Wearable Computing Technology for Assessment of Cognitive Functioning of Bipolar Patients and Healthy Controls

PEGAH HAFIZ, Technical University of Denmark, Denmark
KAMILLA WOZNICA MISKOWIAK, University of Copenhagen, Denmark
ALBAN MAXHUNI, Technical University of Denmark, Denmark
LARS VEDEL KESSING, University of Copenhagen, Denmark
JAKOB EYVIND BARDRAM, Technical University of Denmark, Denmark

Mobile cognitive tests have been emerged to first, bring the assessments outside the clinics and second, frequently measure individuals' cognitive performance in their free-living environment. Patients with Bipolar Disorder (BD) suffer from cognitive impairments and poor sleep quality negatively affects their cognitive performance. Wearables are capable of unobtrusively collecting multivariate data including activity and sleep features. In this study, we analyzed daily attention, working memory, and executive functions of patients with BD and healthy controls by using a smartwatch-based tool called UbiCAT to 1) investigate its concurrent validity and feasibility, 2) identify digital phenotypes of mental health using cognitive and mobile sensor data, and 3) classify patients and healthy controls on the basis of their daily cognitive and mobile data. Our findings demonstrated that UbiCAT is feasible with valid measures for *in-the-wild* cognitive assessments. The analysis showed that the patients responded more slowly during the attention task than the healthy controls, which could indicate a lower alertness of this group. Furthermore, sleep duration correlated positively with participants' working memory performance the next day. Statistical analysis showed that features including cognitive measures of attention and executive functions, sleep duration, time in bed, awakening frequency and duration, and step counts are the digital phenotypes of mental health diagnosis. Supervised learning models was used to classify individuals' mental health diagnosis using their daily observations. Overall, we achieved accuracy of approximately 74% using K-Nearest Neighbour (KNN) method.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing design and evaluation methods**; **Mobile devices**; **Empirical studies in ubiquitous and mobile computing**.

Additional Key Words and Phrases: cognition, wearable technology, mobile sensing, mental health, bipolar disorder, digital phenotype

ACM Reference Format:
Pegah Hafiz, Kamilla Woznica Miskowiak, Alban Maxhuni, Lars Vedel Kessing, and Jakob Eyvind Bardram. 2020. Wearable Computing Technology for Assessment of Cognitive Functioning of Bipolar Patients and Healthy Controls. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 4, Article 129 (December 2020), 22 pages. https://doi.org/10.1145/3432219

1 INTRODUCTION

Cognitive functioning of individuals' attention, memory, and executive skills characterize the quality of their daily tasks. The common practice in psychiatry is to assess patients' cognitive functioning using neuropsychological

Authors' addresses: Pegah Hafiz, pegh@dtu.dk, Technical University of Denmark, Lyngby, Denmark, 2800; Kamilla Woznica Miskowiak, University of Copenhagen, Copenhagen, Denmark, 2100, kamilla.woznica.miskowiak@regionh.dk; Alban Maxhuni, Technical University of Denmark, Lyngby, Denmark, 2800, almax@dtu.dk; Lars Vedel Kessing, University of Copenhagen, Copenhagen, Denmark, 2100, lars.vedel.kessing@regionh.dk; Jakob Eyvind Bardram, Technical University of Denmark, Lyngby, Denmark, 2800, jakba@dtu.dk.

2474-9567/2020/12-ART129
https://doi.org/10.1145/3432219

tests. Such experiments are run in a controlled environment and at a certain time of a day suitable for the clinician and patient. However, a fixed environment and context for taking cognitive tests may negatively impact the validity and reliability of the test results [2] since human cognition fluctuates during the day [55, 68]. In particular, patients with Bipolar Disorder (BD) (mania and depression) suffer from cognitive impairment even during their period of symptom remission [9, 61]. However, time and resource constraints hinder continuous and frequent monitoring of patients' cognitive functioning. Therefore, novel computing technologies are essential to obtain cognitive performance measures over time.

A few smartphone-based tools have been proposed for measuring individuals' cognitive functioning outside the clinic [10, 30, 48, 62]. Although these tools have contributed to mobile assessments, the use of wearables offers two advantages over smartphones. First, wearables can be used to collect reliable data on physical activity (e.g., step count) and physiological data (e.g., heart rate and sleep). Second, wearables are devices that people wear all the time and during most activities such as walking and running. Recently, smartwatch-based tools have been developed to frequently assess cognitive functions [14, 24]. Taken together, wearables provide an opportunity in conducting *in-the-wild* studies for collecting multivariate sensor data in conjunction with cognitive test performance measures. It is, however, essential to evaluate the feasibility of using such tools and their concurrent validity compared with the gold-standard neuropsychological tests.

Active and passive data collected via mobile devices can assist in identifying digital phenotype of human mental health [47]. Wearables are capable of unobtrusively collecting various data types. For instance, daily step counts, physical activities, and sleep duration per cycle are calculated by Fitbit trackers. There is evidence that sleep quality affects individuals' cognitive performance during a day [21, 42]. Particularly, patients with BD can potentially suffer from the negative consequences of their frequent poor sleep quality [5]. To date, digital behavioural phenotypes of individuals' mental health have been investigated (for example, [13, 19, 20, 53, 54, 64, 65]). Yet, we do not know what cognitive, behavioural, and physiological features are significant for mental health diagnosis.

In this study, first, we show the concurrent validity and feasibility of an *in-the-wild* cognitive assessment tool developed for Fitbit Ionic smartwatches. Then, we collect daily cognitive performance measures as well as activity and sleep features using the smartwatch to 1) investigate the impact of sleep on the next-day cognitive performance measures and 2) identify digital phenotypes of individuals' mental health diagnosis and 3) classify patients with BD and healthy controls using supervised learning methods.

2 RELATED WORK

A number of studies have shown the feasibility of mobile cognitive assessments by using Personal Digital Assistants (PDAs), cellphones, or smartphones [45]. Table 1 gives an overview of studies that have used smartphone or smartwatch technology for cognitive testing. These studies have all adopted the Ecological Momentary Assessment (EMA) [59] or Experience Sampling Method (ESM) [37] methodology (which are often mentioned interchangeably). Prior related work has been focusing on collecting self-reports on mood [14, 15, 29, 63], sleep [1, 15, 18, 29], activity [15], location [15, 63], and alertness [1, 18]. No effect of sleep quality, mood, location, or activity stress was found on the cognitive test results in [15]. An *'in-the-wild'* study investigated individuals' alertness and showed the effectiveness of using mobile cognitive tasks in detecting circadian variations [18]. However, self-reports on sleep duration and quality did not affect the cognitive test measures. On the other hand, a similar study reported a negative impact of poor sleep on individuals' alertness [1].

Previous studies mostly collected self-reported, subjective measures of sleep and behavioural features. In this study, we collect objective sleep data using Fitbit smartwatches, which has shown acceptable performance in differentiating sleep and wake cycles [25] and in estimating the accuracy of sleep stage [26]. Activity features (e.g., step count) are also collected passively using the Fitbit smartwatches. Of the recent studies performed to assess *in-the-wild* cognition, two measured objective attention [1, 18], one evaluated working memory [14],

Table 1. Overview of the studies conducted with mobile devices for cognitive assessments.

Study	Device	Participants	Sampling	Duration	Cognitive tasks
Timmers et al. [63]	Smartphone	Young adults (N=26)	4 times daily	1 day	Letter span
Sliwinski et al. [56]	Smartphone	Adults (N=219)	5 times daily	14 days	Symbol search, Dot memory
Abdullah et al. [1]	Smartphone	Students (N=40)	2 times daily	40 days	Psychomotor vigilance test
Dingler et al. [18]	Smartphone	Students (N=12)	1–6 times daily	2–13 days	Psychomotor vigilance, Go No-Go, Multiple object tracking
Daniels et al. [15]	Smartphone	Healthy adults (N=49)	8 times daily	6 days	Visuospatial working memory, Digit symbol substitution
Hung et al. [29]	Smartphone	Depression (N=54)	once per week	8 weeks	Stroop, TMT part B
Cormack et al. [14]	Smartwatch	Depression (N=30)	3 times daily	6 weeks	N-back

and one assessed working memory and psychomotor speed [15]. We extend this work by measuring three key cognitive domains, namely, attention, working memory, and executive functions. Our work reports the findings of a clinical *in-the-wild* feasibility study conducted with healthy controls and patients with BD, and achieves the following contributions:

- Demonstrating the concurrent validity of the smartwatch-based tool in assessing individuals' cognitive functioning.
- Showing the feasibility of a smartwatch-based tool for continuous, daily, *in-the-wild* administration of cognitive assessment tests.
- Investigating the relationship between sleep duration and individuals' cognitive functioning the following day.
- Identifying digital phenotypes of human mental health using daily cognitive tests combined with mobile and wearable sensor data.
- Classifying patients with BD and healthy controls using their daily observations.

