EAI Endorsed Transactions
on Pervasive Health and Technology
Research Article EAI.EU

Not All Errors Are Created Equal: Influence of User Characteristics on Measurement Errors of Consumer Wearable Devices for Sleep Tracking

Zilu Liang[1,2,*] and Mario Alberto Chapa Martell[3]

[1]Kyoto University of Advanced Science, 18 Yamanouchi Gotanda-cho, Ukyo-ku, Kyoto 615-8577, Japan
[2]The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
[3]CAC Corporation, 24-1 Hakozakicho, Chuo-ku, Tokyo 103-0015, Japan

Abstract

Consumer sleep tracking devices are known to be inaccurate, but there is a lack of understanding of how user characteristics may affect the accuracy of these devices. This study aims to examine the effect of age, gender, subjective sleep quality, sleep hygiene and sleep structure on the accuracy of two consumer sleep trackers, i.e. Fitbit Charge 2 and Neuroon. Sleep data were collected from 27 healthy participants using consumer devices and a medical device concurrently. Analysis found that age, sleep hygiene and sleep structure were significantly associated to the accuracy of consumer sleep trackers, whereas no association was found on gender and subjective sleep quality. Both consumer devices had improved accuracy on total sleep time and sleep efficiency for participants who had longer, deeper and less interrupted sleep. Our findings suggest that consumer devices may not be suited for young adults and for people with short and fragmented sleep.

Keywords: wearable, sleep, validation, error analysis, Fitbit, EEG, personal informatics.

Received on 31 April 2018, accepted on 25 June 2018, published on 30 July 2018

Copyright © 2019 Zilu Liang *et al.*, licensed to EAI. This is an open access article distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/3.0/), which permits unlimited use, distribution and reproduction in any medium so long as the original work is properly cited.

doi: 10.4108/eai.24-7-2018.159404

*Corresponding author. Email: z.liang@cnl.t.u-tokyo.ac.jp

1. Introduction

The proliferation of consumer sleep tracking technologies has significantly raised people's awareness of sleep health [1-3]. Compared to polysomnography (PSG), consumer wearable devices provide low-cost and unobtrusive alternatives for individuals to monitor sleep on daily basis without the need for constant technical support. These devices also offer great opportunities for researchers to conduct large-scale longitudinal studies at reasonable cost.

Despite of the multi-fold advantages of consumer sleep trackers, the quality of data measured by these devices has been a major concern for end users, researchers, and clinicians [2, 4, 5]. Low quality data may mislead end users to make wrong decisions. In addition, data quality is of top priority for researchers who intend to use these devices in scientific studies. Therefore, it is important to understand whether, when and with whom these devices could produce accurate measurements. In response to this need, researchers have studied the validity and reliability of popular wearable sleep trackers both in sleep laboratories [6-10] and under free living conditions [11, 25]. The most recent studies revealed that the latest models of consumer sleep trackers could accurately measure total sleep duration and sleep efficiency in health people, whereas detecting sleep stages remained to be the main challenge [6, 11, 25].

Previous studies on clinical actigraphy revealed that user characteristics may affect device accuracy [12, 13]. However, little work has yet looked at whether and how user characteristics associate to the accuracy of consumer sleep tracking devices. This study aims to fill in the knowledge

gap by investigating the influence of a set of user-specific factors on the measurement errors by the latest consumer sleep trackers. Our hypothesis was that the accuracy of consumer sleep-tracking devices may be associated to users' age, gender, sleep hygiene, and sleep structure. We therefore selected a set of independent variables including demographic characteristics (age and gender), subjective sleep quality (PSQI) [14], sleep hygiene (bed time and rise time) and objective sleep structure measured by a medical device. The dependent variables were the absolute percent errors on total sleep time (TST), wake after sleep onset (WASO), and sleep efficiency (SE) measured by two of the most recent consumer sleep trackers, i.e. Fitbit Charge 2 and Neuroon. Binary logistic regression was used to determine whether the odds of measurement error increased or declined as the value of the user-specific factors changed, and Pearson correlation coefficients were computed to quantify the linear relationships between measurement errors and the user-specific factors. The findings of this study served as a reference for individual users and researchers to select the most appropriate devices and provided implications for designing more accurate sleep staging algorithms for consumer sleep trackers.

2. Related Work

Consumer wearable sleep-tracking devices offer a cost-effective and unobtrusive alternative to traditional clinical sleep monitors. The basic mechanism of these consumer devices is similar to that of medical actigraphy. These devices rely on the measurement of limb movements, which is then used to infer sleep stages [12, 15, 16]. Most of the sleep staging algorithms are proprietary and are not made available to the public. These devices are also increasingly used in scientific studies to measure sleep outcomes [17-22].

Regardless of the popularity of these consumer devices, their validity is a main concern for users who plan to use them for sleep tracking. A number of validation studies have investigated the agreement of the consumer devices to clinical sleep monitors [6, 8, 11, 23-26]. It was found that the previous models of sleep-tracking wristbands have the common problem of overestimating time asleep and underestimating time awake, mostly due to the misclassification of motionless awake as sleep. Most of these validation studies were conducted in laboratory settings and a few was conducted in free living conditions. Very recently, a few studies also attempted to enhance the accuracy of Fitbit sleep data through machine learning based approaches [27, 28].

