Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Evaluating depression with multimodal wristband-type wearable device: screening and assessing patient severity utilizing machine-learning

Yuuki Tazawa [a], Kuo-ching Liang [a], Michitaka Yoshimura [a], Momoko Kitazawa [a], Yuriko Kaise [a], Akihiro Takamiya [a], Aiko Kishi [b], Toshiro Horigome [a], Yasue Mitsukura [b], Masaru Mimura [a], Taishiro Kishimoto [a,*]

[a] *Keio University School of Medicine, Tokyo, Japan*
[b] *Faculty of Science and Technology, Keio University, Kanagawa, Japan*

ARTICLE INFO

Keywords:
Psychiatry
Biological psychiatry
Depression
Clinical research
Health informatics
Health technology
Diagnostics
Biomarkers
Machine learning
Wearable electronic devices
Heart rate
Sleep
Body temp

ABSTRACT

Objective: We aimed to develop a machine learning algorithm to screen for depression and assess severity based on data from wearable devices.
Methods: We used a wearable device that calculates steps, energy expenditure, body movement, sleep time, heart rate, skin temperature, and ultraviolet light exposure. Depressed patients and healthy volunteers wore the device continuously for the study period. The modalities were compared hourly between patients and healthy volunteers. XGBoost was used to build machine learning models and 10-fold cross-validation was applied for the validation.
Results: Forty-five depressed patients and 41 healthy controls participated, creating a combined 5,250 days' worth of data. Heart rate, steps, and sleep were significantly different between patients and healthy volunteers in some comparisons. Similar differences were also observed longitudinally when patients' symptoms improved. Based on seven days' data, the model identified symptomatic patients with 0.76 accuracy and predicted Hamilton Depression Rating Scale-17 scores with a 0.61 correlation coefficient. Skin temperature, sleep time-related features, and the correlation of those modalities were the most significant features in machine learning.
Limitations: The small number of subjects who participated in this study may have weakened the statistical significance of the study. There are differences in the demographic data among groups although we performed a correction for multiple comparisons. Validation in independent datasets was not performed, although 10-fold cross validation with the internal data was conducted.
Conclusion: The results indicated that utilizing wearable devices and machine learning may be useful in identifying depression as well as assessing severity.

1. Introduction

In recent years, small sensors that can be attached to a person's body for 24 h, known as "wearable devices", have been widely used (Mazzetta et al., 2018). They can continuously and non-invasively collect a variety of information, including amount of activity, amount of sleep, heart rate, respiratory rate, and physical location (Nieto-Riveiro et al., 2018). In many medical fields such as cardiology, endocrinology, and metabolic medicine, the use of wearable devices in clinical research and application is growing (Kuehn, 2016; Shelgikar et al., 2016). This trend also holds true in psychiatric research, as data related to activity, sleep, heart rate, etc. have been shown in previous studies to have relevance in determining diagnoses and illness severity (Marzano et al., 2015; Reinertsen and Clifford, 2018). Because there is a lack of quantifiable biological markers in psychiatry (Beijers et al., 2019), the ability to non-invasively collect data from wearable devices can improve diagnoses and evaluations of illness severity (Dogan et al., 2017; Marzano et al., 2015; Patel et al., 2017). However, currently there is no sufficient evidence concerning the use of such wearable devices in clinical examinations. Actigraphy, which relies on a device that collects data via an accelerometer, was first used in the psychiatric field as early as the 1970s, and a large number of studies have been conducted since then (Burton et al., 2013; Luik et al., 2015; Martin and Hakim, 2011). In a meta-analysis focusing on such studies that used actigraphy to evaluate mood disorders (Tazawa et al., 2019), significant differences in the daily activity and sleep-related measurements were found between patients with mood disorders and healthy controls.

* Corresponding author.
E-mail address: tkishimoto@keio.jp (T. Kishimoto).

https://doi.org/10.1016/j.heliyon.2020.e03274
Received 26 November 2019; Received in revised form 11 December 2019; Accepted 17 January 2020

Moreover, significant differences were found when comparing pre- and post-treatment periods. Additionally, specific measurement patterns characterizing each mood disorder/status were found.

Thanks to advances in wearable device technology, actigraphy-enabled wearable devices with similar capabilities as those made for research are now available commercially at economical prices. Moreover, some new modalities that were not measurable with previous devices can now be measured with the new devices, such as heart and respiratory rates, skin temperature, location information, etc. If wearable devices are able to collect multi-modal data at an economical cost, they would be viable tools for evaluating mood disorders in clinical settings, and they could even be used in pre-clinical screenings.

Studies that assessed mood disorders with these new modalities reported the following findings. Moraes et al. (2013) found that in patients with depression (n = 20), there was low amplitude for circadian rhythm and light exposure, and high amplitude for peripheral body temperature compared to healthy controls (n = 10). nullvakmfondeknqsyqiykeegy reported that a weak 24-hour periodicity of body temperature was prominent in melancholic depressed patients (n = 41) compared to healthy controls (n = 25). Licht et al. (2008) reported that those with depression had lower heart rate variability than healthy controls. In recent years, there has been an increase in studies that analyze mood disorders using various data collected from wearable devices (Rohani et al., 2018; Wang et al., 2018), and in particular, research on bipolar disorder is moving forward (Puiatti et al., 2011; Valenza et al., 2014, 2015). Valenza et al. used a wearable textile device to analyze heart rate variability, and reported that they were able to predict bipolar disorder patients' emotional states with an accuracy of over 90%.

Along with the improvement in the sensory abilities of wearable devices, machine learning is becoming increasing popular in the medical field, as clinical data often contain complex cross-sectional and longitudinal patterns. Saad et al. (2019) used heart rate variability during sleep from polysomnograms to distinguish 87 major depressive disorder (MDD) patients and 87 healthy controls utilizing machine learning, then reported a classification accuracy of 79.9%. Valenza et al. (2013) used inter-beat interval time series, heart rate, and respiratory dynamics data to create algorithms to predict mood states. They collected over 120 h of data from three subjects with bipolar disorder and reported a mood state prediction accuracy of 97%. However, limitations for this study include a small study population of three people, and the fact that no rating scale was used to assign illness and mood state labels. Cho et al. (2019) used activity, sleep, light exposure, and heart rate data from wearable devices and smartphones to predict mood state within the next three days. They recruited 55 patients with mood disorders, and analyzed their data with machine learning, then reported a prediction accuracy of 64–94%. However, they used their original self-reported mood assessment scale, and predicted the presence of symptoms but not severity.

Given such limitations in previous similar studies, we aimed to investigate the usefulness of wearable devices with sensors for acceleration, heart rate, skin temperature, and ultraviolet (UV) light to identify symptomatic patients as well as measure illness severity through a biostatistical machine learning approach.

2. Methods

2.1. Subjects

This was a multicenter study, and depressed patients were recruited from 10 different medical facilities in five prefectures in Japan. This research was done as part of a larger project called Project for Objective Measures Using Computational Psychiatry Technology (PROMPT). The concept and design of PROMPT is reported elsewhere (Kishimoto et al., 2019). PROMPT protocols have been registered with the University Hospital Medical Information Network (UMIN) (UMIN ID: UMIN000023764). This study was approved by the ethics committee of Keio University Hospital and other participating facilities.

The subjects were depression inpatients and outpatients and healthy controls over 20 years of age. Patient participants in this study were referred to our research team by their attending physician when they were examined as outpatients or hospitalized as inpatients. Then the research team held a detailed interview with each referred patient, wherein it was confirmed whether the patient met the DSM-5 diagnostic criteria for MDD or BD. Healthy controls were confirmed to have no history of psychiatric illness (screening done with Mini-International Neuropsychiatric Interview [MINI]). Exclusion criteria for this study were: (1) patients whose illness could be exacerbated by the study's interview; (2) patients who have comorbidities that could interfere with measurements in the study, such as patients with involuntary movement or quadriplegic palsy.