3 METHODOLOGY

The methodology of our study is adapted from the EMA/ESM approach and aims at collecting active cognitive performance measures and passive mobile sensor data. This study was exempted for ethical approval by the Danish ethics committee (Journal-nr.: H-19086232). All participants were informed about the types of data collected during the study and signed an informed consent before enrolled in the study (see also Figure 2). We used a cognitive assessment tool developed for the Fitbit Ionic smartwatches called UbiCAT [23, 24] that collects daily cognitive performance measures as well as behavioural, contextual, and physiological data. The cognitive tests of this tool are choice reaction time to measure attention, N-back to evaluate working memory, and Stroop color-word test to assess executive functions. Each cognitive test of this tool is a standalone smartwatch-based app. The snapshots and description of the smartwatch-based cognitive tests are presented in Figure 1. The

choice reaction time test has 40 arrows that appear on either right or left side of the watch screen. The arrows are right-hand or left-hand. The participants should select the direction of each arrow by tapping either the right-hand or left-hand rectangle (touch button) as shown in Figure 1a. The N-back test has three difficulty levels determined by the *N* value. This test shows a sequence of 40 letters one by one. Figure 1b shows a 2-back task, thus, the participant should tap on 'Yes', if the letter *G* had appeared 2 letters back in the sequence. The Stroop test displays 30 color names one by one where each stimuli is either congruent or incongruent. Figure 1c represents an incongruent stimuli as "Green" is written in pink. Participants should have responded in 2500 ms to each of the stimuli of the cognitive tests. The time limit in our study was the same as the standard computer-based tools, namely, PsyToolkit [58] and THINC-it [28] that implement the same cognitive tests as the smartwatch-based tool. Furthermore, we examined the viability of this time limit during several cognitive test sessions with some users to ensure that they could respond properly during this time limit.

Table 2 shows the features collected for this study as well as their associated characteristics. As can be seen, the performance measures of the cognitive tests were the number of correct responses and response times (RTs) to the test stimuli. The feature 'missed stimuli' in Table 2 refers to the number of stimuli in the test sessions to which the participant did not respond during a time limit (2500 ms). The hit rates and false alarm rates were also calculated for the N-back test. Hit rate is the number of times that the user correctly identified a match in the N-back divided by the total matches in the sequence. False alarm rate refers to the number of times that the user responded as a match while there was no match in the sequence. It should be noted that the RTs in the choice reaction time tests are the main performance measure to quantify alertness. Therefore, the outliers in the user's RTs can lead to incorrect assessment of alertness. However, the primary performance measure in the N-back and Stroop tests is the number of correct responses. Consequently, the median RT was calculated for the choice reaction time test since the median value is not sensitive to the outliers in the RTs in contrast to the mean values [44]. In addition, previous work [1, 6, 8] considered the median RT for the Psychomotor Vigilance Test, which is one of the cognitive tests in the choice reaction paradigm. As for the N-back and Stroop tests, the mean RT was calculated which is also inline with the previous related work (e.g., [14, 22, 48] in the N-back and [29, 30] in the Stroop test).

In order to determine the validity of the tool, we applied the method of 'concurrent validity', which is used to evaluate the measures of a novel tool against the current practice [40, 46]. Due to the frequent fluctuations in human cognition, the validity of the smartwatch-based tests was assessed immediately after the neuropsychological test sessions held at the clinic. Therefore, relevant cognitive domains of the neuropsychological tests were selected by psychiatrists to evaluate concurrent validity of the smartwatch-based cognitive tests. The tests administered during the follow-up visits included Trail Making Test (TMT) part A and B [50], Repeatable Battery for the Assessment of Neuropsychological Status (RBANS) coding and digit span [17], Wechsler Adult Intelligence Scale Letter-Number Sequencing (WAIS LNS) [33], and verbal fluency [66]. Z-transformation of the raw scores were calculated per neuropsychological test as provided below:

(1) **Composite attention and processing speed scores**: Average of z-transformations of the scores in the RBANS coding and digit span and the TMT part A.
(2) **Composite executive functions scores**: Average of z-transformations of the scores in the verbal fluency and TMT part B.
(3) **Composite working memory scores**: Z-transformations of the scores in the WAIS LNS.
(4) **Global cognitive composite scores**: Average of z-transformations of the composite scores calculated for the attention and processing speed (1) and working memory (3).

The scores of the participants with the smartwatch-based cognitive tests were averaged and their z-transformed scores were used to compare against their global composite scores.

(a) Choice reaction time test.

(b) N-back (N=2) test.

(c) Stroop color-word test.

Fig. 1. UbiCAT Cognitive Tests.

3.1 Study Procedure

Prior to the clinical feasibility study, we conducted a 1-week pilot study with two adults without any previous mental illness and one patient with BD in order to test data collection as well as refining the study procedure according to the participants' feedback. Then, recruitment of participants was commenced at Psychiatric Center Copenhagen for the *in-the-wild* clinical feasibility study. The study included both healthy controls and patients with BD, who were in partial or full remission when they came to the clinic for a follow-up visit. Psychiatrists define partial remission as scores of <=14 on the Hamilton Depression Rating Scale (HAMD) and Young Mania Rating Scale (YMRS) and full remission as scores <=7 on these rating scales. Participants were reimbursed with

Table 2. Features collected throughout the study

#	Name	Category	Type	Features
1	Choice reaction time	Cognitive	Active	Median response time, correct responses, missed stimuli
2	N-back	Cognitive	Active	Mean response time, correct responses, hit rate, false alarm rate, missed stimuli
3	Stroop	Cognitive	Active	Mean response time, correct responses, missed stimuli
4	GPS	Contextual	Passive	Latitudes and longitudes to detect indoor and outdoor environments
5	Time of the day	Contextual	Passive	Time extracted from cognitive test logs
6	Physical Activity	Behavioural	Passive	Step counts, Minutes sedentary, Minutes lightly active, Minutes fairly active, Minutes very active, Activity calories
7	Sleep	Physiological	Passive	Minutes asleep, Minutes awake, Number of awakenings, Time in bed, Minutes REM sleep, Minutes light sleep, Minutes deep sleep

an amount equivalent to 1.5 USD per cognitive test and were additionally reimbursed with an amount equivalent to 3 USD per night in case they wore the smartwatch during their sleeping time.

The study had three stages per participant as illustrated in Figure 2. Participants first underwent neuropsychological testing at the clinic. Upon finishing their follow-up visit, the study leader (PH) explained the goals and requirements to them and asked for voluntarily participation in the study. If a participant volunteered, detailed information on the study was given to them both in writing and orally, and a consent form was signed. Each participant performed the cognitive tests on the smartwatch immediately after finishing the neuropsychological tests at the clinic. Followed by that, a Fitbit Ionic smartwatch was given to the participant and s/he was instructed to perform the cognitive tests for seven days. Activity and sleep data were passively collected using the Fitbit Application Programming Interface (API).

Three alarms were set on the smartwatches; morning, afternoon, and evening. The hours set for the participant was not fixed for all of them due to different working or studying schedules. The study leader asked them to take the cognitive tests one by one in any order when they received an alarm during their earliest suitable time and emphasized on taking the tests in both indoor and outdoor environments. The stimulus of each smartwatch-based cognitive test was generated randomly. Each of the N-back sessions were either 1-back, 2-back, or 3-back such that none of the two consecutive sessions were the same. For instance, if a participant finished a 1-back task, the next time it was either a 2-back or 3-back task. While participants took cognitive tests on their smartwatch, their Global Positioning System (GPS) data was collected through the Fitbit API to detect their location. One read per second was used for the GPS. The Fitbit mobile app was installed on the participant's own smartphone since Fitbit smartwatches collect physical activity data in conjunction with their mobile app. Figure 3 illustrates a participant standing while taking a cognitive test with the smartwatch-based tool.

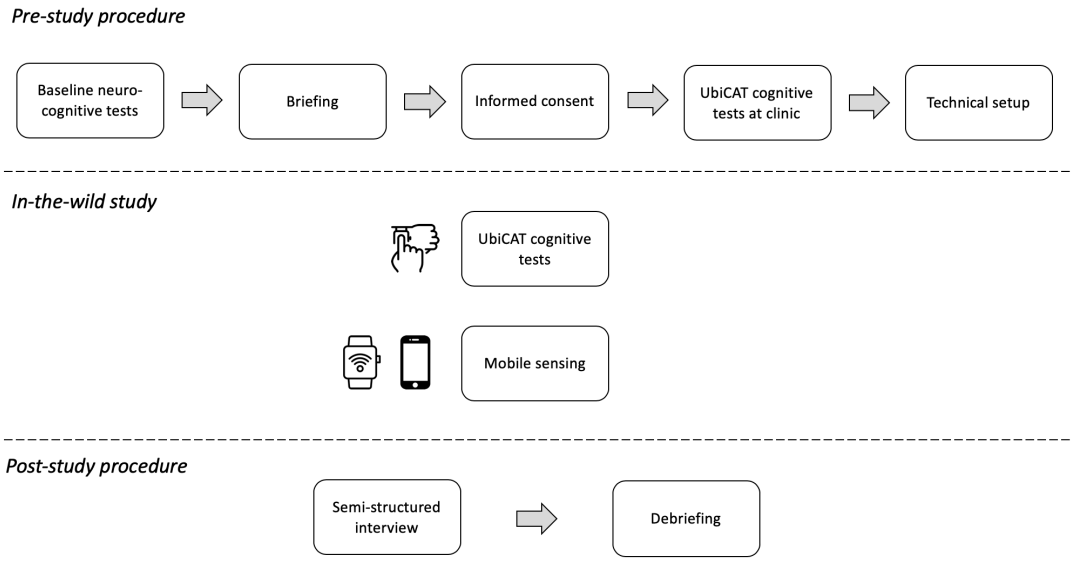


Fig. 2. Study procedure performed in three phases per participant.