A very interesting phenomenon of clinical actigraphy is that measurement accuracy may vary on different populations. A few studies revealed that the misclassification of actigraphy is associated to participant characteristics. For example, actigraphy in proportional integration mode (PIM) was found to be reasonably accurate for people who sleep more than 7 hours [13]. However, it overestimated TST for people with more fragmented sleep (WASO > 90 min or SE < 70%) and underestimated TST for those with sleep disordered breathing. Those with higher BMI showed better agreement of TST, and those with the lower waist circumference showed overestimation. No significant association was found between device accuracy and other participant characteristics such as age, race, medical conditions, sleep onset latency and cognition. Another study found negative correlation between the accuracy of a clinical actigraphy and the fragmentation level of sleep [12]. This study also suggested that the clinical actigraphy was less accurate for older users.

Since consumer sleep tracking devices rely on the same mechanism as clinical actigraphy, it is very likely that the accuracy of these devices is associated to user characteristics as well. This study contributes an understanding of the user-specific factors that affect the accuracy of popular consumer sleep tracking devices (i.e. Fitbit Charge 2 and Neuroon). The findings of this study provide useful hint that help end users select the most suited consumer devices.

3. Methods

3.1. Participants

We recruited participants by distributing posters around the campus of The University of Tokyo. In total 38 people applied, of whom 27 (71%) were eligible to participate in the study. The sample size is comparable with other studies in the field [29-33]. The inclusion criteria required that the participants were healthy adults (age > 18) and were free of chronic conditions, severe sleep problems or mental diseases. Neither gender nor nationality was a prerequisite for participation in this study.

Table 1. Denotations and definitions of sleep metrics

Sleep metrics	Denotation	Definition
Total sleep time (min)	*TST*	Total amount of time asleep after sleep onset.
Wake after sleep onset (min)	*WASO*	Total amount of time awake after sleep onset.
Sleep efficiency	*SE*	$\frac{TST}{SOL + TST + WASO}$[a]

[a] SOL: sleep onset latency

Participants filled in a PSQI (Pittsburgh Sleep Quality Index) [14] questionnaire to establish a baseline of their sleep quality. The PSQI is a widely used instrument for assessing subjective sleep quality averaged over the past one month, and a PSQI≥5 is indicative of poor sleep. This research was approved by the ethical committee of the University of Tokyo (Ethics ID: KE16-83). All participants provided informed consent.

3.2. Sleep Metrics

Human sleep can be characterized in multiple dimensions [15, 34-36]. In this study, we focused on the measurement errors overall three sleep metrics, i.e. total sleep time (TST), wake after sleep onset (WASO), and sleep efficiency (SE). These parameters are closely related to people's physical and mental health [37], and many people use them parameters as indicators of their sleep quality [20]. Table 1 summarizes the denotations and definitions of the sleep metrics following the literatures in sleep science [38-41].

Table 2. Denotations and data collection methods of user-specific factors

Sleep metrics	Denotation	Data collection methods
Age (year)	*Age*	Self-reported
Gender	*Sex*	Self-reported
PSQI	*PSQI*	PSQI questionnaire
Total sleep time (min)	*TST*	Sleep Scope (medical device)
Wake after sleep onset (min)	*WASO*	Sleep Scope
Sleep onset latency (min)	*SOL*	Sleep Scope
Sleep efficiency (%)	*SE*	Sleep Scope
Ratio of Sleep stage N1 (%)	*N1R*	Sleep Scope
Ratio of Sleep stage N2 (%)	*N2R*	Sleep Scope
Ratio of Deep sleep (%)	*DSR*	Sleep Scope
Ratio of REM sleep (%)	*RSR*	Sleep Scope
Average sleep cycle (min)	T_{avg}	Sleep Scope
Bedtime	t_{start}	Sleep Scope
Rise time	t_{end}	Sleep Scope

3.3. Data Collection

Devices

In this study we focused on two consumer devices, i.e., Fitbit Charge 2 and Neuroon. Both devices were the most recent models in the market at the time the study was conducted, and they were readily available for individual consumers to purchase. Fitbit Charge 2 (Fitbit Inc., San Francisco, CA, USA) is a wearable activity wristband with an embedded tri-axial accelerometer. It estimates sleep stages for each 30s period by integrating a user's movement and heart rate data. With advances in software and hardware, Fitbit Charge 2 has overcome some problems of previous models and is able to measure total sleep time and sleep efficiency with good accuracy [11]. Neuroon (Inteliclinic Co., San Francisco, CA, USA) is a wearable EEG eye mask with an embedded single channel EEG sensor. Similar to polysomnography test, Neuroon can measure a number of sleep-concurrent physiological parameters including brainwave, heart rate, eyeball movements, body temperature and body movement. These data were used to estimate overall sleep metrics and sleep stages using the company's proprietary algorithms. These two devices were selected in this study due to their popularity, affordability and the potential for accurate sleep tracking compared to other consumer devices [11].

A portable medical sleep monitor named Sleep Scope (Sleep Well Co., Osaka, Japan) was used to obtain accurate measurements on sleep metrics. Sleep Scope is a clinical-grade single-channel EEG (Japanese Medical Device Certification 225ADBZX00020000), which was validated against PSG (agreement = 86.9%, average Cohen's Kappa value = 0.753) [42, 43]. Sleep Scope was chosen over PSG because it enabled data collection in participants' homes rather than in a sleep laboratory, thus minimalizing the possible disruption of sleep by unfamiliar environment.