2.2. Clinical assessment

Demographic characteristics such as age, sex, duration of illness, diagnoses, and medication were collected. Participants were assessed by fully-trained psychologists or psychiatrists whose inter-rater reliability was over 0.9 using the following clinical rating scales: Hamilton Depression Rating Scale (HAMD; Hamilton, 1960), Montgomery-Asberg Depression Rating Scale (MADRS; Montgomery and Asberg, 1979), Young Mania Rating Scale (YMRS). Participants were asked to complete the Beck Depression Inventory-II (BDI-II) and Pittsburgh Sleep Quality Index (PSQI) by themselves as well.

Evaluation intervals for outpatients were aligned with their hospital visits to prevent any additional burden. The evaluation interval for inpatients was roughly every 1–2 weeks and evaluations were conducted up to 10 times.

2.3. Wearable device acquisition

During the study period, subjects were asked to continuously wear a Silmee W20 (TDK, Inc., Tokyo, Japan) wristband biosensor device to measure step count, energy expenditure, body movement, sleep time, heart rate, skin temperature, and UV light exposure during their daily activities.

Step count, energy expenditure, body movement, and sleep time were calculated based on accelerator data. Heart rate, skin temperature, and UV exposure data were each collected directly from their respective sensors on the device.

Heart rate data was collected for a one-minute period every five minutes. Other data were measured by calculating the total sum of a one-minute interval every one-minute. The validation data is available upon request to the corresponding author. Heart rate data were measured by processing pulse wave signals collected by a photoplethysmographic sensor.

This device was also used in a previous study which investigated the relationship between the device's measurements and cognitive function in elderly people (Kimura et al., 2019).

2.4. Wearable device data preprocessing

Before conducting the analyses in the following sections, preprocessing was applied to the acquired data. We prepared the raw data taken from the wearable devices into hourly and daily data for biostatistical and machine learning analysis. The process and inclusion/exclusion criteria of the data are as described below.

a. In order to analyze data collected during periods when the devices were worn properly by the subjects, we used heart rate as the inclusion criterion. We included data if a patient's mean heart rate was between 30-200 beats per minute for hourly data and discarded any data outside those parameters. The reason we chose heart rate as our control measure is because, unlike acceleration-type data, etc., regardless of lifestyle and how the device is worn, heart rate data have a biologically-defined acceptable range, and it is easy to determine data that deviate from this standard.

b. Heart rate data were collected 12 times per hour, every five minutes. When capturing heart rate data, data that records the heart rate six or less times per hour is discarded. This is because, in order for one hour's worth of data to properly represent a patients' heart rate, we concluded that the necessary minimum data collection frequency for the majority of the data was seven or more times per hour.
c. The hourly data detailed in the process description above was incorporated into what we labeled as “daily data”, which comprised 20 or more hours of data in one 24-hour period. Any data that did not fit those parameters were discarded. This was decided because we believed that if four or more hours of data were missing in a one-day period, there was a possibility that the device was not worn properly, causing a loss of sleep data, etc., which would not be representative of one day's worth of data.
d. The wearable device data that was included in our daily data sets for analysis were collected within seven days of the day that a patient's symptom evaluation was done.

2.5. Biostatistical analysis

Descriptive statistics were used to describe the study participants. Statistical significance was set at two-tailed $p < 0.05$, and we used false discovery rate (FDR) to control for multiple comparisons.

Because the mood disorders group and healthy control group had significant differences in age distribution, we used multiple linear regression to control for age when comparing those groups. When comparing the subsets of patients with depression and healthy controls with no significant differences in age, we used Student's t-test. During the course of this study, each participant underwent symptom evaluations up to 10 times, but for our analysis, we included only one evaluation per person in the analysis in the following manner: for patients with depression, we used data from the evaluation session wherein the most severe symptoms of an individual patient were observed; for healthy controls, we used data from the session that showed the least severe symptoms for that participant.

Depression data was collected longitudinally, and during the study, there were more than a few patients with depression who went into remission. When we compared patients with healthy controls, we extracted the data for when the patients' depression symptoms were strongest; in other words, when the patients were experiencing their depression symptoms the most. We also collected the healthy controls' data longitudinally, and we extracted data for when the healthy controls were estimated to be most healthy.

Next, we performed a longitudinal comparison of the data from patients with depression. We compared data from the patients' evaluation sessions based on their HAMD-17 scores, using data from the session with the highest HAMD-17 score and the session with the lowest HAMD-17 score. As the purpose of this analysis is to examine if there are any modalities that are reflective of symptom change, we only included data from patients whose HAMD-17 score had shifted four or more points between the highest and lowest sessions. Since the time between the two sessions varies from person to person, we used the number of days between the two sessions as a control variable in a multiple regression analysis.

2.6. Screening and severity assessment using machine learning

We used machine learning to create an evaluation model to determine the presence of depression symptoms. For this machine learning analysis, we included all HAMD rating data sets for all participants. For each rating, we divided subjects (including patient and healthy control datasets) into two groups based on their HAMD-17 score: datasets with scores of eight or more were placed in the symptomatic group, and datasets with scores of seven or less were placed in the asymptomatic group.

First, we used data taken from the wearable devices to create features for use in machine learning models. We used the following method for this process:

a. Collect the per-hour data of the seven data types (step count, energy expenditure, body movement, sleep time, heart rate, skin temperature, and UV exposure) for up to one day prior to the current timestamp. We chose per-hour data as our feature value because per-minute data were too noisy, and per-day data would not provide enough information for an accurate analysis.
b. For each data type, compute the 5th, 25th, 50th, 75th, and 95th percentiles of the distribution of the per-hour data from the collected one-day data.
c. For each data type, compute the standard deviation of the per-hour data from the collected one-day data.
d. For each unique pair of data types, compute the Pearson's correlation coefficient of the per-hour data of the two data types from the collected one-day data.

Overall, we extracted a total of 63 features for building machine learning models. Using these extracted feature values, we followed the below process to create machine learning models to evaluate the presence of depressed state. For machine learning, we chose XGBoost. Using our initial smaller dataset, we built models with different machine learning algorithms such as SVM, Random Forest, and XGBoost, and the results showed that XGBoost provided the best results out of all of the algorithms that we tried. Therefore, we used only XGBoost for the final dataset:

a. For each subject's rating assessment, extract the 63 depression status screening features from the per-hour data for up to either three or seven days prior to the clinical assessment.
b. Train an XGBoost (Chen and Guestrin, 2016) classifier model using all the depression status screening features as the input feature vector, along with physician-assessed HAMD scores.
c. Grid search is used to find the optimal parameters for XGBoost.

We used a sample in the model training if the sample had at least one day of data collected within three or seven days prior to the clinical assessment for a 3-day or 7-day model, respectively. We excluded any data that exceeded our 3- and 7-day periods for analysis. In cases where data were not recorded for a full day, those data sets were also excluded from our analysis. For the models we created, we used the following 10-fold cross-validation (Zhang and Yang, 2015) and calculated the accuracy, sensitivity, and specificity levels of the models.

We used machine learning to create prediction models for depression severity as well. The HAMD-17 score of each rating session was used as a target output for the training datasets, and we trained machine learning models for the prediction of HAMD-17 scores based on the data taken from the wearable devices.

2.7. Feature importance using machine learning

To investigate how important each feature is to the prediction of HAMD scores, we calculated the feature importance of each feature, averaged over the 10 models of the 10-fold cross-validation. Feature importance, or Gini importance of a feature, is related to the amount that the feature's split points contribute to the improvement of the performance metric in a decision tree (Breiman, 2001). The feature importance of all the features have a total sum of 1, and this allows us to compare the relative importance of features with one another.

