
- 659 [56] C. V. Molgat and S. B. Patten, "Comorbidity of Major Depression and Migraine — A  
660 Canadian Population-Based Study," *Can. J. Psychiatry*, vol. 50, no. 13, pp. 832–837, Nov.  
661 2005, doi: 10.1177/070674370505001305.
- 662 [57] E. Fuller-Thomson, M. Battiston, T. M. Gadalla, and S. Brennenstuhl, "Bouncing back:  
663 remission from depression in a 12-year panel study of a representative Canadian  
664 community sample," *Soc. Psychiatry Psychiatr. Epidemiol.*, vol. 49, no. 6, pp. 903–910,  
665 Jun. 2014, doi: 10.1007/s00127-013-0814-8.

## DETECTION OF NATURALISTIC DEPRESSION SYMPTOM VARIABILITY

- 666 [58] C.-I. Hung, C.-Y. Liu, Y.-Y. Juang, and S.-J. Wang, "The Impact of Migraine on Patients  
667 With Major Depressive Disorder," *Headache J. Head Face Pain*, vol. 46, no. 3, pp. 469–  
668 477, 2006, doi: 10.1111/j.1526-4610.2006.00378.x.
- 669 [59] "Headache Classification Committee of the International Headache Society (IHS) The  
670 International Classification of Headache Disorders, 3rd edition," *Cephalalgia Int. J.*  
671 *Headache*, vol. 38, no. 1, pp. 1–211, Jan. 2018, doi: 10.1177/0333102417738202.
- 672 [60] D. M. Fergusson, J. M. Boden, and L. J. Horwood, "Recurrence of major depression in  
673 adolescence and early adulthood, and later mental health, educational and economic  
674 outcomes," *Br. J. Psychiatry*, vol. 191, no. 4, pp. 335–342, Oct. 2007, doi:  
675 10.1192/bjp.bp.107.036079.
- 676 [61] S. A. Everson, S. C. Maty, J. W. Lynch, and G. A. Kaplan, "Epidemiologic evidence for the  
677 relation between socioeconomic status and depression, obesity, and diabetes," *J.*  
678 *Psychosom. Res.*, vol. 53, no. 4, pp. 891–895, Oct. 2002, doi: 10.1016/S0022-  
679 3999(02)00303-3.
- 680 [62] M. Melchior *et al.*, "Socioeconomic position predicts long-term depression trajectory: a 13-  
681 year follow-up of the GAZEL cohort study," *Mol. Psychiatry*, vol. 18, no. 1, Art. no. 1, Jan.  
682 2013, doi: 10.1038/mp.2011.116.
- 683 [63] M. J. Roche and N. C. Jacobson, "Elections Have Consequences for Student Mental  
684 Health: An Accidental Daily Diary Study," *Psychol. Rep.*, vol. 122, no. 2, pp. 451–464, Apr.  
685 2019, doi: 10.1177/0033294118767365.
- 686 [64] D. English, S. F. Lambert, and N. S. Ialongo, "Longitudinal Associations Between  
687 Experienced Racial Discrimination and Depressive Symptoms in African American  
688 Adolescents," *Dev. Psychol.*, vol. 50, no. 4, pp. 1190–1196, Apr. 2014, doi:  
689 10.1037/a0034703.
- 690 [65] L. Torres and A. D. Ong, "A daily diary investigation of latino ethnic identity, discrimination,  
691 and depression," *Cultur. Divers. Ethnic Minor. Psychol.*, vol. 16, no. 4, pp. 561–568, 2010,  
692 doi: 10.1037/a0020652.
- 693 [66] N. Tsuno, A. Besset, and K. Ritchie, "Sleep and Depression," *J. Clin. Psychiatry*, vol. 66,  
694 no. 10, p. 19685, Oct. 2005.
- 695 [67] R. Tavernier and T. Willoughby, "Bidirectional associations between sleep (quality and  
696 duration) and psychosocial functioning across the university years.," *Dev. Psychol.*, vol. 50,  
697 no. 3, pp. 674–682, Mar. 2014, doi: 10.1037/a0034258.
- 698 [68] W. Li, J. Yin, X. Cai, X. Cheng, and Y. Wang, "Association between sleep duration and  
699 quality and depressive symptoms among university students: A cross-sectional study,"  
700 *PLOS ONE*, vol. 15, no. 9, p. e0238811, Sep. 2020, doi: 10.1371/journal.pone.0238811.
- 701 [69] K. Bi and S. Chen, "Sleep profiles as a longitudinal predictor for depression magnitude and  
702 variability following the onset of COVID-19," *J. Psychiatr. Res.*, vol. 147, pp. 159–165, Mar.  
703 2022, doi: 10.1016/j.jpsychires.2022.01.024.
- 704 [70] Y. Fang, D. B. Forger, E. Frank, S. Sen, and C. Goldstein, "Day-to-day variability in sleep  
705 parameters and depression risk: a prospective cohort study of training physicians," *Npj*  
706 *Digit. Med.*, vol. 4, no. 1, Art. no. 1, Feb. 2021, doi: 10.1038/s41746-021-00400-z.
- 707 [71] V. L. Amelia, H.-J. Jen, T.-Y. Lee, L.-F. Chang, and M.-H. Chung, "Comparison of the  
708 Associations between Self-Reported Sleep Quality and Sleep Duration Concerning the  
709 Risk of Depression: A Nationwide Population-Based Study in Indonesia," *Int. J. Environ.*  
710 *Res. Public Health*, vol. 19, no. 21, p. 14273, Nov. 2022, doi: 10.3390/ijerph192114273.
- 711 [72] R. Furihata *et al.*, "Association of short sleep duration and short time in bed with  
712 depression: A Japanese general population survey: Short time in bed and depression,"  
713 *Sleep Biol. Rhythms*, vol. 13, no. 2, pp. 136–145, Apr. 2015, doi: 10.1111/sbr.12096.