Measurement Protocol

We held a briefing with each participant individually before the start of the experiment. In this meeting, we installed the Fitbit and Neuroon applications on participants' smartphones and gave them the following items for the self-tracking experiment: a Fitbit Charge 2, a Neuroon eye mask, a medical device Sleep Scope, and necessary accessories. The participants then tracked their sleep for three consecutive nights in their homes. All three devices were used concurrently to ensure that measurement differences were derived under the same conditions. Participants received a $54 shopping card as appreciation when they complete the experiment.

3.4. Data Analysis

Data Pre-Processing

We retrieved and analysed the sleep data of one night for each participant and obtained a dataset with 27 observations. Following the common practice in sleep science, we analysed the second night for each participant to remove "the first night effect" [44-46]. If the data of the second night was not valid, then the data of the third night was analysed. The data of the first night was only selected when neither the second night nor the third night was valid.

Fitbit sleep data was retrieved through Fitbit public API using a web application that we developed in our previous study [20]. Neuroon sleep data was manually retrieved from the dashboard as no public API was available. The data of the medical device was analysed by the Sleep Well Company using proprietary automatic scoring algorithms,

followed by visual inspection epoch-by-epoch by specialists according to established standards [47] and corrections were added if needed.

User-specific Factors and Error Measures

In this study we investigated the influence of 14 user-specific factors, including age, gender, subjective sleep quality, sleep hygiene and sleep patterns. Table 2 summarizes the denotations of the user-specific factors and how they were collected during the study.

Taking the data derived by the medical device as the ground truth, we calculated absolute percent error $|e_i^{sleep}|\ (sleep \in \{TST, WASO, SE\})$ as indicators of measurement errors by the consumer devices using Equation (1) [48, 49], where $y_i^{sleep_C}$ epresents the sleep data of participant i measured by a consumer device, and $y_i^{sleep_M}$ represents the corresponding ground truth derived by the medical device.

$$|e_i^{sleep}| = \frac{|y_i^{sleep_C} - y_i^{sleep_M}|}{y_i^M} \times 100 \quad (1)$$

Statistical Analysis

In line with previous studies [50, 51], we defined the acceptable error range as $|e_i^{sleep}| \leq 5\%$ since this approximates a widely acceptable standard for statistical significance in health science research [52]. Based on this criterion, we divided the dataset into two subsets according to the magnitude of the measurement errors as is shown in Equation (2) ~ (5). Each of the observations in the dataset was classified as either good agreement or poor agreement.

Good agreement:

$$S_{e^0}^{sleep} = \{s_i \,|\, |e_i^{sleep}| \leq 5\%,\ i = 1,2,\dots,N_{e^0}\}, \quad (2)$$

Poor agreement:

$$S_{e^1}^{sleep} = \{s_j \,|\, |e_j^{sleep}| > 5\%,\ j = 1,2,\dots,N_{e^1}\}, \quad (3)$$

$$s_i \in \{y_i^{TST}, y_i^{WASO}, y_i^{SE}\} \quad (4)$$

$$N_{e^0}^{sleep} + N_{e^1}^{sleep} = N^{sleep}, \quad (5)$$

where s_i represents the i-th observation in each subset, $|e_i^{sleep}|$ is the absolute percent error of either *TST*, *WASO*, or *SE* in observation s_i, $N_{e^0}^{sleep}$ and $N_{e^1}^{sleep}$ represent the number of observations in the two subsets, and N^{sleep} is the number of observations in the whole dataset ($N^{sleep} = 27$).

Logistic regression [53] was used to examine the contribution of user-specific factors to measurement errors because of the binary nature of the dependent variable. A value 1 of the dependent variable denotes the occurrence of measurement error, while a value 0 denotes the absence of measurement error (accurate measurement). Therefore, all observations in the subset $S_{e^0}^{sleep}$ were correspondent to an output of 0 while those in the subset $S_{e^1}^{sleep}$ were correspondent to an output of 1. Pearson product-moment correlation was used to investigate the overall linear relationships between user-specific factors (except gender which is binary) and absolute percent errors. The analysis results are described in detail in the next section.

4. Results

4.1. User Statistics

The demographic information and sleep baseline (measured by a medical EEG device) of the participants are summarized in Table 3. All participants were in their 20s or 30s. Ten out of the 27 participants had a PSQI higher than 5, which was indicative of poor sleep quality. Male participants on average had longer wake time, more awakenings, higher ratio of sleep stage N1, lower ratio of deep sleep and lower ratio of REM sleep. Taking $|e_i^{sleep}| \leq 5\%$ as the acceptable error range, the dataset was divided into two subsets according to the magnitude of the measurement errors. Table 4 presented the number of observations in each subset for the concerned sleep metrics. Wake after sleep onset (WASO) was excluded from the subsequent logistic regression analysis due to the small number of observations in $S_{e^0}^{WASO}$ for both consumer sleep tracking devices.

4.2. Results of Logistic Regression

Table 5 and Table 6 contain the results of logistic regression for Fitbit and Neuroon respectively. Notably, gender and subjective sleep quality (*PSQI*) were not associated to the measurement errors on any sleep metric. Age was strongly associated to the measurement accuracy of *TST* by both consumer devices. The probability of obtaining measurement errors by Fitbit significantly decreased for participants aged over 25-year-old on *TST* (OR = 0.05, 95% CI = 0.01-0.39, P = 0.004) and *SE* (OR = 0.17, 95% CI = 0.03-0.93, P = 0.041) compared to younger people.