3. Results

Our study included 30 MDD patients, 15 BD patients, and 41 healthy controls, producing a combined 5,250 days' worth of data. The breakdown of patients from each participating facility is as follows: Keio Hospital, 23 people; Oizumi Hospital, 2; Oizumi Mental Clinic, 2; Tsurugaoka Garden Hospital, 10; Nagatsuta Ikoi Forest Clinic, 8; Komagino Hospital, 8; Asaka Hospital, 18; Biwako Hospital, 8; Sato Hospital, 12.

From each individual patient, we collected evaluation data at least one time, and at most nine times, for an average collection rate of 4.2 times per subject. This provided a total of 241 datasets, with 133 sets from MDD patients, 49 sets from BD patients, and 59 sets from healthy controls. Of these 241 datasets, there were 114 datasets with ≥8 on HAMD, four datasets with ≥8 on YMRS, one dataset with ≥8 on both HAMD and YMRS, and 122 datasets with asymptomatic state.

The demographic characteristics for each group are shown in Table 1. Healthy controls were significantly older and there was no significant difference in gender.

3.1. Biostatistical comparisons of patients vs. healthy controls

Per-hour and per-day data comparisons are shown in Figure 1. Even after performing multiple comparison corrections with FDR, the step count, energy expenditure, and body movement all showed similar tendencies and were likely to be significantly high in the healthy control group during roughly the hours between 8:00 am–6:00 pm and 10:00 pm–11:00 pm. Sleep time for the patient group was significantly longer during the hours of 7:00 pm–11:00 pm. The skin temperature of the patient group was significantly higher during the hours of 7:00 pm–10:00 pm. Heart rate results were significantly higher in the patient group during almost all time periods, but particularly during the hours of 1:00 am–9:00 am. UV exposure was significantly higher in the patient group only during the hour between 7:00 am–8:00 am. In a per-day comparison, only energy expenditure was significantly higher in the control group.

Given that there were significant age differences between patient and healthy control populations, we selected subpopulations of samples from both the patient and control populations with similar age distributions and compared the two groups to see the robustness of the results, as a post hoc analysis. The result of this comparison showed that significant differences in the outcomes recorded by the accelerometer, such as step count, energy expenditure, body movement, and sleep time, remained. However, significant differences in heart rate disappeared.

3.2. Longitudinal comparison of patients using biostatistics

Per-hour and per-day data comparisons are shown in Figure 2. Before performing multiple comparison corrections with FDR, the step count, energy expenditure, and body movement were found to be significantly higher during less severe symptom periods within the hours of 10:00 am–4:00 pm. For sleep time, sleep was significantly longer within the hour of 3:00 am–4:00 am during less severe symptom periods, and during severer symptom periods, sleep was significantly longer from 8:00 pm–9:00 pm. During less severe symptom periods, significantly high results were seen for skin temperature from 11:00 am–12:00 pm; for heart rate, from 12:00 am–1:00 am; and for UV exposure, from 2:00 pm–3:00 pm. However, these differences disappeared once multiple comparison corrections were done. In a per-day data comparison, only body movement was found to be significantly higher during less severe symptom periods, even after multiple comparison corrections.

3.3. Screening and severity prediction using machine learning

For each symptom evaluation session, we calculated the screening accuracy of the trained model using three days' worth of wearable device data, which consists of 109 symptomatic samples and 119 asymptomatic samples, for a total of 228 datasets. On average, 2.97 ± 0.35 days' worth of data was obtained per subject. For the 3-day dataset, we achieved an accuracy of 74%, sensitivity of 66%, and specificity of 81%. Similarly, for the 7-day dataset, there are 112 symptomatic samples and 124 asymptomatic samples, for a total of 236 samples. On average, 6.66 ± 1.23 days' worth of data was obtained per subject. From the 7-day model, we achieved an accuracy of 76%, sensitivity of 73%, and a specificity of 79%.

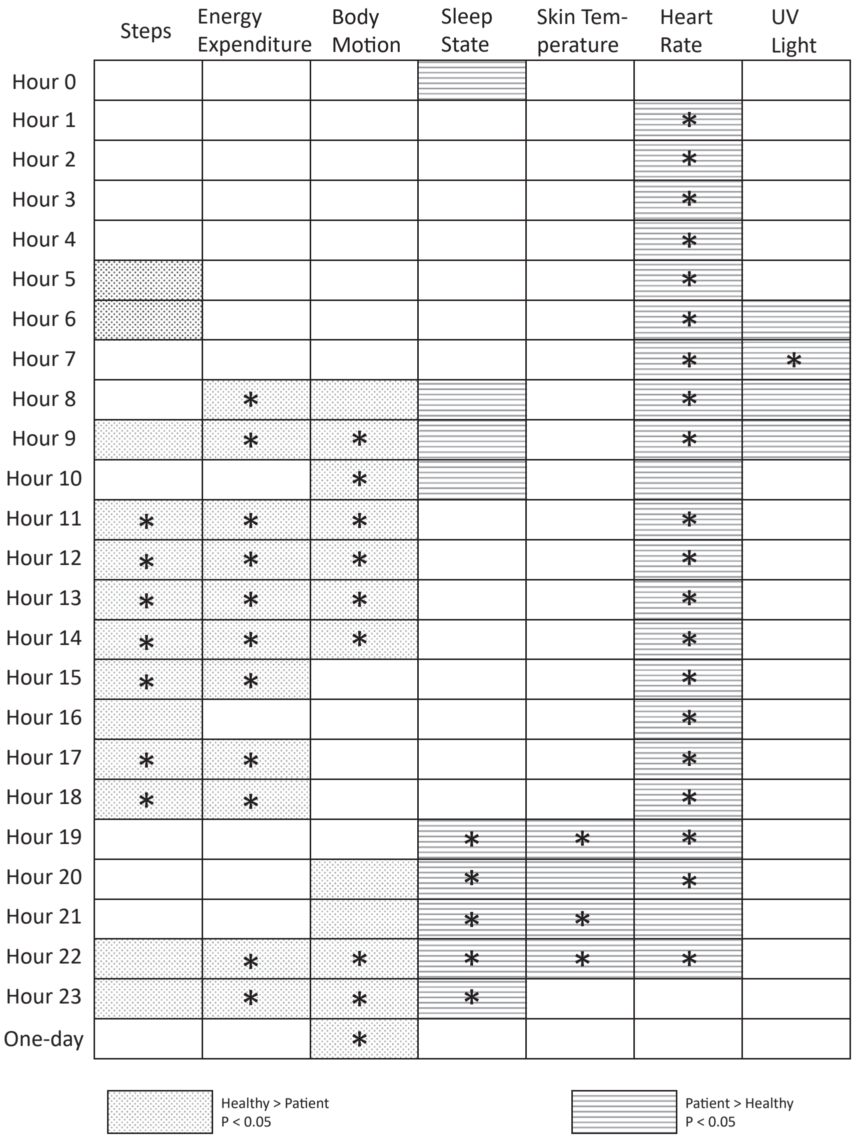
When measuring the accuracy of symptom severity predictions using the HAMD-17 score, we made calculations using three days' worth of wearable device data per evaluation session, in which we included 109 symptomatic samples and 119 asymptomatic samples, for a total of 228 samples. The mean absolute error was 5.33, and the correlation between HAMD-17 scores as rated by a clinician and predicted by machine learning was $r = 0.47$ ($p = 8.95e\text{-}14$), $R2 = 0.22$. Similarly, when using seven days' worth of data to calculate the accuracy of the illness severity prediction model, we included 112 symptomatic samples and 124 asymptomatic samples, for a total of 236 samples. The mean absolute error of the model was 4.94, and the correlation was $r = 0.61$ ($p = 2.20e\text{-}16$), $R2 = 0.37$ (Figure 3).

3.4. Feature importance using machine learning

Through 10-fold cross-validation, we built a list of features that consistently show high feature importance in all the folds. Based on the results, skin temperature had the highest importance, and sleep time had the next highest. Furthermore, the correlation between skin temperature and sleep time also appeared in most models as one of the most important features (Table 2).