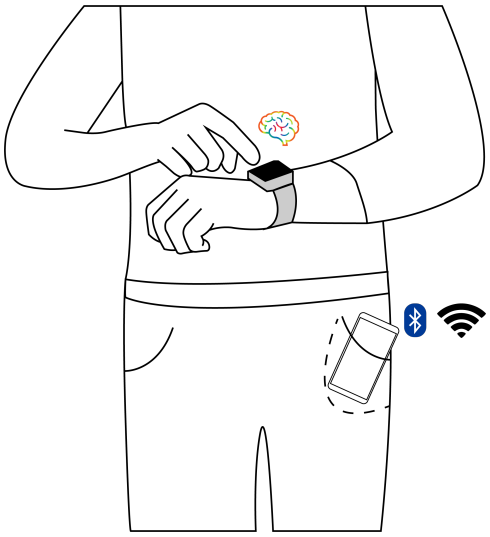


Fig. 3. A sketch of a participant taking a cognitive test of UbiCAT on the smartwatch while standing. The participant's smartphone concurrently collects GPS data.

Participants were asked to return the smartwatch upon finishing the seven-day experiment. A short semi-structured interview was held with them to explore possible issues and how comfortable or uncomfortable it was to take the tests indoor or outdoor. In addition, the contexts in which they took the tests as well as their

positions were inquired. The interview questions are presented in Appendix A. It should be noted that we were not allowed to record these interview. Thus, the study leader took notes on the participants' responses during the interviews. Finally, participants were debriefed about the further analysis of the data in the study.

3.2 Data Collection and Pre-processing

The cognitive tests results were stored locally on the smartwatch. When the smartwatch was handed back at the end of the study, the test logs were extracted and transferred to spreadsheets. Activity and sleep data were stored in the Fitbit server. We screened our dataset for missing values and outliers in the cognitive, contextual, behavioural, and physiological features. It should be noted that according to a previous work conducted on the smartwatch-based tool [24], the fastest RTs around 200 ms of the study participants were considered as an outlier. As such, the RTs to the tests stimuli below 200 ms were removed from our dataset. One participant did not wear the smartwatch during sleep and another participant only took one cognitive test daily and took off the watch during sleeping hours in most of the nights. The samples of these participant was excluded from data analysis.

The relative RTs of the participants were calculated to obtain the degree to which their alertness increased or decreased. The positive and negative values of the relative RTs shows an increase or decrease in their own alertness, respectively. Given that $MRT_{s,p}$ is the median RT of participant p in session s of the choice reaction time test, we calculated $MMRT_{s,p}$ as the mean of $MRT_{s,p}$ to obtain the $RRT_{s,p}$ (relative RT) as shown in eq. (1) (taken from [1]).

$$RRT_{s,p} = (1 - MRT_{s,p}/MMRT_{s,p}) * 100 \tag{1}$$

Two datasets were prepared for analysis of the results of this paper. A dataset was used for analyzing the impact of sleep on the next-day cognitive performance in Section 4.3. To create this dataset, first, the accuracy of the participants were averaged separately per test for each day. Then, the participant's sleep data during one night before was added to the daily cognitive observations. The second dataset was prepared by adding participants' corresponding daily activity features to the first dataset for the purpose of digital phenotyping and training supervised models as reported in Section 4.4. This combined dataset is referred as *daily cognitive and mobile data* in the rest of the paper.

3.3 Data Analysis

Several methods were used to calculate the results of this study. Pearson correlation was used for evaluating concurrent validity of the smartwatch-based tool and the association between sleep duration and cognition. T-test was performed to compare the healthy and patient groups with each other. Analysis of Variance (ANOVA) was applied to assess feasibility of the smartwatch-based tool by measuring the impact of environment on participants' cognitive performance measures. Principal Component Analysis (PCA) [70] was performed using the *factoextra* package [32] to derive the contribution of the dataset features to the principal components with an eigenvalue >1. The rest of the figures in this paper were created using the *ggplot2* [69] and *ggstatplot* [49] packages in R studio.

Supervised predictive models were used to classify healthy controls and patients with BD. Random Forest (RF) [39], Extreme Gradient Boosting (XGBoost) [12], Support Vector Machines (SVM) (radial kernel) [57], and KNN [3] were used to build the predictive models in *caret* package [36]. The label assigned to the observations was their mental health diagnosis (healthy or patient) and the positive class for the predictive models was the patient's class. Each model was trained and tested using the *leave-one-subject-out* cross validation method such that daily observations of a participant was put only in the test set and the rest of the observations were put in the train set. We applied 5-fold cross validation to train each model. The performance evaluation metrics are accuracy, precision, recall, and F1-measure. Accuracy calculates the number of correctly-classified samples. Precision quantifies how many observations were classified as 'patient' which were actually a patient. Recall

shows the extent to which a classifier can identify observations in the 'patient' class. F1-measure calculates the harmonic mean of precision and recall [72].

4 RESULTS

We initially recruited 10 healthy controls and 8 patients with BD. Of the participants, 9 healthy controls and 6 patients completed their seven-day experiment ($N = 15$). Table 3 reports gender, age, education years of the participants, HAMD and YMRS clinical ratings, and verbal intelligence quotients for each group. In total, we collected the following number of observations per smartwatch-based cognitive test: 318 for the choice reaction time, 294 for the N-back, and 309 for the Stroop test. The corresponding mean and standard deviation of the completion times (in second) per cognitive test session are 30.02±4.54, 33.02±8.30, and 36.99±6.37. Participants used the smartwatch between 6 and 18 days (8.60±2.80). It should be noted that a participant took the tests for 18 days kept the device longer than the rest of the participants due to a problem in handing back the device.

Table 3. Characteristics of study participants reported separately for patients and controls.

Characteristic	**Measure**	**Statistics**	
		Healthy Control	**Bipolar Patient**
Gender	Female	5	5
	Male	4	1
Age	Mean±SD	34±13	32±6
Years of education	Mean±SD	16±1.8	15±2.04
HAMD	Mean±SD	1.1±1.3	5.2±3.5
YMRS	Mean±SD	0.7±2	2.3±3.2
Verbal Intelligence Quotient	Mean±SD	115±5	108±3

4.1 Validity and Feasibility of the In-the-wild Tool

The participants' smartwatch-based test results were used to investigate the concurrent validity of the tool. Table 4 shows the correlation coefficients applied between the neuropsychological tests and the UbiCAT tests per cognitive domain as well as the global cognition. We found a strong, significant correlation between the average scores obtained from the smartwatch-based and neuropsychological tests *(r=0.77)* indicating adequate concurrent validity of the smartwatch-based tool. The cognitive domains also correlated significantly *(r=0.58-0.64)*. It should be noted that correlation analysis for attention and processing speed was performed between two cognitive test scores of the smartwatch-based tool namely choice reaction time and Stroop's score of congruent stimuli. Although the analysis for the Stroop's congruent scores did not reveal a significant correlation coefficient, the choice reaction time test showed a significant p-value indicating validity of this test for measuring attention and processing speed.

Feasibility of our study instrument was examined to demonstrate the viability of smartwatches in assessing *in-the-wild* cognitive functioning considering the impact of indoor and outdoor places on participants' cognitive performance measures together with the interviews conducted with them. Of all participants, five patients and seven healthy individuals took the smartwatch-based tests both in indoor and outdoor environments according to their GPS data. Table 5 shows that their performance measures in all of the tests were statistically not different

Table 4. Pearson correlation analysis between neuropsychological tests and UbiCAT test scores.

Cognitive Function	Neuropsychological Test	UbiCAT Test	Pearson's r	p
Executive functions	Verbal fluency and TMT-B	Stroop's score to incongruent stimuli	0.58	**0.024**
Working memory	WAIS LNS	N-Back	0.63	**0.011**
Attention and processing speed	TMT-A and RBANS	Choice reaction time	0.64	**0.010**
		Stroop's score to congruent stimuli	-0.11	0.686
Global cognition	Working memory and attention	Composite scores	0.77	**<0.001**

in the indoor and outdoor places demonstrating the feasibility of the smartwatch-based tests in individuals' free-living context. Seven participants could allocate time for the post-study interview. None of them mentioned any issue except for one participant: "I felt uncomfortable when I was together with people". Their position while taking the cognitive tests were also investigated; four participants reported *sitting* as their most common position while two reported that they sometimes took the tests while *traveling* (e.g., on the bus). Two participants took the tests while *walking*. One patient and one healthy participant reported that they were motivated by using the smartwatch, stating that: "It motivated me to track my data and I checked my activities all the time", and "The watch motivated me to walk more".