Table 3. Descriptive statistics of the sleep dataset.

	All (n = 27)	Men (n = 16)	Women (n = 11)
age (years)	25.2 ± 3.4	25.5 ± 3.7	25.4 ± 3.5
PSQI	4.4 ± 2.3	4.7 ± 2.7	3.8 ± 1.3
TST (min)	359.9 ± 96.6	361.5 ± 97.9	367.0 ± 100.1
WASO (min)	17.9 ±12.7	21.4 ± 13.9	12.7 ± 8.0
NAWK (count)	17.9 ± 8.5	19.7 ± 8.5	15.3 ± 7.8
SOL (min)	14.4 ± 17.4	12.7 ± 15.7	16.3 ± 19.8
SE (%)	90.1 ± 8.5	90.3 ± 4.8	89.8 ± 12.3
N1R (%)	12.0 ± 8.4	14.3 ± 9.0	7.9 ± 5.6
N2R (%)	54.6 ± 8.5	54.7 ± 8.5	54.8 ± 8.3
DSR (%)	5.7 ± 7.2	4.6 ± 6.2	7.7 ± 8.1
RSR (%)	22.9 ± 5.8	20.9 ± 5.3	26.3 ± 4.9

Table 4. The number of observations in subset $S_{e^0}^{sleep}$ and $S_{e^1}^{sleep}$ for *TST*, *WASO* and *SE*.

	TST		*WASO*		*SE*	
	$N_{e^0}^{TST}$	$N_{e^1}^{TST}$	$N_{e^0}^{WASO}$	$N_{e^1}^{WASO}$	$N_{e^0}^{SE}$	$N_{e^1}^{SE}$
Fitbit	10	17	2	25	13	14
Neuroon	7	20	0	27	7	20

Similarly, the probability of obtaining measurement errors by Neuroon significantly decreased for participants aged over 26-year-old in comparison to younger participants (OR = 0.07, 95% CI = 0.01-0.55, P = 0.011). In addition, a higher ratio of sleep stage N1 (>10%) was associated to higher probability of obtaining accurate measurements on *SE* by Neuroon (OR = 0.09, 95% CI = 0.01-0.90, P = 0.041).

4.3. Results of Correlation Analysis

Pearson correlation coefficients were calculated to examine the linear relationships between user-specific factors and absolute percent errors by consumer sleep trackers. Gender was excluded from the analysis due to its binary nature. Table 7 presents the results of correlation analysis for Fitbit and Neuroon. No correlation was found from measurement errors to PSQI, average sleep cycle, rise time, the ratio of sleep stage N2, and the ratio of REM sleep.

Interestingly, absolute percent errors tend to negatively correlate to the underlying sleep metrics being measured. This phenomenon was observed for Fitbit on *TST* (r = -.46, P = 0.016), *WASO* (r = -.65, P < 0.001), *SE* (r = -.93, P < 0.001), and was observed for Neuroon on *WASO* (r = -.45, P = 0.020). For Fitbit, *SE* was found to be strongly and negatively associated to both $|e^{TST}|$ and $|e^{SE}|$, and *Age* was found to be moderately and negatively associated to $|e^{TST}|$ and $|e^{WASO}|$. In addition, *SOL* was moderately and positively associated to $|e^{TST}|$ and $|e^{SE}|$. Bedtime and deep sleep ratio were moderately and positively associated to $|e^{WASO}|$. In contrast, only two factors were found to be significantly associated to the measurement errors by Neuroon. Both *WASO* and the ratio of sleep stage N1 (*N1R*) were moderately and negatively associated to $|e^{WASO}|$.

5. Discussions

5.1. Principal Findings

Fitbit has been one of the main wearable vendors in the global market. Fitbit devices and smartphone apps enable users to monitor sleep in an unobtrusive way. On the other hand, Neuroon relies on embedded EEG sensors to enhance the accuracy of home sleep tracking and is increasingly gaining popularity. Previous validation studies have revealed the strength of Fitbit and Neuroon in measurement sleep duration and sleep efficiency as well as their weakness in measurement sleep stages [6, 11]. This study expanded our understanding on how device accuracy may be influenced by user-specific factors.

The analysis results found no influence from gender and subjective sleep quality measured by PSQI, whereas age and sleep structures were found to be significantly associated to device accuracy. Logistic regression analysis found that young adults above 26-year-old were more likely to obtain accurate data on total sleep time and sleep efficiency by both consumer devices. Correlation analysis also found significant and moderate negative relationship between age and absolute percent errors on *TST* by Fitbit. We therefore do not recommend the consumer devices to young people below 25-year-old when accurate estimates of *TST* and *SE* are important.

Table 5. Associations between user-specific factors and risk of measurement errors (Fitbit).