Table 1. Demographic characteristics of patients and healthy controls.

	Patients (n = 45)	Healthy Controls (n = 41)	p
Age, in years, Mean (SD)	52.1 (13.2)	69.1 (14.2)	<0.05
Female, n (%)	21 (46.7)	19 (46.3)	0.97
Illness duration, in years, Mean (SD)	9.7 (9.3)	-	
Interview-Based Assessment Score			
Hamilton Depression Scale-17 score, Mean (SD)	14.6 (9.3)	2.29 (2.70)	<0.05
Montgomery Asberg Depression Rating Scale score, Mean (SD)	17.47 (12.93)	1.51 (2.93)	<0.05
Young Mania Rating Scale score, Mean (SD)	2.0 (4.4)	0.2 (0.71)	<0.05
Self-Rating Assessment Score			
Beck Depression Inventory, Mean (SD)	18.35 (12.55)	5.78 (5.55)	<0.05
Pittsburgh Sleep Quality Index, Mean (SD)	9.49 (4.43)	5.59 (3.44)	<0.05
Medication			
Any antidepressant, n (%)	30 (66.7)	-	
Any antipsychotic, n (%)	25 (55.6)	-	
Any mood stabilizer, n (%)	16 (35.6)	-	
Any anxiolytic/hypnotic, n (%)	39 (86.7)	-	



	Steps	Energy Expenditure	Body Motion	Sleep State	Skin Temperature	Heart Rate	UV Light
Hour 0				▤			
Hour 1						▤ *	
Hour 2						▤ *	
Hour 3						▤ *	
Hour 4						▤ *	
Hour 5	░					▤ *	
Hour 6	░					▤ *	▤
Hour 7						▤ *	▤ *
Hour 8		░ *	░	▤		▤ *	▤
Hour 9	░	░ *	░ *	▤		▤ *	▤
Hour 10			░ *	▤		▤	
Hour 11	░ *	░ *	░ *			▤ *	
Hour 12	░ *	░ *	░ *			▤ *	
Hour 13	░ *	░ *	░ *			▤ *	
Hour 14	░ *	░ *	░ *			▤ *	
Hour 15	░ *	░ *				▤ *	
Hour 16	░					▤ *	
Hour 17	░ *	░ *				▤ *	
Hour 18	░ *	░ *				▤ *	
Hour 19				▤ *	▤ *	▤ *	
Hour 20			░	▤ *	▤	▤ *	
Hour 21			░	▤ *	▤ *	▤	
Hour 22	░	░ *	░ *	▤ *	▤ *	▤ *	
Hour 23	░	░ *	░ *	▤ *			
One-day			░ *				

░ Healthy > Patient P < 0.05

▤ Patient > Healthy P < 0.05

* Significant difference remained after false discovery rate (FDR) for multiple comparisons.

Each cell indicates the result of comparisons for each outcome during each time interval.

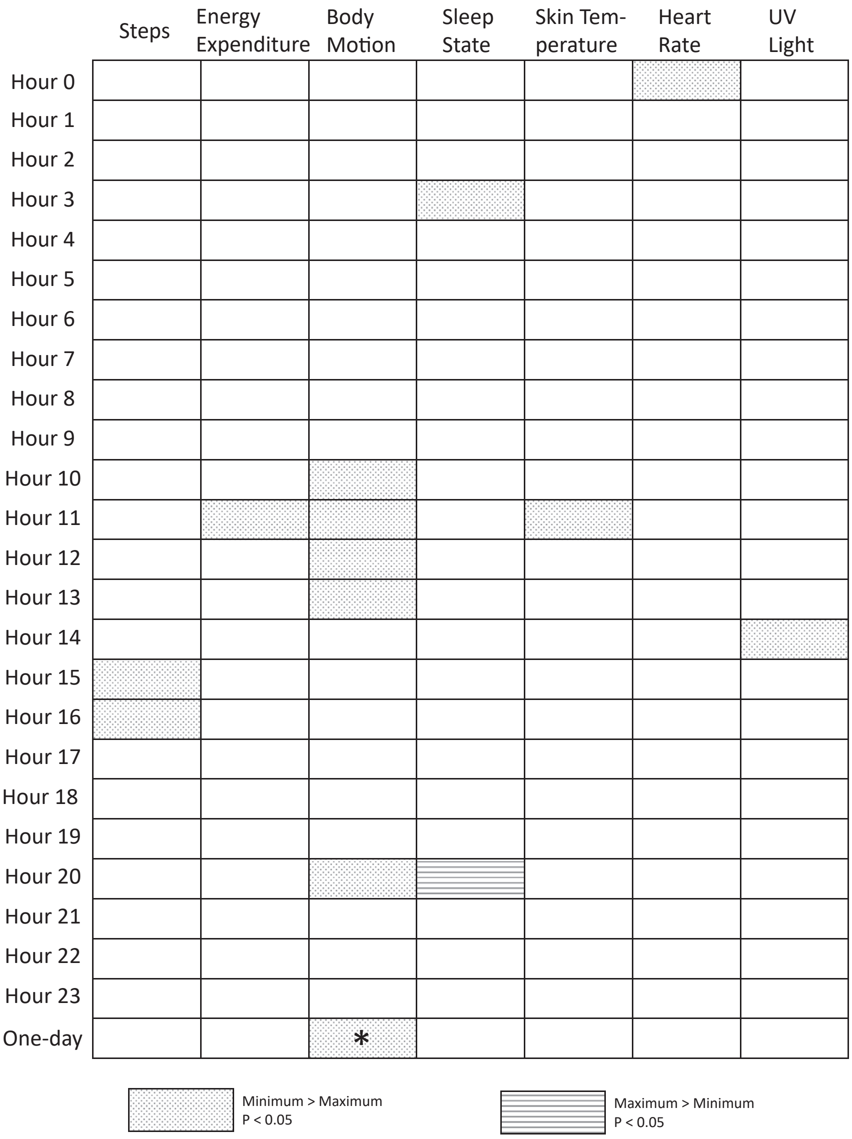
Figure 1. Results of Biostatistical Comparisons of Patients vs Healthy Controls During Each Time Interval.

4. Discussion

In this study, a multimodal wristband-type wearable device was used to estimate the presence and severity of depression. Previous studies have investigated the relationship between depression and each modality measured in this study, but they did not reflect the illnesses in question well enough to be applied as diagnostic tools in clinical practice. Our research combined multiple datasets with a machine learning approach to create a practical model for estimating both the presence and severity of depressive states.

First, using the seven modalities, we compared each dataset among mood disorder patients and healthy controls. For activity level indicators, we used step count and energy expenditure, which we calculated from accelerometer readings. Both indicators differed significantly between the patient and control study groups. There are already many studies and meta-analyses that have demonstrated the significant difference between healthy controls and mood disorder patients based on activity levels measured from a three-axis actigraphy device (Teychenne et al., 2008). On the other hand, step count is an activity marker that is normally measured by a single up/down axis accelerometer called a pedometer, but among studies comparing mood disorder patients and healthy controls, there are few that have measured step count using actigraphy. McKercher et al. (2009) reported that a pedometer was used to measure the steps taken per day by young adults, and that the prevalence of depression was higher when one's step count was lower. In our study, we found that healthy controls had significantly greater step counts and energy use during the hours of 11:00 am–6:00 pm, which agrees with the results of previous research.