Table 5. Analysis of variance applied to examine the impact of indoor and outdoor places on the tests measures

UbiCAT Test	Observations (Nr.)	Indoor/Outdoor (ratio)	Performance Measure	Mean Square	F	p
Choice-reaction time	249	6.32	Median response times	4505.44	0.45	0.501
			Num. of correct responses	0.41	0.35	0.553
N-Back	217	5.78	Mean response times	15478.79	0.34	0.560
			Num. of correct responses	9.38	0.09	0.759
Stroop	227	5.88	Mean response times	14294.25	0.30	0.583
			Num. of correct responses	2.50	0.57	0.452

4.2 Alertness Per Hour

Individuals' alertness fluctuates during the day [55] and the choice reaction time paradigm tests are typically used for measuring individuals' alertness. Hence, we used the choice reaction time test of the smartwatch-based tool to measure and compare the alertness of the patients and healthy controls. The records of one healthy participant in the choice reaction time test were removed since this participant had some issues with the touch sensitivity of the smartwatch screen. Figure 4 shows the median RTs per group and shows that the median RTs of the Healthy Control (HC) group were statically lower than the Bipolar Patient (BP) group *(t(230.30)=5.24, p<0.001)* indicating better alertness of the healthy controls.

Figure 5 represents the participants' median RTs for each hour that they performed the tests. Both groups did the choice reaction time test mostly at 9AM (patient:16, healthy:7), 1PM (patient:14, healthy:17), 2PM (patient:18, healthy:16), and 6PM (patient:20, healthy:12). Independent samples t-test along with Bonferroni adjusted p-values (p_{bonf}) revealed that patients responded significantly slower at 2PM *(t(32)=3.52, mean difference= 113.67, SE=*

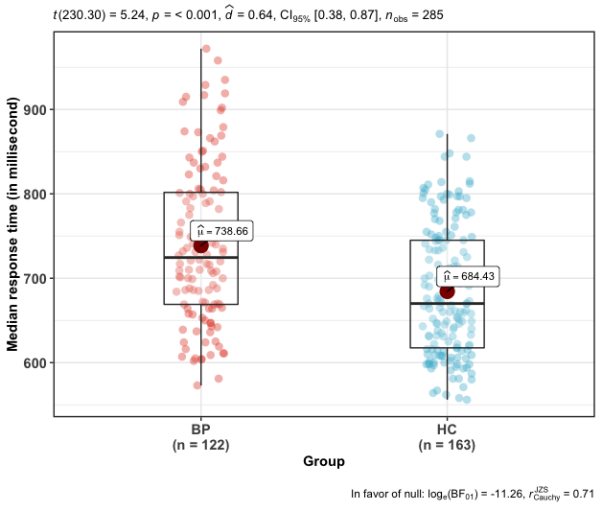


Fig. 4. Median response times of the healthy and patient groups in the choice reaction time test.

32.26, p_{bonf} =0.001) and 6PM *(t(30)=4.10, mean difference= 104.17, SE= 25.41, p_{bonf} <0.001*). The relative RTs were calculated per participant similar to the approach used in previous work [1, 34, 67]. Figure 6 shows the percentage to which the alertness of the individuals in the BP and HC groups increased or decreased compared with their own median RTs in each hour, showing an almost balanced number of positive (N=153) and negative (N=132) samples. The ratio of negative samples of the patients (48%) was higher than the healthy controls (45%) while the ratio of the positive samples of healthy controls (55%) was higher than the patients (52%). Thus, the drop in alertness was higher in the patients while the rise in alertness was higher in the controls.

Fig. 5. Hourly representation of the median response times in the choice reaction time test.

Fig. 6. Relative response times of the participants in the choice reaction time test per hour.

4.3 Correlation between Sleep Duration and Cognition

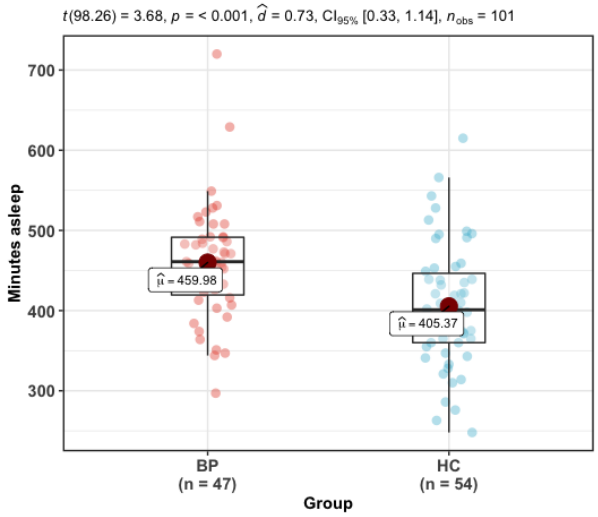


Fig. 7. Sleep duration in minutes visualized for healthy and patient groups.

Sleep duration of the participants is visualized in Figure 7 showing that patients slept more than healthy controls *(t(98.26)=3.68, p<0.001)*. Correlation analysis between sleep duration and the next-day cognitive performance measures revealed a significant coefficient in terms of the N-back hit rates as a measure of working memory performance *(r=0.26, p=0.026)*. It can be inferred that more sleeping led to higher accuracy in recalling the letters

during the N-back test sessions. The rest of the cognitive test measures did not reveal any significant correlation with sleep duration. Minutes of light and deep sleep also did not correlate with the cognitive test performance measures. The difference in the participants' N-back hit rates and sleep duration are depicted in Figure 8 per group. The ratios in both plots were calculated by taking the first value (day 1) as the basis to compare to the next days. While the difference in the sleep and hit rates of the two groups do not have the same pattern, there is a similar trend in their hit rates and sleep duration within groups.

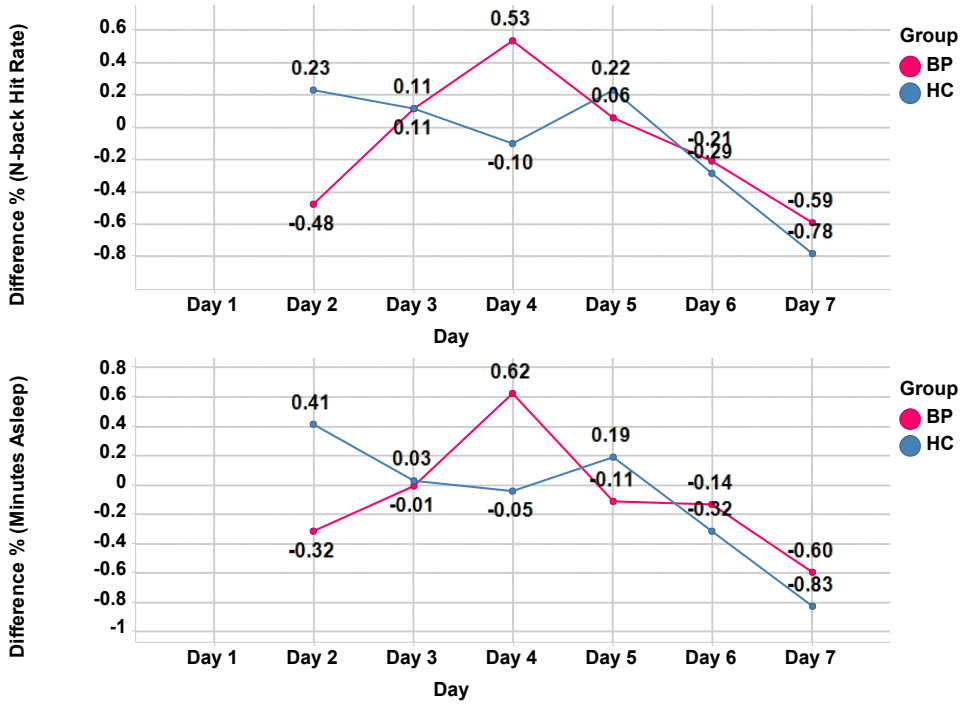


Fig. 8. Overall difference of N-back hit rates and minutes asleep per group.

4.4 Daily Cognitive and Mobile data

The daily cognitive and mobile data had $N = 81$ observations. Five incomplete cases were removed from this dataset such that $N = 76$ observations remained (bipolar:40, healthy: 36). The features of this dataset were *time in bed, sleep duration, number of awakenings, minutes awake, step counts, mean RTs during the Stroop and choice reaction time tests, and average accuracy in the N-back, Stroop, and choice reaction time tests.* Since some of the healthy participants did not wear the smartwatch during sleep, their observations were incomplete. Moreover, the patient who used the smartwatch for 18 days had more observations than the rest of the participants. Therefore, the total number of daily observations of the patients is higher than the healthy participants although the number

of healthy participants is more than the patients. It should be noted that the RTs of the N-back tests were not included since the mean RTs of the tests with various difficulty levels were not comparable .