Factors		**Risk of errors on *TST***			**Risk of errors on *SE***		
		OR[a]	95% CI[b]	*P*	OR	95% CI	*P*
Age	≤26-year-old	Ref			Ref		
	>26-year-old	0.05	[0.01, 0.39]	**0.004**	0.17	[0.03, 0.93]	**0.04**
Sex	Female	Ref			Ref		
	Male	1.22	[0.24,6.11]	0.81	0.59	[0.12, 2.89]	0.52
PSQI	<5	Ref			Ref		
	≥5	1.56	[0.28, 8.53]	0.61	1.71	[0.34, 8.68]	0.52
SOL	≤10 min	Ref			Ref		
	>10 min	2.07	[0.40;10.84]	0.39	4.4	[0.84;23.58]	0.08
TST	<7 hours	Ref			Ref		
	7-9 hours	0.63	[0.12, 3.22]	0.57	2.50	[0.47, 13.27]	0.28
WASO	≤25 min	Ref			Ref		
	>25 min	1.23	[0.18;8.33]	0.83	0.38	[[0.06;2.52]	0.31
T_{avg}	≤90 min	Ref			Ref		
	>90 min	0.61	[0.12, 3.23]	0.56	0.59	[0.12, 2.89]	0.52
SE	≤90%	Ref			Ref		
	>90%	0.46	[0.07;2.89]	0.41	0.54	[0.10;2.93]	0.47
t_{start}	Before 0:00 am	Ref			Ref		
	After 0:00 am	3.6	[0.70, 18.56]	0.12	1.54	[0.33, 7.23]	0.58
t_{end}	Before 7:00 am	Ref			Ref		
	After 7:00 am	1.43	[0.30;6.88]	0.66	2.10	[0.45;9.84]	0.35
N1R	≤10%	Ref			Ref		
	>10%	1.23	[0.18;8.33]	0.83	0.91	[0.15;5.58]	0.92
N2R	≤60%	Ref			Ref		
	>60%	2.18	[0.35;1376]	0.41	1.85	[0.34;10.05]	0.47
DSR	≤5%	Ref			Ref		
	>5%	2.33	[0.44;12.40]	0.32	2.00	[0.41;9.84]	0.39
RSR	≤25%	Ref			Ref		
	>25%	0.31	[0.06;1.64]	0.17	0.32	[0.06;1.71]	0.18

[a]OR: odds ratio.
[b]CI: confidence interval.

Table 6. Associations between factors and risks of measurement errors (Neuroon)

Factors		**Risk of errors on *TST***			**Risk of errors on *SE***		
		OR[a]	95% CI[b]	*P*	OR	95% CI	*P*
Age	≤26-year-old	Ref			Ref		
	>26-year-old	0.07	[0.01;0.55]	**0.01**	0.44	[0.07;2.71]	0.38
Sex	Female	Ref			Ref		
	Male	0.60	[0.09;3.89]	0.59	0.00	[0.00;Inf]	0.99
PSQI	<5	Ref			Ref		
	≥5	3.27	[0.31;34.72]	0.32	1.45	[0.21;9.98]	0.70
SOL	≤10 min	Ref			Ref		
	>10 min	2.05	[0.32;13.16]	0.45	6.00	[0.61;59.30]	0.13
TST	<7 hours	Ref			Ref		
	7-9 hours	0.25	[0.04;1.52]	0.13	0.57	[0.10;3.38]	0.54
WASO	≤25 min	Ref			Ref		
	>25 min	2.00	[0.19;20.90]	0.56	0.63	[0.09;4.49]	0.64
T_{avg}	≤90 min	Ref			Ref		
	>90 min	0.60	[0.09, 3.89]	0.59	0.60	[0.09;3.89]	0.59
SE	≤90%	Ref			Ref		
	>90%	0.93	[0.14;6.23]	0.94	0.31	[0.03;3.11]	0.32
t_{start}	Before 0:00 am	Ref			Ref		
	After 0:00 am	2.48	[0.43;14.34]	0.31	2.48	[0.43;14.34]	0.31
t_{end}	Before 7:00 am	Ref			Ref		
	After 7:00 am	2.00	[0.35;11.44]	0.44	4.64	[0.71;30.42]	0.11
N1R	≤10%	Ref			Ref		
	>10%	0.27	[0.04;1.73]	0.16	0.09	[0.01;0.90]	**0.04**
N2R	≤60%	Ref			Ref		
	>60%	1.07	[0.16;7.15]	0.94	3.23	[0.32;32.48]	0.32
DSR	≤5%	Ref			Ref		
	>5%	6.67	[0.67;66.51]	0.11	2.25	[0.35;14.61]	0.40
RSR	≤25%	Ref			Ref		
	>25%	1.35	[0.21;8.82]	0.76	4.00	[0.40;39.83]	0.24

[a]OR: odds ratio.
[b]CI: confidence interval.

Table 7. Pearson correlation coefficients between user-specific factors and absolute percent errors.

	Fitbit			Neuroon		
	$\|e^{TST}\|$	$\|e^{WASO}\|$	$\|e^{SE}\|$	$\|e^{TST}\|$	$\|e^{WASO}\|$	$\|e^{SE}\|$
Age	**-0.45**[a]	**-0.54**	-0.27	-0.21	-0.31	-0.15
PSQI	0.22	-0.09	0.08	0.04	-0.12	0.00
SOL	**0.49**	0.31	**0.66**	0.29	-0.08	0.28
TST	**-0.46**	-0.16	-0.34	0.05	0.08	0.07
WASO	0.01	**-0.65**	-0.35	-0.15	**-0.45**	-0.13
T_{avg}	-0.34	-0.26	-0.31	-0.23	-0.49	-0.23
SE	**-0.75**	-0.15	**-0.73**	-0.09	0.30	-0.09
t_{start}	0.33	**0.39**	0.19	0.09	0.15	0.09
t_{end}	-0.06	0.15	-0.03	0.19	0.14	0.22
N1R	0.24	-0.19	0.17	-0.20	**-0.42**	-0.23
N2R	-0.32	-0.03	-0.19	0.08	0.10	0.11
DSR	0.01	**0.46**	0.13	0.13	0.32	0.10
RSR	-0.01	0.10	0.01	0.10	0.33	0.12

[a]Bold indicates statistically significant correlations.