Regarding sleep, our results showed that sleep time was particularly long among patients during the nighttime hours of 9:00 pm–12:00 am. Insomnia is common in depression (Benca and Peterson, 2008; Riemann and Voderholzer, 2003). Lack of sleep is very significant and is a major concern when treating depression, but there is also evidence that subjective complaints of insomnia and objective measurements of sleep time do not always align (Argyropoulos et al., 2003). Our results suggest that objective measurements for sleep showed higher levels of physical



	Steps	Energy Expenditure	Body Motion	Sleep State	Skin Temperature	Heart Rate	UV Light
Hour 0						░	
Hour 1							
Hour 2							
Hour 3				░			
Hour 4							
Hour 5							
Hour 6							
Hour 7							
Hour 8							
Hour 9							
Hour 10			░				
Hour 11		░	░		░		
Hour 12			░				
Hour 13			░				
Hour 14							░
Hour 15	░						
Hour 16	░						
Hour 17							
Hour 18							
Hour 19							
Hour 20			░	▤			
Hour 21							
Hour 22							
Hour 23							
One-day			░ *				

░ Minimum > Maximum $P < 0.05$

▤ Maximum > Minimum $P < 0.05$

* Significant difference remained after false discovery rate (FDR) for multiple comparisons.

Each cell indicates the result of comparisons for each outcome during each time interval.

Figure 2. Results of biostatistical comparisons within patient groups during each time interval.

calmness during nighttime hours in depressed patients than healthy controls. This may be because patients calm their physical movement earlier in the day compared to healthy controls, or because patients' social activity levels are lower.

Our results also showed a significant difference between mood disorder patients and healthy controls in the heart rate data collected by the wearable devices' sensors. In particular, heart rates captured during the sleep hours of 1:00 am–9:00 am show that patients have significantly higher heart rates than healthy controls. Upon investigation, we found numerous studies on depression and heart rate variability (Bassett, 2016; Kemp et al., 2010; Kwon et al., 2019; Stapelberg et al., 2012; Udupa et al., 2007; Wang et al., 2013), but found very little prior research regarding the relationship between heart rate itself and depression. For example, Kemp et al. (2014) reported that there was no significant difference between depressed patients and healthy controls based on heart rate data taken from 10-minute resting-state ECG tests. Additionally, Carney et al. (2016) reported that when observing depression symptoms in coronary artery disease patients, patients with high heart rates at night had a poorer response to depression treatment. From a biological standpoint, the reason that a depression patient's resting heart rate rises may be due to the fact that the automatic nervous system manages one's resting heart rate. Based on other studies, it is well known that as stress increases, the hypothalamic-pituitary-adrenal (HPA) axis is activated, which throws the automatic nervous system into disarray causing a rise in heart rate (Agelink et al., 2004; Juruena et al., 2018; Nederhof et al., 2015; Ulrich-Lai and Herman, 2009). In this study, as a result of collecting heart rate data over seven days (including nights), we observed a significant difference in heart rates between patients and healthy controls, although this difference did not remain in the age-balanced subpopulation.

In regards to skin temperature, Avery et al. (1999) reported that patients with depression have higher body temperatures at night than healthy controls, and that after recovery, patients' temperatures decrease. Our results also suggest that around the time period of 7:00 pm–11:00 pm, patients' skin temperature increases significantly, which means skin temperature is a possible indicator for depression.

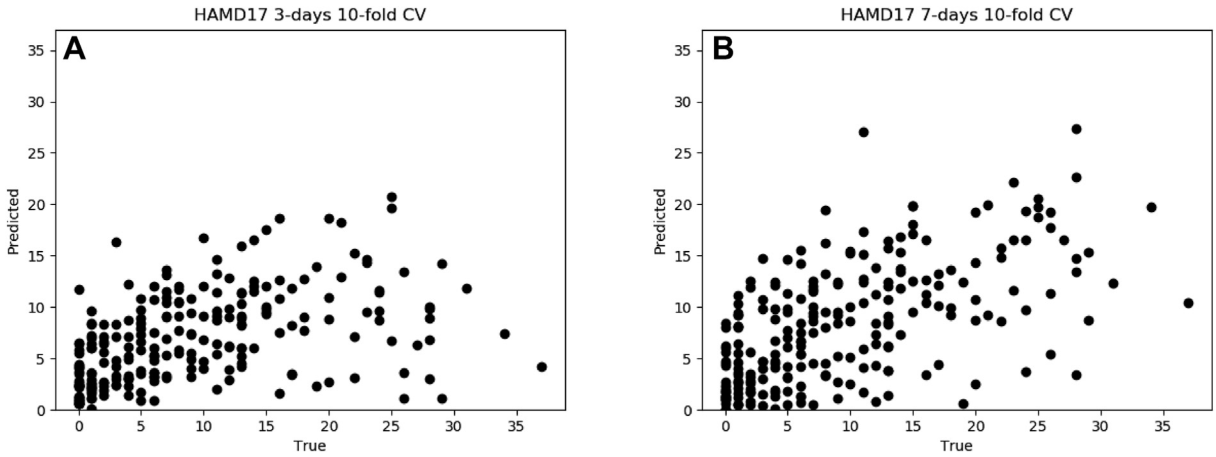


Figure 3. A. Machine learning severity predictions based on three days of wearable data. B. Machine learning severity predictions based on seven days of wearable data.

Table 2. Feature importance across 10-fold cross validation.

Model	Screening using 7 days' data		Severity prediction using 7 days' data		Screening using 3 days' data		Severity prediction using 3 days' data	
	Feature	Importance	Feature	Importance	Feature	Importance	Feature	Importance
1st	Skin - 95%	0.0462	Skin- 95%	0.0340	Corr - Sleep & Skin	0.0453	Corr - Sleep & Skin	0.0427
2nd	Skin - 75%	0.0443	Corr- Sleep & Skin	0.0288	Skin - 50%	0.0439	Skin - 50%	0.0419
3rd	Sleep - SD	0.0416	Skin - 50%	0.0284	Motion - SD	0.0434	Skin - 5%	0.0403
4th	Corr - HR & UV	0.0415	Corr - Sleep & HR	0.0283	Step - 95%	0.0388	Step - 95%	0.0305
5th	Corr - Step & Energy	0.0401	Corr - Energy & Motion	0.0275	Corr - Sleep & UV	0.0364	Corr - Motion & Skin	0.0290
6th	UV - SD	0.0392	Corr- HR & UV	0.0268	Motion - 50%	0.0312	Corr - Motion & Sleep	0.0279
7th	Corr - Energy & UV	0.0377	Energy - 95%	0.0259	Sleep - 50%	0.0296	Sleep - 50%	0.0278
8th	Corr - Step & HR	0.0343	Motion - 50%	0.0250	Skin - 5%	0.0292	HR - 75%	0.0271
9th	Energy - 25%	0.0325	Skin - 5%	0.0249	Corr - Energy & HR	0.0278	Motion - 50%	0.0265
10th	Corr - Sleep & HR	0.0317	Energy - 50%	0.0249	Corr - Skin & HR	0.0271	Corr - Sleep & HR	0.0258

Note: % = percentile, corr = correlation, energy = energy expenditure, HR = heart rate, motion = body motion, SD = standard deviation, skin = skin temperature, sleep = sleep time, step = step count, UV = ultraviolet light exposure.

For UV light exposure, we found only a small difference for UV light exposure between depressed patients and healthy controls. There are several studies that have reported that UV light exposure has an effect on people's emotional state, and UV light exposure has been used in the treatment of depression (Veleva et al., 2018). Additionally, UV light is almost completely absent from interior light sources, but makes up a large part of light from the sun. Therefore, we believe that UV light exposure can indicate when someone goes outside, and so we used it as one of our measures in this study. It is possible that the lack of a significant difference in UV light exposure is due to the fact that much of modern life takes place indoors these days, which causes UV light exposure to be low overall regardless of depressive state.

Next, using individual longitudinal patient data, we compared data taken from the wearable devices regarding the worsening and lessening of symptoms. From those results, we observed a tendency for daytime step count, energy expenditure, and body motion to be high during times of less severe symptoms, which shows that when depression symptoms are less severe, patients are more active during the day.