4.4.1 Statistics and Comparison. Table 6 reports the descriptive statistics of the total daily missed counts during the cognitive test sessions and the average accuracy of the participants in each test. The missed count is the number of times that a participant did not respond to the test stimuli throughout a test session. As such, daily missed count is the sum of missed counts calculated for the test sessions that the participant completed each day. T-test revealed that the patients had more daily missed counts compared with the healthy controls *(t(72)=3.24, p<0.001)*, indicating inability of the patients with BD in responding during the time limit of the cognitive tests. The average accuracy of the healthy controls in their daily cognitive tests with the smartwatch was higher than the patients although t-test analysis did not give significant *p*-values. The RTs during the Stroop test was averaged per day for each participant and t-test showed a significantly higher RTs of the patients *(t(72)=1.93, p=0.029)*. Hence, patients were slower in selecting the correct ink color of the stimuli in the Stroop test (see Figure 1c). The participants'

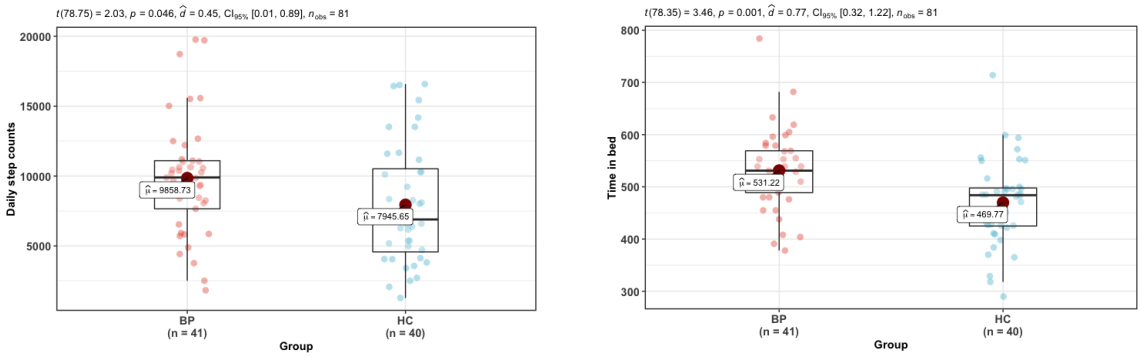
Table 6. Descriptive statistics of the daily missed counts and average accuracy in the cognitive tests

	Missed Counts		CRT accuracy (%)		N-back accuracy (%)		Stroop accuracy (%)	
	BP	HC	BP	HC	BP	HC	BP	HC
Mean	3.17	1.35	98.98	99.20	90.59	92.49	96.15	96.80
SD	3.67	1.66	1.35	1.47	5.24	4.26	2.75	2.87
Minimum	0.00	0.00	95.00	92.00	78.67	83.50	88.00	87.00
Maximum	18.00	7.00	100.00	100.00	100.00	100.00	100.00	100.00

daily step counts are represented in Figure 9a, showing that patients' mobility was significantly higher than the controls *(t(78.75)=2.03, p=0.046)*. Results also showed significant differences between sleep features of both groups. Patients spend more time in bed compared with the controls *(t(78.35)=3.46, p=0.001)* as can be observed in Figure 9b. The number of awakenings per sleeping cycle was higher in the patients as shown in Figure 10a *(t(93.75)= 3.70, p<0.0001)*. Moreover, the duration of awakenings was longer in the patients *(t(88.92)=3.20, p=0.001)* as can be seen in Figure 10b. Taken together, we found significant differences between multivariate features of the patients and healthy controls through their daily observations. These features determine the digital phenotypes of individuals' mental health diagnosis as presented in Table 7.

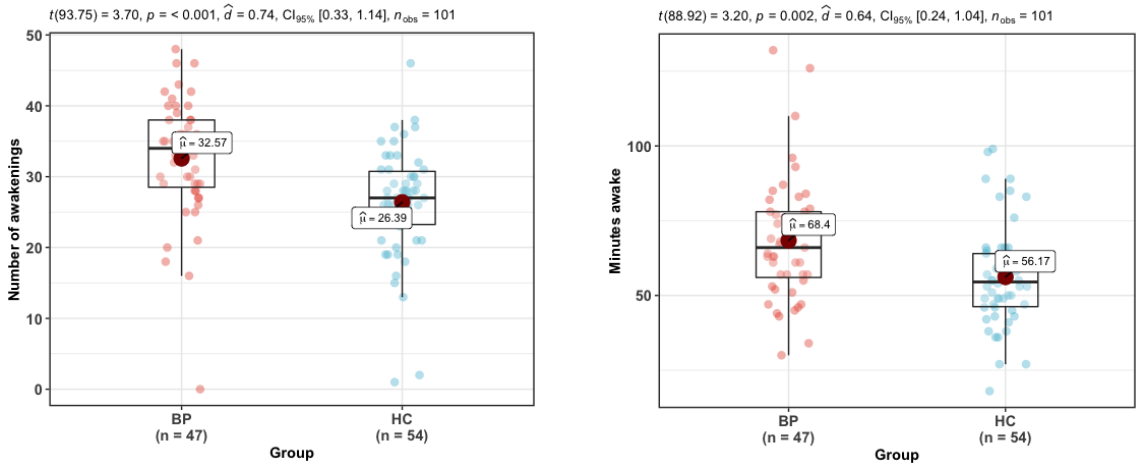
Table 7. Digital phenotypes of mental health diagnosis

Category	**Features**
Cognitive	-Response times in the choice reaction time test -Response times in the Stroop test -Missed stimuli
Physiological	-Minutes asleep -Time in bed -Number of awakenings -Minutes awake
Behavioural	-Step counts



(a) Daily step counts. (b) Time in bed.

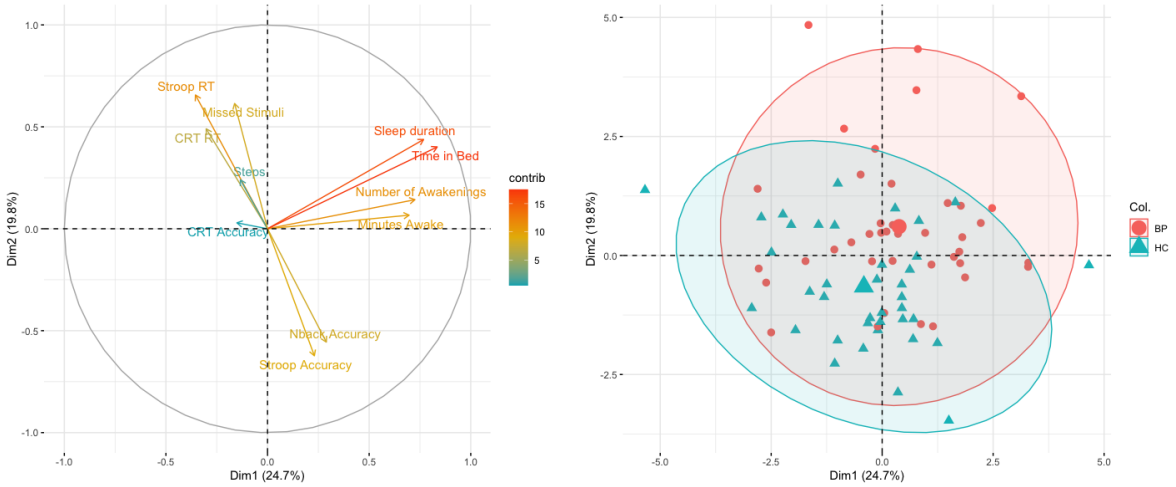
Fig. 9. Time in bed and daily step counts of each group.



(a) Number of awakenings. (b) Minutes awake.

Fig. 10. Number of awakenings and duration of awakenings of each group.

4.4.2 Feature Analysis and Predictive Models. PCA was performed to analyze the features in the dataset. The contribution of the dataset features to the Principal Component (PC) 1 and PC 2 are plotted in Figure 11a. The dataset observations were projected according to the direction of the largest variance as represented in Figure 11b. The PCs with an eigenvalue greater than 1 were PC1 (2.72), PC2 (2.18), PC3 (1.42), and PC4 (1.16). As such, we derived the features with a contribution higher than the expected average by dividing 1 by the number of features. As we have N=11 features int he daily mobile and cognitive dataset, the expected value is 9%. According to the contribution scores generated by the functions in *factorextra* package, the features that contributed around the expected value were *time in bed, sleep duration, accuracy and RT to the choice reaction time test, step counts, and RT to the Stroop*. The supervised models were then trained and tested using these features. The performance evaluation results of the supervised models are reported in Table 8. As can be seen, the average accuracies of the KNN and SVM are higher than RF and XGBoost. KNN could more accurately classify the patients' observations



(a) Dataset variables and their contributions.

(b) Projection of the dataset observations.

Fig. 11. Principal Component Analysis.

than the rest of models due to the highest average precision. KNN also gave the highest F1-measure compared to the rest of the models.