The accuracy of consumer sleep tracking devices was also significantly associated to the sleep structure of users. In general, measurement errors on TST and SE by Fitbit were more pronounced in people with shorter sleep duration, longer sleep onset latency and lower sleep efficiency as estimated by the medical device. This finding was consistent with previous studies on clinical sleep monitors [13, 30, 33, 54] and older models of wearable trackers [24]. As for Neuroon, analysis results showed that people with higher ratio of sleep stage N1 were more likely to obtain accurate measurements on total sleep time.

Additionally, our study found that measurement errors on WASO by both devices decreased as WASO increased. This characteristic differentiates Fibit Charge 2 and Neuroon from older models of consumer sleep tracking devices that demonstrated the opposite behaviour [7]. A recent study found that Fitbit Charge 2 overestimated wake time compared to a medical device and attributed it to its tendency of misclassifying sleep epochs as awake [11]. Longer period of wakefulness was equivalent to more wake epochs and fewer sleep epochs, which were then translated into lower chance of misclassifying sleep as wake. This may explain the improved accuracy of Fitbit Charge 2 as the wake time increased. From the perspective of sensor characteristics, this counterintuitive phenomenon may also be attributed to the inconsistent sensitivity of consumer sleep tracking devices [55-57].

Overall, our findings have demonstrated that more emphasis should be placed on eliminating systematic measurement errors of total sleep time and sleep efficiency for young adults and for people with short and interrupted sleep patterns. Moreover, improving epoch wise classification accuracy between sleep and awake may help reduce measurement errors of wake time. The findings pointed out promising directions to designing new algorithms for accurate home sleep tracking.

5.2. Limitations

Our study into the associations between device accuracy and user-specific factors has several limitations. Firstly, the participants mostly included young healthy adults, thus limiting the generalizability of the findings to a more heterogeneous population such as teenagers, the elderly, and people with chronic conditions. Second, this study examined the relationships between accuracy and many factors only at the population level without considering individual differences. As such, the results may not be generalized for intrapersonal analysis. Third, the list of user-specific factors considered in this study was not exhaustive and the pathways thereby these factors affect measurement errors were still not thoroughly understood. Future researches are needed to address these limitations.

6. Conclusions

Wearable consumer sleep trackers are increasingly gaining popularity because they are unobtrusive, affordable, and have the potential to provide longitudinal monitoring. We have investigated the characteristics of the measurement errors of Fitbit Charge 2 and Neuroon under the influence of several user-specific factors. Age and sleep structure were significantly associated to the accuracy of consumer sleep trackers. Both devices had improved accuracy in measurement total sleep time and sleep efficiency for people above 26-year-old and for

people with longer sleep duration, less fragmented and deeper sleep. In addition, measurement accuracy on wake time was negatively correlated to the total duration of wake, which may due to the tendency of misclassifying sleep epochs as wake. Notably, we also found that gender and subjective sleep quality measured by PSQI were not associated to the measurement errors of neither device. Our study suggested that consumer sleep trackers may be less accurate for young adults and for people with poor sleep (especially when accurate estimates of total sleep time and sleep efficiency are important.). These characteristics should be accounted for in selecting devices and in designing new sleep tracking technologies.

Acknowledgements.

This work was supported by JSPS KAKENHI Grant-in-Aid for Research Activity Start-up (Grant Number 16H07469) and JSPS KAKENHI Grant-in-Aid for Early Career Scientists (Grant Number 19K20141). The authors would like to thank the participants of the study.

References

[1] Piwek, L., Ellis, D. A., Andrews, S., and Joinson, A. (2016) The rise of consumer health wearables: promises and barriers. *PLoS Med* **13**(2): e1001953.

[2] Liang, Z., and Ploderer, B. (2016) Sleep tracking in the real world: a qualitative study into barriers for improving sleep. In *Proceedings of the 28th Australian Conference on Computer-Human Interaction, Launceston*, Tasmania, Australia, 537-541.

[3] Liang, Z., and Chapa-Martell, M. A. (2015) Framing self-quantification for individual-level preventive health care. In *Proceedings of the International Joint Conference on Biomedical Engineering Systems and Technologies*, Lisbon, Portugal, 336-343.

[4] Liang, Z., Ploderer, B, and Chapa-Martell, MA, "Is fitbit fit for sleep-tracking? sources of measurement errors and proposed countermeasures. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*, Barcelona, Spain, 476-479.

[5] West, P, Van Kleek, M., Giordano, R., et al. (2017) Information quality challenges of patient-generated data in clinical practice. *Front Public Health* **5**:284.

[6] De Zambotti, M., Goldstone, A., Claudatos, S., et al. (2017) A validation study of Fitbit Charge 2 compared with polysomnography in adults. *Chronobiology International* **35**(4): 465-476.