For sleep time, we found that during more severe symptom periods, patients slept comparatively more during earlier nighttime hours, whereas during less severe periods, patients slept more during later nighttime hours. We believe these results show that patients with less severe symptoms are sleeping well late at night, and when compared with healthy controls, both groups maintain similarly high regular activity levels earlier at night.

Additionally, using machine learning, we created algorithms to evaluate the presence and severity of depression symptoms, and tested those algorithms' accuracy. For evaluating the presence of depression, we achieved an accuracy value of over 0.7 using data measured over a three-day period. When we increased the data collection period to seven days, we did not observe an increase in accuracy. On the other hand, when using machine learning to evaluate illness severity, we reached a correlation coefficient of 0.48 for three days of data, whereas we achieved a correlation coefficient of 0.61 when analyzing seven days of data. Based on these results, we believe that data collection over three days is adequate for diagnosing, but longer periods are needed for estimating illness severity.

Until now, there have been few studies attempting to use machine learning to analyze multimodal data collected from wearable devices in order to evaluate mood disorders. As stated in the introduction, Valenza et al. (2013) and Cho et al. (2019) used machine learning to analyze wearable device data to predict mood states. Bourla et al. (2018) looked at the usefulness of unipolar depression evaluations utilizing wearable devices, and found that features like HRV and body temperature are useful in diagnosing depressive episodes.

There are relatively many more studies so far that have evaluated mood disorders using biological data taken from sources other than wearable devices. Lee et al. (2018) conducted a meta-analysis of 26 studies that used a variety of predictors to create machine learning algorithms for estimating the effectiveness of depression treatments. According to this meta-analysis, most studies used the following types of predictors: neuroimaging, phenomenological (e.g., psychometric, neurocognitive, anthropometric, sociodemographic, psychiatric history), genetic (e.g., single nucleotide polymorphisms [SNPs]), or a combination of the above. The meta-analysis reported that the combined results of these studies produced a prediction accuracy of 0.82. In a separate study, Ramasubbu et al. (2016) attempted to differentiate between MDD patients and healthy controls using fMRI data and a support vector machine. They reported that while they were able to discern between patients with the heaviest symptoms and healthy controls (accuracy 66%, $p = 0.012$ corrected), they were unable to discern between patients with heavy symptoms (accuracy 52%, $p = 1.0$ corrected) and those with mild to moderate symptoms (accuracy 58%, $p = 1.0$ corrected). Thus, the majority of previous research using machine learning has focused on labor-intensive brain image studies and genetic studies that involve large amounts of data inspection. But the wristband-type wearable device used in our study allows data for evaluating the presence and severity of depression to be collected more easily, which may be beneficial in real world clinical practice.

Furthermore, there are few studies that have used machine learning to estimate depression severity. Jiang et al. (2016) predicted illness severity based on HAMD scores using machine learning to analyze magnetoencephalography (MEG) data, and they reported a correlation coefficient of $r = 0.38–0.68$. However, since only 22 datasets were analyzed with machine learning in that study, it is possible that it lacks reliability and generalizability. Therefore, the results of our study, which included 236 datasets and estimates illness severity with a correlation coefficient of 0.61, may be meaningful.

We found that skin temperature-related features consistently showed the greatest contribution to the machine learning algorithm's predictive ability throughout all the models we built. Following skin temperature, sleep time-related features had the next highest significance in making predictions. In biostatistical tests, instead of skin temperature or sleep time, heart rate was shown to have significant differences between healthy and depressed samples most frequently. Therefore, it is interesting to find that skin temperature, sleep time, and the correlation between skin temperature and sleep time contribute more to machine learning predictions than heart rate. One possible explanation is that the statistical tests that we employed in this study test the differences in mean and standard deviation of each feature individually. It is likely that the nonlinear combinatory relationships between skin temperature, sleep time, and depression state were not discovered in the statistical tests, but were uncovered by the machine learning model.

Regarding the relationship between body temperature and sleep, it is reported that body temperature shifts with sleep, and the relationship can be disturbed by depression (Avery et al., 1999; Elsenga and Van den Hoofdakker, 1988; Lorenz et al., 2019). Hasler et al. (2010) compared 18 MDD patients and 19 healthy controls, and found that a larger phase angle difference between midsleep and the core body temperature minimum was associated with greater depression.

One common drawback to machine learning models is that results are often not easily explained in biological or clinical terms. However, the predictive capabilities of machine learning models can still be extremely valuable. Also, in this case, we averaged the feature importance of features across the 10-fold cross-validation to find the features that are important in all of the models that we built, and referenced these features to findings in previously published research works. The results show that the consistently important features have been collaborated by previous research, which further confirms the validity of the machine learning models.

5. Limitations

The results of this study have to be interpreted in the context of a number of limitations. First, the small number of subjects who participated in this study may have weakened the statistical significance of the study. In particular, the number of patients included in the longitudinal comparison was small.

Second, demographic data (e.g., age, sex, etc.) was not matched between patients and healthy controls. Therefore, differences in the demographic data may have had some impact on the study results, although age was included as a covariate in a regression model. While this is the case, we did not have the same issue as stated above with longitudinal comparisons when comparing patients' symptomatic and asymptomatic periods. Furthermore, because all data is included in the machine learning analysis – including data from both times when depressive symptoms are present or absent – there may be a lower likelihood of the results being affected by demographic data.

Third, hospitalized patients live in circumstances unique to life in a hospital. It is possible that such uncommon daily patterns had an effect on the data collected by the wearable device. However, we did not find any significant differences in depression severity between inpatients and outpatients in this study, so we believe hospitalization did not have a large effect on the results. Additionally, we did not exclude participants who had unique occupational circumstances (e.g., night shift worker, etc.), which may have affected the data results. Similarly, medication may have influenced the data acquired by the wearable device, such as heart rate and sleep status, because this study collected the data in an observational manner.

Fourth, because we excluded patients with complications or illnesses other than depression, it is possible that in actual clinical practice, such additional illnesses or symptoms could cause prediction accuracy to decrease.

Fifth, there are limitations associated with the machine learning technique that we used for this study. Due to the small sample size, the trained model is likely to be overfitted. We have tried to alleviate the effect of overfitting through 10-fold cross-validation and tuning the sparse parameter in XGBoost to choose a simpler model; however, cross-validation is a method that makes predictions on data that is already known, so it is still likely that the model would perform worse than the validation accuracy on new samples (Zhang and Yang, 2015).

Finally, validation data for the mechanical performance of the wearable device used in this study was disclosed in part by the device's maker, but there have been no third-party trials for the device.

6. Conclusion

In this study, we used a wristband-type wearable device to record various outcomes, and then utilized machine learning to predict the presence and severity of depressive state based on that data. In doing so, we were able to achieve results with a relatively highly accurate prediction rate. The current accuracy level of our models may not make them suitable for immediate implementation in psychiatric clinical practice, but advantages include using the algorithms to create a wearable device for 24-hour non-invasive monitoring, using them in a family doctor environment, or using them as a screening tool. Moreover, by investigating the differences between the patients and healthy control groups, we found that skin temperature may contribute greatly to predicting depressive state. In the future, it would be desirable to gather a larger study population to research whether wearable devices can be used to judge depressive state in clinical settings and other contexts, while simultaneously considering the possible effects of demographic data such as age, medication, etc.

Declarations

Author contribution statement

T. Kishimoto: conceived and designed the experiments; performed the experiments; wrote the paper.

Y. Tazawa: conceived and designed the experiments; performed the experiments; analyzed and interpreted the data; contributed reagents, materials, analysis tools, or data; wrote the paper.

K. Liang: analyzed and interpreted the data; contributed reagents, materials, analysis tools, or data; wrote the paper.

M. Yoshimura, M. Kitazawa, Y. Kaise: performed the experiments; contributed reagents, materials, analysis tools, or data.