Table 8. Mean and standard deviations of the performance evaluation metrics for classification of healthy and patient groups.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Measure (%)
KNN	74.73±1.72	75.78±2.17	76.50±1.37	76.13±1.46
RF	67.64±1.18	68.25±0.95	72.00±2.09	70.06±1.30
SVM	72.10±1.71	72.92±1.80	74.83±2.91	73.84±1.75
XGBoost	68.74±3.94	72.12 ±4.22	67.35±4.49	69.60±3.68

XGBoost: Extreme Gradient Boosting; KNN: K-Nearest Neighbour; RF: Random Forest; SVM: Support Vector Machines

5 DISCUSSION

The results from this study show that smartwatch-based assessment of cognitive functioning correlated moderately to strongly *(r=0.58-0.77)* with state-of-art neuropsychological tests, thus, demonstrating adequate concurrent validity of the smartwatch-based cognitive tests. Furthermore, the study shows the feasibility of *'in-the-wild'* assessment of cognitive functioning since participants' cognitive performance measures were statistically the same in both indoor and outdoor environments. To our knowledge, this study is the first that 1) demonstrated feasibility of assessing key cognitive functions of the patients with BD and healthy controls in their free-living context and 2) identified digital phenotypes of individuals' mental health diagnosis from a dataset including their daily cognitive, behavioural, and physiological features.

5.1 Sleep and Cognitive Functions

One of the cognitive performance measures of the N-back test was hit rates. The higher the hit rate, the better ability of the user in keeping the letters in his/her mind and thus better working memory performance. We showed a significant correlation between sleep duration and the next-day working memory performance in terms of the N-back hit rates (see Section 4.3). Such finding is in line with the results of the study by Russo et al. [52] who demonstrated a negative impact of sleep disturbance on working memory for patients with BD. Moreover, cognitive load is induced by learning tasks and it directly involves working memory ability [16] while too much cognitive load adversely affects learnability [31, 60, 71]. The significant correlation coefficient between sleep duration and working memory of the participants may inform the Ubicomp community about adjusting learnability of the smartphone-based tasks according to the user's sleep duration the night before.

Similar to Dingler et al. [18], we did not find a significant impact of sleep on alertness although we used objective sleep measures. However, another study [1] did find a significant impact of sleep variation on individuals' alertness while their study aimed to systematically measure circadian rhythm during 40 days, their sample size was larger ($N = 20$) and a different target population (young college students) took part in their study.

5.2 Digital Phenotypes of Mental Health Diagnosis

The statistical comparison between the cognitive performance measures of the patients with BD and healthy controls determined the processing speed of the participants during the choice reaction time as one of the cognitive, digital phenotypes of human mental health. Patients with BD responded more slowly compared with healthy controls during the choice reaction time test although the patients were in remission. This is in line with prior research showing that cognitive impairment remains in patients with BD despite being in remission [9, 61]. We also showed that 1) patients' responses were statistically slower than the controls at 2PM and 6PM and 2) there were more drop and less rise in patients' alertness compared with the healthy controls (see Figure 6). The meaningful difference between alertness level of the healthy and patient groups together with hourly-basis analysis of their alertness may inform the Ubicomp community to consider individuals' mental health diagnosis as a potential feature for managing attention-demanding tasks per hour. Prior work has established the threshold for an impairment in alertness using smartphones as 500 ms in [7]. According to Figure 4, the results from this study seems to imply that the threshold of median RTs for distinguishing patients and healthy controls is 800 ms when using a smartwatch-based tool. The difference between the modality of smartphones and smartwatches justifies the difference between individuals' RTs. Participants' processing speed during the Stroop test is another cognitive, digital phenotype due to the statistically-significant difference found between healthy and patient groups. Taken together, moment-by-moment assessment of cognitive functioning using the smartwatch-based tool can uncover an onset of a mental disorder through slow RTs in the choice reaction time and Stroop tests.

The daily average cognitive test accuracy of the healthy group was slightly higher than the patients although the t-test did not reveal any significant p-value (see Table 6). Daily missed counts in the cognitive tests sessions were calculated per participant. Patients on average missed more stimuli than the healthy participants. The higher number of daily missed count indicates an inability to respond within a time limit. While none of the previous related work investigated individuals' missed stimuli particularly patients with BD, we showed that this objective measure has a great potential in distinguishing patients with BD and healthy controls from each other. Therefore, missed stimuli is another cognitive, digital phenotype of mental health diagnosis concerning cognitive ability of the individuals.

Significant differences were found between sleep data of the patients and healthy controls such that patients with BD slept more, stayed longer in bed, were awakened more often and for a longer duration per sleeping cycle. Hence, these features are considered as the physiological, digital phenotypes of the patients with BD and the healthy controls. The higher sleep duration of the patients is inline with the findings of a study conducted with

controls and patients with BD using actigraphy [43]. Furthermore, the higher awakening duration and frequency in the patients with BD may justify the longer sleep duration of the patients compared with the controls. Figure 9b also showed that the patients stayed more time in bed compared with the healthy controls. Sleep disturbance of the patients with BD was previously demonstrated in [51] and more time in bed and higher awakenings may inform about sleep disturbance of the patients.

Mobility of the patients in terms of their step counts was also higher than the controls that determines step counts as the behavioural, digital phenotypes of human mental health diagnosis. While we observed higher mobility in patients compared with the controls, it did not seem to be explained by more subsyndromal mania symptoms, which was examined using their YMRS ratings. Such finding should be interpreted with caution given the small sample size of our study. Moreover, some of the participants mentioned that they were motivated by the Fitbit smartwatches to walk more according to the interviews conducted with them in Section 4.1.

5.3 Implications of the Supervised Models

Various models were trained and tested on the daily cognitive and mobile dataset. We showed the potential of supervised learning methods in classifying patients with BD and healthy controls using a dataset including daily observations of cognitive functioning, sleep, and activity features captured by the smartwatches. KNN is a simple and non-parametric classifier without any assumption about the underlying data [27] while it is susceptible to the noise in the data. However, the impact of noise is less significant when using simple classifiers like KNN rather than more complicated methods like RF and SVM [73]. SVM is suitable when classes are separable but does not perform well in case of overlapped classes. This method performed comparably better than the XGBoost and RF techniques. XGBoost uses an ensemble of decision trees and is an enhanced algorithm of gradient boosting [12]. However, it did not perform as expected. RF is also a tree-based method that similarly did not perform well. Above all, the KNN model gave the highest average accuracy of nearly 74% to show the potential of supervised modelling in classifying patients with BD and healthy controls using multivariate active and passive data.

5.4 Perspectives

The results of this study justify the feasibility of utilising objective sleep and activity data in cognition-aware systems to help in managing the demand on users' working memory the following day by, for example, reducing task load in case of poor sleep quality. Furthermore, our findings paves the way for building clinical decision support systems using wearable and mobile sensor data for timely detection of the mental disorders in particular BD. Continuing this line of research will also enable researchers to include mobile and wearable sensor data in their studies to identify other digital phenotypes of cognition. One such feature is ambient noise, which can be collected via the smartphone's microphone. Possible associations between cognitive performance measures and moment-by-moment stress ratings may also provide new knowledge. Moreover, phone interaction features such as gestures and accelerometer measures can be integrated with the features we collected in our study toward a more comprehensive identification of digital phenotypes of mental health. Future studies may collect variance in the participants' RTs to investigate significant differences between responses of the patients and healthy groups.

5.5 Limitation

This study have some limitations. First, the final sample size (number of participants) was smaller than planned due to the COVID-19 outbreak, which required recruitment to be stopped before the study ended. Nevertheless, our findings still provide a statistically-significant analysis on cognitive performance measures and daily mobile data in particular related to the concurrent validity and feasibility of the smartwatch-based tool and the identification of digital phenotypes of mental health. Second, the fluctuations in the RTs over the course of the day depend on several factors including the chronotype of the individuals (for example, morningness vs, eveningness), which

we did not control for this study. Third, the golden standard for sleep assessment in clinical studies are typically self-reported sleep assessments [11], activity patterns of wrist-worn actigraphy [4], or polysomnography for sleep monitoring [35, 38, 41]. Even though acceptable performance of the Fitbit device in collecting sleep data have been demonstrated [25, 26], such consumer sleep tracking devices are not medical devices and might not be as accurate.

6 CONCLUSION

This study showed that a smartwatch-based cognitive assessment tool is a valid instrument for measuring attention, working memory, and executive functions. Moreover, this tool is feasible for frequent assessment of the key cognitive functions *'in-the-wild'*, i.e. in both indoor and outdoor environments, as well as when users are taking different positions such as *sitting, standing, and walking*. We also showed the potential of wearable computing technology in collecting daily multivariate, active and passive data for digital phenotyping and supervised modelling. Patients with BD responded slower in the attention test compared to healthy controls, indicating lower alertness level of the patients. Sleep duration correlated positively with the next-day working memory performance, which may help inform the design of cognition-aware computing system when cognitive load is managed in accordance with sleep duration. Digital phenotypes of mental health were determined by comparing cognitive performance measures and sleep and activity data of the patients with the healthy controls. We conclude that using mobile and wearable technology for ambulatory collection of individuals' physiological, behavioural and cognitive features provides the basis for assisting clinicians in continuously monitoring patients' symptoms for early diagnosis and treatments of mental disorders.