[7] Meltzer, L., Hiruma, L., Avis, K., et al. (2015) Comparison of a commercial accelerometer with polysomnography and actigraphy in children and adolescents. *Sleep* **38**(8): 1323-1330.

[8] De Zambotti, M., Baker, F.C., and Colrain, I.M. (2015) Validation of sleep-tracking technology compared with polysomnography in adolescents. *Sleep* **38**(9):1461-1468.

[9] De Zambotti, M., Claudatos, S., Inkelis, S., et al. (2015) Evaluation of a consumer fitness-tracking device to assess sleep in adults. *Chronobiology International* **32**(7): 1024-1028.

[10] Montgomery-Downs, H., Insana, S., and Bond, J. (2012) Movement toward a novel activity monitoring device. *Sleep and Breathing* **16**(3):913-917.

[11] Liang, Z., and Chapa-Martell, M.A. (2018) Validity of consumer activity wristbands and wearable EEG for measurement overall sleep parameters and sleep structure in free-living conditions. *Journal of Healthcare Informatics Research* **2**(1-2):152-178.

[12] Marino, M., Li, Y., Rueschman, M., et al. (2013) Measurement sleep accuracy, sensitivity, and specificity of wrist actigraphy compared to polysomnography. *Sleep* **36**(11):1747-1755.

[13] Blackwell, T., Ancoli-Israel, S., Redline, S., et al. (2011) Factors that may influence the classification of sleep-wake by wrist actigraphy: the MrOS sleep study. *J Clin Sleep Med* **7**(4): 357-367.

[14] Buysse, D., Reynolds, C., Monk, T., et al. (1989) The Pittsburgh sleep quality index: a new instrument for psychiatric practice and research. *Psychiatry Res* **28**(2): 193-213.

[15] Natale, V., Leger, D., Martoni, M., et al. (2014) The role of actigraphy in the assessment of primary insomnia: a retrospective study. *Sleep Medicine* **15**(1): 111-115.

[16] Martin, J., and Hakim, A. (2011) wrist actigraphy. *CHEST* **139**(6): 1514-1527.

[17] Liang, Z., Ploderer, B., Chapa-Martell, M. A., and Nishimura, T. (2016) A cloud-based intelligent computing system for contextual exploration on personal sleep-tracking data using association rule mining. *Intelligent Computing Systems. Communications in Computer and Information Science*, Springer, Cham.

[18] Liang, Z., Chapa-Martell, M. A., and Nishimura, T. (2016) Mining hidden correlations between sleep and lifestyle factors from quantified-self data. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, Heidelberg, Germany, 547-552.

[19] Liang, Z., Chapa-Martell, M. A., and Nishimura, T. (2016) A personalized approach for detecting unusual sleep from time series sleep-tracking data. In *Proceedings of the IEEE International Conference on Health Informatics (ICHI)*, Chicago, US.

[20] Liang, Z., Ploderer, B., Liu, W., et al. (2016) SleepExplorer: a visualization tool to make sense of correlations between personal sleep data and contextual factors. *Personal Ubiquitous Comput* **20**(6): 985-1000.

[21] Weatherall, J., Paprocki, Y., Meyer, T. M., Kudel, I., and Witt, E. A. (2018) Sleep tracking and exercise in patients with type 2 diabetes mellitus (step-D): pilot study to determine correlations between Fitbit data and patient-reported outcomes. *JMIR Mhealth Uhealth* **6**(6): e131.

[22] Haeuber, E., Shaughnessy, M., Forrester, L., Coleman, K. L., and Macko, R. F. (2004) Accelerometer monitoring of home- and community-based ambulatory activity after stroke. *Archives of Physical Medicine and Rehabilitation* 85(12):1997-2001.

[23] Kanga, S.-G., Kanga, J. M., and Kob, K.-P. (2017) Validity of a commercial wearable sleep tracker in adult insomnia disorder patients and good sleepers. *Journal of Psychosomatic Research* **97**:38-44.

[24] Dickinson, D., Cazier, J., and Cech, T. (2016) A practical validation study of a commercial accelerometer using good and poor sleepers. *Health Psychol Open* **3**(2): 1-10.

[25] Liang, Z., and Chapa-Martell, M. A. (2019) Accuracy of Fitbit wristbands in measurement sleep stage transitions and the effect of user-specific factors. *JMIR Mhealth Uhealth* **7**(6): e13384.

[26] Liang, Z., and Chapa-Martell, M. A. (2019) Combining numerical and visual approaches in validating sleep data

quality of consumer wearable wristbands. In *Proceedings of PerCom Workshops*, Kyoto, Japan, 777-782.

[27] Liang, Z., and Chapa-Martell, M. A. (2019) Achieving accurate ubiquitous sleep sensing with consumer wearable activity wristbands using multi-class imbalanced classification. In *Proceedings of the 17th IEEE Intl Conf Perv Intel Comp (PICOM 2019)*, Fukuoka, Japan.

[28] Liang, Z., and Chapa-Martell, M. A. (2019) Combining resampling and machine learning to improve sleep-wake detection of Fitbit wristbands. In *Proceedings of the 7th International Conference on Healthcare Informatics*, Xi'an, China.