A. Takamiya: conceived and designed the experiments; contributed reagents, materials, analysis tools, or data; wrote the paper.

A. Kishi, Y. Mitsukura: analyzed and interpreted the data; contributed reagents, materials, analysis tools, or data.

T. Horigome, M. Mimura: conceived and designed the experiments.

Funding statement

Masaro Mimura has received grants and/or speaker's honoraria from Daiichi Sankyo, Dainippon-Sumitomo Pharma, Eisai, Eli Lilly, Fuji Film RI Pharma, Janssen Pharmaceutical, Mochida Pharmaceutical, MSD, Nippon Chemipher, Novartis Pharma, Ono Pharma, Otsuka Pharmaceutical, Pfizer, Takeda Pharma, Tsumura, and Yoshitomi Pharma within the past three years. Taishiro Kishimoto has received consultant fees from Otsuka, Pfizer, and Dainippon Sumitomo, and speaker's honoraria from Banyu, Eli Lilly, Dainippon Sumitomo, Janssen, Novartis, Otsuka, and Pfizer. This research was supported by the Japan Agency for Medical Research and Development (AMED) under Grant Number JP18he1102004.

Competing interest statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

References

Agelink, M.W., Klimke, A., Cordes, J., Sanner, D., Kavuk, I., Malessa, R., Klieser, E., Baumann, B., 2004. A functional-structural model to understand cardiac autonomic nervous system (ANS) dysregulation in affective illness and to elucidate the ANS effects of antidepressive treatment. Eur. J. Med. Res. 9, 37–50.

Argyropoulos, Spilios V., Hicks, Jane A., Nash, Jon R., 2003. Correlation of subjective and objective sleep measurements at different stages of the treatment of depression. Psychiatr. Res. 120, 179–190.

Avery, D., Shah, S., Eder, D., Wildschisdtz, G., 1999. Nocturnal sweating and temperature in depression. Acta Psychiatr. Scand. 100, 295–301.

Bassett, D., 2016. A literature review of heart rate variability in depressive and bipolar disorders. Aust. N. Z. J. Psychiatr.

Beijers, L., Wardenaar, K.J., Bosker, F.J., Lamers, F., van Grootheest, G., de Boer, M.K., Penninx, B.W.J.H., Schoevers, R.A., 2019. Biomarker-based subtyping of depression and anxiety disorders using Latent Class Analysis. A NESDA study. Psychol. Med. 49, 617–627.

Benca, R.M., Peterson, M.J., 2008. Insomnia and depression. Sleep Med. 9, S3–S9.

Bourla, A., Ferreri, F., Ogorzelec, L., Guinchard, C., Mouchabac, S., 2018. Assessment of mood disorders by passive data gathering: the concept of digital phenotype versus psychiatrist's professional culture. Encephale 44, 168–175.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Burton, C., McKinstry, B., Szentagotai Tătar, A., Serrano-Blanco, A., Pagliari, C., Wolters, M., 2013. Activity monitoring in patients with depression: a systematic review. J. Affect. Disord. 145, 21–28.

Carney, R.M., Freedland, K.E., Steinmeyer, B.C., Rubin, E.H., Stein, P.K., Rich, M.W., 2016. Nighttime heart rate predicts response to depression treatment in patients with coronary heart disease. J. Affect. Disord. 200, 165–171.

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System.

Cho, C.-H., Lee, T., Kim, M.-G., In, H.P., Kim, L., Lee, H.-J., 2019. Mood prediction of patients with mood disorders by machine learning using passive digital phenotypes based on the circadian rhythm: prospective observational cohort study. J. Med. Internet Res. 21, e11029.

Dogan, E., Sander, C., Wagner, X., Hegerl, U., Kohls, E., 2017. Smartphone-based monitoring of objective and subjective data in affective disorders: where are we and where are we going? Systematic review. J. Med. Internet Res. 19, e262.

Elsenga, S., Van den Hoofdakker, R.H., 1988. Body core temperature and depression during total sleep deprivation in depressives. Biol. Psychiatr. 24, 531–540.

Hamilton, M., 1960. A rating scale for depression. J. Neurol. Neurosurg. Psychiatr.

Hasler, B.P., Buysse, D.J., Kupfer, D.J., Germain, A., 2010. Phase relationships between core body temperature, melatonin, and sleep are associated with depression severity: further evidence for circadian misalignment in non-seasonal depression. Psychiatr. Res. 178, 205–207.

Jiang, H., Popov, T., Jylänki, P., Bi, K., Yao, Z., Lu, Q., Jensen, O., van Gerven, M.A.J., 2016. Predictability of depression severity based on posterior alpha oscillations. Clin. Neurophysiol. 127, 2108–2114.

Juruena, M.F., Bocharova, M., Agustini, B., Young, A.H., 2018. Atypical depression and non-atypical depression: is HPA axis function a biomarker? A systematic review. J. Affect. Disord. 233, 45–67.

Kemp, A.H., Quintana, D.S., Gray, M.A., Felmingham, K.L., Brown, K., Gatt, J.M., 2010. Impact of depression and antidepressant treatment on heart rate variability: a review and meta-analysis. Biol. Psychiatr. 67, 1067–1074.

Kemp, A.H., Brunoni, A.R., Santos, I.S., Nunes, M.A., Dantas, E.M., Carvalho de Figueiredo, R., Pereira, A.C., Ribeiro, A.L., Mill, J.G., Andreão, R.V., Thayer, J.F., Benseñor, I.M., Lotufo, P.A., 2014. Effects of depression, anxiety, comorbidity, and antidepressants on resting-state heart rate and its variability: an ELSA-Brasil cohort baseline study. Am. J. Psychiatr.

Kimura, N., Aso, Y., Yabuuchi, K., Ishibashi, M., Hori, D., Sasaki, Y., Nakamichi, A., Uesugi, S., Fujioka, H., Iwao, S., Jikumaru, M., Katayama, T., Sumi, K., Eguchi, A., Nonaka, S., Kakumu, M., Matsubara, E., 2019. Modifiable lifestyle factors and cognitive function in older people: a cross-sectional observational study. Front. Neurol. 10, 401.

Kishimoto, T., Takamiya, A., Liang, K., Funaki, K., Fujita, T., Kitazawa, M., Yoshimura, M., Tazawa, Y., Horigome, T., Eguchi, Y., Kikuchi, T., Tomita, M., Bun, S., Murakami, J., Sumali, B., Warnita, T., Kishi, A., Yotsui, M., Toyoshiba, H., Mitsukura, Y., Shinoda, K., Sakakibara, Y., Mimura, M., 2019. The project for objective measures using computational psychiatry technology (PROMPT): rationale, design, and methodology. medRxiv 19013011.

Kuehn, B.M., 2016. Wearable biosensors studied for clinical monitoring and treatment. JAMA 316, 255.

Kwon, H. Bin, Yoon, H., Choi, S.H., Choi, J.W., Lee, Y.J., Park, K.S., 2019. Heart rate variability changes in major depressive disorder during sleep: fractal index correlates with BDI score during REM sleep. Psychiatr. Res. 271, 291–298.

Lee, Y., Ragguett, R.M., Mansur, R.B., Boutilier, J.J., Rosenblat, J.D., Trevizol, A., Brietzke, E., Lin, K., Pan, Z., Subramaniapillai, M., Chan, T.C.Y., Fus, D., Park, C., Musial, N., Zuckerman, H., Chen, V.C.H., Ho, R., Rong, C., McIntyre, R.S., 2018. Applications of machine learning algorithms to predict therapeutic outcomes in depression: a meta-analysis and systematic review. J. Affect. Disord.

Licht, C.M.M., de Geus, E.J.C., Zitman, F.G., Hoogendijk, W.J.G., van Dyck, R., Penninx, B.W.J.H., 2008. Association between major depressive disorder and heart rate variability in The Netherlands study of depression and anxiety (NESDA). Arch. Gen. Psychiatr. 65, 1358.