ACKNOWLEDGMENTS

This research is funded by the European Union's Horizon Sklodowska-Curie grant agreement as part of the Technology-Enabled Mental Health for Young People (TEAM) Initial Training Network (No. 722561) and the Copenhagen Center for Health Technology (CACHET). We wish to thank Malte Kampmark for graphical assistance and our study participants.

REFERENCES

[1] Saeed Abdullah, Elizabeth L Murnane, Mark Matthews, Matthew Kay, Julie A Kientz, Geri Gay, and Tanzeem Choudhury. 2016. Cognitive rhythms: unobtrusive and continuous sensing of alertness using a mobile phone. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 178–189.

[2] Michèle Allard, Mathilde Husky, Gwénaëlle Catheline, Amandine Pelletier, Bixente Dilharreguy, Hélène Amieva, Karine Pérès, Alexandra Foubert-Samier, Jean-François Dartigues, and Joel Swendsen. 2014. Mobile technologies in the early detection of cognitive decline. *PLoS One* 9, 12 (2014), e112197.

[3] Naomi S Altman. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician* 46, 3 (1992), 175–185.

[4] Sonia Ancoli-Israel, Jennifer L Martin, Terri Blackwell, Luis Buenaver, Lianqi Liu, Lisa J Meltzer, Avi Sadeh, Adam P Spira, and Daniel J Taylor. 2015. The SBSM guide to actigraphy monitoring: clinical and research applications. *Behavioral sleep medicine* 13, sup1 (2015), S4–S38.

[5] S Ancoli-Israel and Th Roth. 1999. Characteristics of insomnia in the United States: results of the 1991 National Sleep Foundation Survey. I. *Sleep* 22 (1999), S347–53.

[6] Noa Avni, Isaac Avni, Erez Barenboim, Bella Azaria, David Zadok, REUVEN KOHEN-RAZ, and Yair Morad. 2006. Brief posturographic test as an indicator of fatigue. *Psychiatry and clinical neurosciences* 60, 3 (2006), 340–346.

[7] Mathias Basner, Daniel Mollicone, and David F Dinges. 2011. Validity and sensitivity of a brief psychomotor vigilance test (PVT-B) to total and partial sleep deprivation. *Acta astronautica* 69, 11-12 (2011), 949–959.

[8] S Bihari, A Venkatapathy, S Prakash, E Everest, D McEvoy R, and A Bersten. 2020. ICU shift related effects on sleep, fatigue and alertness levels. *Occupational medicine* 70, 2 (2020), 107–112.

[9] E Bora and A Özerdem. 2017. Meta-analysis of longitudinal studies of cognition in bipolar disorder: comparison with healthy controls and schizophrenia. *Psychological medicine* 47, 16 (2017), 2753–2766.

[10] Robert M Brouillette, Heather Foil, Stephanie Fontenot, Anthony Correro, Ray Allen, Corby K Martin, Annadora J Bruce-Keller, and Jeffrey N Keller. 2013. Feasibility, reliability, and validity of a smartphone based application for the assessment of cognitive function in the elderly. *PloS one* 8, 6 (2013), e65925.

[11] Colleen E Carney, Daniel J Buysse, Sonia Ancoli-Israel, Jack D Edinger, Andrew D Krystal, Kenneth L Lichstein, and Charles M Morin. 2012. The consensus sleep diary: standardizing prospective sleep self-monitoring. *Sleep* 35, 2 (2012), 287–302.

[12] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.

[13] Chul-Hyun Cho, Taek Lee, Min-Gwan Kim, Hoh Peter In, Leen Kim, and Heon-Jeong Lee. 2019. Mood prediction of patients with mood disorders by machine learning using passive digital phenotypes based on the circadian rhythm: prospective observational cohort study. *Journal of medical Internet research* 21, 4 (2019), e11029.

[14] Francesca Cormack, Maggie McCue, Nick Taptiklis, Caroline Skirrow, Emilie Glazer, Elli Panagopoulos, Tempest A van Schaik, Ben Fehnert, James King, and Jennifer H Barnett. 2019. Wearable Technology for High-Frequency Cognitive and Mood Assessment in Major Depressive Disorder: Longitudinal Observational Study. *JMIR Mental Health* 6, 11 (2019), e12814.

[15] NEM Daniëls, SL Bartels, SJW Verhagen, RJM Van Knippenberg, ME De Vugt, and Ph AEG Delespaul. 2020. Digital assessment of working memory and processing speed in everyday life: Feasibility, validation, and lessons-learned. *Internet Interventions* 19 (2020), 100300.

[16] Robin Deegan. 2013. Mobile Learning Application Interfaces: First Steps to a Cognitive Load Aware System. *International Association for Development of the Information Society* (2013).

[17] Faith Dickerson, John J Boronow, Cassie Stallings, Andrea E Origoni, Sara K Cole, and Robert H Yolken. 2004. Cognitive functioning in schizophrenia and bipolar disorder: comparison of performance on the Repeatable Battery for the Assessment of Neuropsychological Status. *Psychiatry research* 129, 1 (2004), 45–53.

[18] Tilman Dingler, Albrecht Schmidt, and Tonja Machulla. 2017. Building cognition-aware systems: A mobile toolkit for extracting time-of-day fluctuations of cognitive performance. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 47.

[19] Maria Faurholt-Jepsen, Jonas Busk, Helga Þórarinsdóttir, Mads Frost, Jakob Eyvind Bardram, Maj Vinberg, and Lars Vedel Kessing. 2019. Objective smartphone data as a potential diagnostic marker of bipolar disorder. *Australian & New Zealand Journal of Psychiatry* 53, 2 (2019), 119–128.

[20] Maria Faurholt-Jepsen, Mads Frost, Maj Vinberg, Ellen Margrethe Christensen, Jakob E Bardram, and Lars Vedel Kessing. 2014. Smartphone data as objective measures of bipolar disorder symptoms. *Psychiatry research* 217, 1-2 (2014), 124–127.

[21] Marcos G Frank. 2006. The mystery of sleep function: current perspectives and future directions. *Reviews in the Neurosciences* 17, 4 (2006), 375–392.

[22] Lars Frings, Kathrin Wagner, Thomas Maiwald, Astrid Carius, Anika Schinkel, Christiane Lehmann, and Andreas Schulze-Bonhage. 2008. Early detection of behavioral side effects of antiepileptic treatment using handheld computers. *Epilepsy & Behavior* 13, 2 (2008), 402–406.

[23] Pegah Hafiz and Jakob E Bardram. 2019. Design and formative evaluation of cognitive assessment apps for wearable technologies. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. 1162–1165.

[24] Pegah Hafiz and Jakob Eyvind Bardram. 2020. The Ubiquitous Cognitive Assessment Tool for Smartwatches: Design, Implementation, and Evaluation Study. *JMIR mHealth and uHealth* 8, 6 (2020), e17506.

[25] Shahab Haghayegh, Sepideh Khoshnevis, Michael H Smolensky, Kenneth R Diller, and Richard J Castriotta. 2019. Accuracy of Wristband Fitbit Models in Assessing Sleep: Systematic Review and Meta-Analysis. *Journal of medical Internet research* 21, 11 (2019), e16273.

[26] Shahab Haghayegh, Sepideh Khoshnevis, Michael H Smolensky, Kenneth R Diller, and Richard J Castriotta. 2020. Performance assessment of new-generation Fitbit technology in deriving sleep parameters and stages. *Chronobiology International* 37, 1 (2020), 47–59.

[27] D Hand, H Mannila, and P Smyth. 2001. Principles of Data Mining". The MIT Press. In *A comprehensive, highlytechnical look at the math and science behind extracting useful information from large databases*. Vol. 546.

[28] John E Harrison, Harry Barry, Bernhard T Baune, Michael W Best, Christopher R Bowie, Danielle S Cha, Larry Culpepper, Philippe Fossati, Tracy L Greer, Catherine Harmer, et al. 2018. Stability, reliability, and validity of the THINC-it screening tool for cognitive impairment in depression: A psychometric exploration in healthy volunteers. *International journal of methods in psychiatric research* 27, 3 (2018), e1736.

[29] Shan Hung, Min-Shan Li, Yen-Lin Chen, Jung-Hsien Chiang, Ying-Yeh Chen, and Galen Chin-Lun Hung. 2016. Smartphone-based ecological momentary assessment for Chinese patients with depression: An exploratory study in Taiwan. *Asian journal of psychiatry* 23 (2016), 131–136.

[30] Susan Jongstra, Liselotte Willemijn Wijsman, Ricardo Cachucho, Marieke Peternella Hoevenaar-Blom, Simon Pieter Mooijaart, and Edo Richard. 2017. Cognitive testing in people at increased risk of dementia using a smartphone app: the iVitality proof-of-principle study. *JMIR mHealth and uHealth* 5, 5 (2017), e68.