## DETECTION OF NATURALISTIC DEPRESSION SYMPTOM VARIABILITY

- 714 [73] T. Zhou *et al.*, “The Associations between Sleep Duration, Academic Pressure, and  
715 Depressive Symptoms among Chinese Adolescents: Results from China Family Panel  
716 Studies,” *Int. J. Environ. Res. Public. Health*, vol. 18, no. 11, p. 6134, Jun. 2021, doi:  
717 10.3390/ijerph18116134.
- 718 [74] Y. Hu *et al.*, “The relationship between sleep pattern and depression in Chinese shift  
719 workers: A mediating role of emotional exhaustion,” *Aust. J. Psychol.*, vol. 72, no. 1, pp.  
720 68–81, Mar. 2020, doi: 10.1111/ajpy.12253.
- 721 [75] M. Philipp and M. Fickinger, “The Definition of Remission and Its Impact on the Length of a  
722 Depressive Episode.,” *Arch Gen Psychiatry*, vol. 50, no. 5, pp. 407–408, 1993, doi:  
723 doi:10.1001/archpsyc.1993.01820170093013.  
724

725

**Figure and Table Legend(s)**

726 *Figure 1. A flow diagram, representing the selection and exclusion of participants, which led to the 939-*  
 727 *participant sample in the present work. From top to bottom: Headers at the top of the diagram reflect*  
 728 *projects, with citations, from which the data originated. Below the headers, we show the absolute*  
 729 *numbers of participants, changing with further exclusion, during each stage of the project. Dialogue*  
 730 *bubbles provide detail at a stage where participants were excluded. Large rectangular dialogue boxes*  
 731 *contain high level detail regarding features included at each stage. Gradient arrows indicate feature*  
 732 *change or subsetting that occurred to produce the feature set used in the present work.*

733

734 *Figure 2. Model(s) actual versus predicted values plotted with respective correlative strength and the top*  
 735 *five most influential features for the models predictions. In the respective SHAP plots, the individual dot*  
 736 *color corresponds to the value of the variable, and location on the plot's x-axis corresponds to that*  
 737 *point's relative impact on the model output (e.g., a high-feature value (red) with a corresponding high x-*  
 738 *axis value (SHAP value) represents a point that strongly, positively influences the model's prediction of*  
 739 *depression symptom variability). (A) Baseline biodemographic variables. (B) Passively-collected*  
 740 *movement and sleep variables. (C) A composite model, using biodemographic and passively-collected*  
 741 *movement and sleep variables.  $r$  = Pearson's correlation coefficient. For binary features, presence of*  
 742 *comorbid migraines, male sex, required financial assistance, and white race represented a higher feature*  
 743 *value (red SHAP value color).*

744

745 *Figure 3. Comparative analysis incorporating or transforming all originally collected variables for the*  
 746 *three respective models. Model(s) actual versus predicted values plotted with respective correlative*  
 747 *strength and the top five most influential features for the models predictions. In the respective SHAP*  
 748 *plots, the individual dot color corresponds to the value of the variable, and location on the plot's x-axis*  
 749 *corresponds to that point's relative impact on the model output (e.g., a high-feature value (red) with a*  
 750 *corresponding high x-axis value (SHAP value) represents a point that strongly, positively influences the*  
 751 *model's prediction of depression symptom variability). (A) Baseline biodemographic variables. (B)*  
 752 *Passively-collected movement and sleep variables. (C) A composite model, using biodemographic and*  
 753 *passively-collected movement and sleep variables.  $r$  = Pearson's correlation coefficient.*

754

755 *Table 1A. Model performance of the theory-organized and full variable set stacked ensemble machine*  
 756 *learning approaches for the validation and held-out test set(s) for the three model types, reported as*  
 757 *correlation  $\pm$  standard deviation.*

758

759 *Table 1B. Model performance of the theory-organized and full variable set stacked ensemble machine*  
 760 *learning approaches for the validation and held-out test set(s) for the three model types, reported as*  
 761 *normalized mean absolute error  $\pm$  standard deviation.*

762



763

**Supplementary Material**

764 Supplementary Table 1.

765

766 Name and description of all features included in the theory-organized modeling approach.

767

768 Supplementary Table 2.

769

770 Machine learning architecture including algorithm(s) and hyperparameter(s) for the present analyses.

771

772

773