Lorenz, N., Spada, J., Sander, C., Riedel-Heller, S.G., Hegerl, U., 2019. Circadian skin temperature rhythms, circadian activity rhythms and sleep in individuals with self-reported depressive symptoms. J. Psychiatr. Res. 117, 38–44.

Luik, A.I., Zuurbier, L.A., Direk, N., Hofman, A., Van Someren, E.J.W., Tiemeier, H., 2015. 24-HOUR activity rhythm and sleep disturbances IN depression and anxiety: a population-based study OF middle-aged and older persons. Depress. Anxiety 32, 684–692.

Martin, J.L., Hakim, A.D., 2011. Wrist actigraphy. Chest 139, 1514–1527.

Marzano, L., Bardill, A., Fields, B., Herd, K., Veale, D., Grey, N., Moran, P., 2015. The application of mHealth to mental health: opportunities and challenges. Lancet Psychiatry 2, 942–948.

Mazzetta, I., Gentile, P., Pessione, M., Suppa, A., Zampogna, A., Bianchini, E., Irrera, F., Mazzetta, I., Gentile, P., Pessione, M., Suppa, A., Zampogna, A., Bianchini, E., Irrera, F., 2018. Stand-alone wearable system for ubiquitous real-time monitoring of muscle activation potentials. Sensors 18, 1748.

McKercher, C.M., Schmidt, M.D., Sanderson, K.A., Patton, G.C., Dwyer, T., Venn, A.J., 2009. Physical activity and depression in young adults. Am. J. Prev. Med. 36, 161–164.

Montgomery, S., Asberg, M., 1979. A new depression scale designed to be sensitive to change. Br. J. Psychiatry.

Moraes, C.T., Cambras, T., Diez-Noguera, A., Schimitt, R., Dantas, G., Levandovski, R., Hidalgo, M.P., 2013. A new chronobiological approach to discriminate between acute and chronic depression using peripheral temperature, rest-activity, and light exposure parameters. BMC Psychiatr. 13.

Nederhof, E., Marceau, K., Shirtcliff, E.A., Hastings, P.D., Oldehinkel, A.J., 2015. Autonomic and adrenocortical interactions predict mental health in late adolescence: the TRAILS study. J. Abnorm. Child Psychol. 43, 847–861.

Nieto-Riveiro, L., Groba, B., Miranda, M.C., Concheiro, P., Pazos, A., Pousada, T., Pereira, J., 2018. Technologies for participatory medicine and health promotion in the elderly population. Medicine (Baltim.) 97, e10791.

Patel, M.S., Foschini, L., Kurtzman, G.W., Zhu, J., Wang, W., Rareshide, C.A.L., Zbikowski, S.M., 2017. Using wearable devices and smartphones to track physical activity: initial activation, sustained use, and step counts across sociodemographic characteristics in a national sample. Ann. Intern. Med. 167, 755.

Puiatti, A., Mudda, S., Giordano, S., Mayora, O., 2011. Smartphone-centred wearable sensors network for monitoring patients with bipolar disorder. Conf. Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Conf. 2011, 3644–3647.

Ramasubbu, R., Brown, M.R.G., Cortese, F., Gaxiola, I., Goodyear, B., Greenshaw, A.J., Dursun, S.M., Greiner, R., 2016. Accuracy of automated classification of major depressive disorder as a function of symptom severity. NeuroImage Clin. 12, 320–331.

Reinertsen, E., Clifford, G.D., 2018. A review of physiological and behavioral monitoring with digital sensors for neuropsychiatric illnesses. Physiol. Meas. 39, 05TR01.

Riemann, D., Voderholzer, U., 2003. Primary insomnia: a risk factor to develop depression? J. Affect. Disord. 76, 255–259.

Rohani, D.A., Faurholt-Jepsen, M., Kessing, L.V., Bardram, J.E., 2018. Correlations between objective behavioral features collected from mobile and wearable devices and depressive mood symptoms in patients with affective disorders: systematic review. JMIR mHealth uHealth 6, e165.

Saad, M., Ray, L.B., Bujaki, B., Parvaresh, A., Palamarchuk, I., De Koninck, J., Douglass, A., Lee, E.K., Soucy, L.J., Fogel, S., Morin, C.M., Bastien, C., Merali, Z., Robillard, R., 2019. Using heart rate profiles during sleep as a biomarker of depression. BMC Psychiatr. 19, 168.

Shelgikar, A.V., Anderson, P.F., Stephens, M.R., 2016. Sleep tracking, wearable technology, and opportunities for research and clinical care. Chest 150, 732–743.

Stapelberg, N.J., Hamilton-Craig, I., Neumann, D.L., Shum, D.H., McConnell, H., 2012. Mind and heart: heart rate variability in major depressive disorder and coronary heart disease - a review and recommendations. Aust. N. Z. J. Psychiatr.

Tazawa, Y., Wada, M., Mitsukura, Y., Takamiya, A., Kitazawa, M., Yoshimura, M., Mimura, M., Kishimoto, T., 2019. Actigraphy for evaluation of mood disorders: a systematic review and meta-analysis. J. Affect. Disord.

Teychenne, M., Ball, K., Salmon, J., 2008. Physical activity and likelihood of depression in adults: a review. Prev. Med. (Baltim) 46, 397–411.

Udupa, K., Sathyaprabha, T.N., Thirthalli, J., Kishore, K.R., Lavekar, G.S., Raju, T.R., Gangadhar, B.N., 2007. Alteration of cardiac autonomic functions in patients with major depression: a study using heart rate variability measures. J. Affect. Disord. 100, 137–141.

Ulrich-Lai, Y.M., Herman, J.P., 2009. Neural regulation of endocrine and autonomic stress responses. Nat. Rev. Neurosci. 10, 397–409.

Valenza, G., Gentili, C., Lanatà, A., Scilingo, E.P., 2013. Mood recognition in bipolar patients through the PSYCHE platform: preliminary evaluations and perspectives. Artif. Intell. Med. 57, 49–58.

Valenza, G., Nardelli, M., Lanatà, A., Gentili, C., Bertschy, G., Paradiso, R., Scilingo, E.P., 2014. Wearable monitoring for mood recognition in bipolar disorder based on history-dependent long-term heart rate variability analysis. IEEE J. Biomed. Health Inf. 18, 1625–1635.

Valenza, G., Citi, L., Gentili, C., Lanata, A., Scilingo, E.P., Barbieri, R., 2015. Characterization of depressive States in bipolar patients using wearable textile technology and instantaneous heart rate variability assessment. IEEE J. Biomed. Health Inf. 19, 263–274.

van Londen, L., Goekoop, J.G., Kerkhof, G.A., Zwinderman, K.H., Wiegant, V.M., De Wied, D., 2001. Weak 24-h periodicity of body temperature and increased plasma vasopressin in melancholic depression. Eur. Neuropsychopharmacol. 11, 7–14.

Veleva, B.I., van Bezooijen, R.L., Chel, V.G.M., Numans, M.E., Caljouw, M.A.A., 2018. Effect of ultraviolet light on mood, depressive disorders and well-being. Photodermatol. Photoimmunol. Photomed. 34, 288–297.

Wang, Y., Zhao, X., O'Neil, A., Turner, A., Liu, X., Berk, M., 2013. Altered cardiac autonomic nervous function in depression. BMC Psychiatr. 13.

Wang, R., Wang, W., daSilva, A., Huckins, J.F., Kelley, W.M., Heatherton, T.F., Campbell, A.T., 2018. Tracking depression dynamics in college students using mobile phone and wearable sensing. Proc. ACM Interact. Mobile Wearable Ubiquitous Technol. 2, 1–26.

Zhang, Y., Yang, Y., 2015. Cross-validation for selecting a model selection procedure. J. Econom. 187, 95–112.