[31] Slava Kalyuga. 2011. Cognitive load theory: How many types of load does it really need? *Educational Psychology Review* 23, 1 (2011), 1–19.
[32] Alboukadel Kassambara and Fabian Mundt. 2020. *factoextra: Extract and Visualize the Results of Multivariate Data Analyses.* https://CRAN.R-project.org/package=factoextra R package version 1.0.7.
[33] Alan S Kaufman and Elizabeth O Lichtenberger. 2005. *Assessing adolescent and adult intelligence.* John Wiley & Sons.
[34] Matthew Kay, Kyle Rector, Sunny Consolvo, Ben Greenstein, Jacob O Wobbrock, Nathaniel F Watson, and Julie A Kientz. 2013. PVT-touch: adapting a reaction time test for touchscreen devices. In *2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops.* IEEE, 248–251.
[35] Bhanu Prakash Kolla, Subir Mansukhani, and Meghna P Mansukhani. 2016. Consumer sleep tracking devices: a review of mechanisms, validity and utility. *Expert review of medical devices* 13, 5 (2016), 497–506.
[36] Max Kuhn. 2020. *caret: Classification and Regression Training.* https://CRAN.R-project.org/package=caret R package version 6.0-86.
[37] Reed Larson and Mihaly Csikszentmihalyi. 2014. The experience sampling method. In *Flow and the foundations of positive psychology.* Springer, 21–34.
[38] Zilu Liang and Mario Alberto Chapa Martell. 2018. Validity of consumer activity wristbands and wearable EEG for measuring overall sleep parameters and sleep structure in free-living conditions. *Journal of Healthcare Informatics Research* 2, 1-2 (2018), 152–178.
[39] Andy Liaw, Matthew Wiener, et al. 2002. Classification and regression by randomForest. *R news* 2, 3 (2002), 18–22.
[40] Wei-Ling Lin and Grace Yao. 2014. *Concurrent Validity.* Springer Netherlands, Dordrecht, 1184–1185. https://doi.org/10.1007/978-94-007-0753-5_516
[41] Janna Mantua, Nickolas Gravel, and Rebecca Spencer. 2016. Reliability of sleep measures from four personal health monitoring devices compared to research-based actigraphy and polysomnography. *Sensors* 16, 5 (2016), 646.
[42] Emmanuel Mignot. 2008. Why we sleep: the temporal organization of recovery. *PLoS biology* 6, 4 (2008).
[43] Audrey Millar, Colin A Espie, and Jan Scott. 2004. The sleep of remitted bipolar outpatients: a controlled naturalistic study using actigraphy. *Journal of affective disorders* 80, 2-3 (2004), 145–153.
[44] Wendy G Mitchell, Yvonne Zhou, John M Chavez, and Bianca L Guzman. 1992. Reaction time, attention, and impulsivity in epilepsy. *Pediatric neurology* 8, 1 (1992), 19–24.
[45] Raeanne C Moore, Joel Swendsen, and Colin A Depp. 2017. Applications for self-administered mobile cognitive assessments in clinical research: A systematic review. *International journal of methods in psychiatric research* 26, 4 (2017), e1562.
[46] Kevin R Murphy and Charles O Davidshofer. 1988. Psychological testing. *Principles, and Applications, Englewood Cliffs* (1988).
[47] Jukka-Pekka Onnela and Scott L Rauch. 2016. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *Neuropsychopharmacology* 41, 7 (2016), 1691–1696.
[48] Reshmi Pal, John Mendelson, Odile Clavier, Mathew J Baggott, Jeremy Coyle, and Gantt P Galloway. 2016. Development and testing of a smartphone-based cognitive/neuropsychological evaluation system for substance abusers. *Journal of psychoactive drugs* 48, 4 (2016), 288–294.
[49] Indrajeet Patil. 2018. *ggstatsplot: "ggplot2" Based Plots with Statistical Details.* https://doi.org/10.5281/zenodo.2074621
[50] Ralph M Reitan. 1958. Validity of the Trail Making Test as an indicator of organic brain damage. *Perceptual and motor skills* 8, 3 (1958), 271–276.
[51] Paulo Marcos Brasil Rocha, Fernando Silva Neves, and Humberto Corrêa. 2013. Significant sleep disturbances in euthymic bipolar patients. *Comprehensive psychiatry* 54, 7 (2013), 1003–1008.
[52] Manuela Russo, Katie Mahon, Megan Shanahan, Elizabeth Ramjas, Carly Solon, Shaun M Purcell, and Katherine E Burdick. 2015. The relationship between sleep quality and neurocognition in bipolar disorder. *Journal of affective disorders* 187 (2015), 156–162.
[53] Sohrab Saeb, Emily G Lattie, Stephen M Schueller, Konrad P Kording, and David C Mohr. 2016. The relationship between mobile phone location sensor data and depressive symptom severity. *PeerJ* 4 (2016), e2537.
[54] Sohrab Saeb, Mi Zhang, Christopher J Karr, Stephen M Schueller, Marya E Corden, Konrad P Kording, and David C Mohr. 2015. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *Journal of medical Internet research* 17, 7 (2015), e175.
[55] Christina Schmidt, Fabienne Collette, Christian Cajochen, and Philippe Peigneux. 2007. A time to think: circadian rhythms in human cognition. *Cognitive neuropsychology* 24, 7 (2007), 755–789.
[56] Martin J Sliwinski, Jacqueline A Mogle, Jinshil Hyun, Elizabeth Munoz, Joshua M Smyth, and Richard B Lipton. 2018. Reliability and validity of ambulatory cognitive assessments. *Assessment* 25, 1 (2018), 14–30.
[57] Ingo Steinwart and Andreas Christmann. 2008. *Support vector machines.* Springer Science & Business Media.
[58] Gijsbert Stoet. 2017. PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology* 44, 1 (2017), 24–31.
[59] Arthur A Stone and Saul Shiffman. 1994. Ecological momentary assessment (EMA) in behavorial medicine. *Annals of Behavioral Medicine* (1994).
[60] John Sweller. 1994. Cognitive load theory, learning difficulty, and instructional design. *Learning and instruction* 4, 4 (1994), 295–312.

[61] Alejandro Szmulewicz, Marina P Valerio, and Diego J Martino. 2019. Longitudinal analysis of cognitive performances in recent-onset and late-life Bipolar Disorder: A systematic review and meta-analysis. *Bipolar disorders* (2019).
[62] Zoë Tieges, Antaine Stíobhairt, Katie Scott, Klaudia Suchorab, Alexander Weir, Stuart Parks, Susan Shenkin, and Alasdair MacLullich. 2015. Development of a smartphone application for the objective detection of attentional deficits in delirium. *International psychogeriatrics* 27, 8 (2015), 1251–1262.
[63] Corrie Timmers, Anne Maeghs, Michiel Vestjens, Charlie Bonnemayer, Huub Hamers, and Arjan Blokland. 2014. Ambulant cognitive assessment using a smartphone. *Applied Neuropsychology: Adult* 21, 2 (2014), 136–142.
[64] John Torous, JP Onnela, and Matcheri Keshavan. 2017. New dimensions and new tools to realize the potential of RDoC: digital phenotyping via smartphones and connected devices. *Translational psychiatry* 7, 3 (2017), e1053–e1053.
[65] John Torous, Patrick Staples, Ian Barnett, Luis R Sandoval, Matcheri Keshavan, and Jukka-Pekka Onnela. 2018. Characterizing the clinical relevance of digital phenotyping data quality with applications to a cohort with schizophrenia. *NPJ digital medicine* 1, 1 (2018), 1–9.
[66] Eirini Tsitsipa and Konstantinos N Fountoulakis. 2015. The neurocognitive functioning in bipolar disorder: a systematic review of data. *Annals of general psychiatry* 14, 1 (2015), 42.
[67] Céline Vetter, Myriam Juda, and Till Roenneberg. 2012. The influence of internal time, time awake, and sleep duration on cognitive performance in shiftworkers. *Chronobiology international* 29, 8 (2012), 1127–1138.
[68] Robert West, Kelly J Murphy, Maria L Armilio, Fergus IM Craik, and Donald T Stuss. 2002. Effects of time of day on age differences in working memory. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 57, 1 (2002), P3–P10.
[69] Hadley Wickham. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org
[70] Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems* 2, 1-3 (1987), 37–52.
[71] Bin Xie and Gavriel Salvendy. 2000. Review and reappraisal of modelling and predicting mental workload in single-and multi-task environments. *Work & stress* 14, 1 (2000), 74–99.
[72] Yiming Yang and Xin Liu. 1999. A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. 42–49.
[73] Yan Zhang and Xindong Wu. 2010. Integrating induction and deduction for noisy data mining. *Information Sciences* 180, 14 (2010), 2663–2673.

A QUESTIONS ASKED DURING THE POST-STUDY INTERVIEW

(1) How was your experience with the cognitive tests on the smartwatch?
(2) What positions did you take while doing a test on the smartwatch (e.g. sitting, walking, standing)?
(3) Did you take the cognitive tests outside (e.g. on the street)? If yes, describe it and explain how you perceived the situation.