[29] Bellone, G., Plano, S., Cardinali, D., Perez Chada, D., Vigo, D., and Golombek, D. (2016) Comparative analysis of actigraphy performance in healthy young subjects. *Sleep Science* **9**:272-279.

[30] Lucey, B., McLeland, J., Toedebusch, C., Boyd, J., et al (2016) Comparison of a single-channel EEG sleep study to polysomnography. *J Sleep Res* **25**(6):625-635.

[31] Ferguson, T., Rowlands, A. V., Olds, T., and Maher, C. (2015) The validity of consumer-level, activity monitors in healthy adults worn in free-living conditions: a cross-sectional study. *International Journal of Behavioral Nutrition and Physical Activity* **12**(42): 1-9.

[32] O'Hara, E., Flanagan, D., Penzel, T., Garcia, C., Frohberg, D., and Heneghan, C. (2015) A comparison of radio-frequency biomotion sensors and actigraphy versus polysonography for the assessment of sleep in normal subjects. *Sleep Breath* **19**:91-98.

[33] Taibi, D., Landis, C., and Vitiello, M. (2013) Concordance of polysomnogaphic and actigraphic measurement of sleep and wake in older women with insomnia. *J Clin Sleep Med* **9**(3):217-225.

[34] Suh, S., Nowakowski, S., Bernet, R., et al. (2012) Clinical significance of night-to-night sleep variability in insomnia. *Sleep Medicine* **13**(5): 469-475.

[35] Buysse, D., Cheng, Y., Germain, A., et al. (2010) Night-to-night sleep variability in older adults with and without chronic insomnia. *Sleep Medicine* **11**(1): 56-64.

[36] Natale, V., Plazzi, G., and Martoni, M. (2009) Actigraphy in the assessment of insomnia: a quantitative approach. *Sleep* **32**(6): 767-771.

[37] Buysse, DJ. (2014) Sleep Health: Can We Define It? Does It Matter? *Sleep* **37**(1): 9-17.

[38] Yoda, K., Inaba, M., Hamanoto, K., et al. (2015) Association between poor Glycemic control, impaired sleep quality, and increased arterial thickening in type 2 diabetic patents. *Plus One* **10**(4): e0122521.

[39] Shrivastava, D., Jung, S., Saadat, M., et al. (2014) How to interpret the results of a sleep study. *J Community Hosp Intern Med Perspect* **4**:24983.

[40] Fung, M., Peters, K., Ancoli-Israel, I., et al. (2013) Total sleep time and other sleep characteristics measured by actigraphy do not predict incident hypertension in a cohort of community-dwelling older men. *J Clin Sleep Med* **9**(6): 585-591.

[41] Harvey, A., Stinson, K., Whitaker, K., et al. (2008) The subjective meaning of sleep quality: a comparison of individuals with and without insomnia. *Sleep* **31**(3): 383-393.

[42] Matsuo, M., Masuda, F., Sumi, Y., et al. (2016) Comparison of portable sleep monitors of different modalities: potential as naturalistic sleep recorders. *Front Neurol* **7**:110.

[43] Yoshida, M., Shinohara, H., and Kodama, H. (2015) Assessment of nocturnal sleep architecture by actigraphy and one-channel electroencephalography in early infancy," *Early Human Development* **91**(9): 519-526.

[44] McCall, C., and McCall, V. (2012) Objective vs subjective measurements of sleep in depressed insomniacs: first night effect or reverse first night effect? *J Clin Sleep Med* **8**(1): 59-65.

[45] Ahmadi, N., Shapiro, G., Chung, S., et al. (2009) Clinical diagnosis of sleep apnea based on single night of polysomnography vs two nights of polysomnography," *Sleep and Breathing* **13**(3): 221-226.

[46] Tworoger, S., Davis, S., Vitiello, M., et al. (2005) Factors associated with objective (actigraphic) and subjective sleep quality in young adult women. *Journal of Psychosomatic Research* **59**(1): 11-19.

[47] Ancoli-Israel, I., Chesson, A., and Quan, S. (2017) for the American Academy of Sleep Medicine. The AASM manual for the scoring of sleep and associated events rules, terminology and technical specifications. *Darien, IL: American Academy of Sleep Medicine*, Version 2.4.

[48] Shcherbina, A., Mattsson, M., Waggott, D., et al. (2017) Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. *J Pers Med* **7**(2): 3.

[49] Takacs, J., Pollock, C., Guenther, J., et al. (2014) Validation of the Fitbit One activity monitor device during treadmill walking. *Journal of Science and Medicine in Sport* **17**(5): 496-500.

[50] Meltzer, L., Walsh, C., Traylor, J., et al. (2012) Direct comparison of two new actigraphs and polysomnography in children and adolescents. *Sleep* **35**(1):159-166.

[51] Werner, H., Molinari, L., Guyer, C., et al. (2008) Agreement rates between actigraphy, diary, and questionnaire for children's sleep patterns. *JAMA Pediatrics* **162**(4): 350-358.

[52] Rosenberger, M., Buman, M., Haskell, W., et al. (2016) Twenty-four hours of sleep, sedentary behavior, and physical activity with nine wearable devices. *Medicine & Science in Sports & Exercise* **48**(3): 457-465.

[53] Cox, D. (1958) The regression analysis of binary sequences. *JSTOR* **20**(2): 215-242.

[54] Blackwell, T., Redline, S., Ancoli-Israel, I., et al. (2008) Comparison of sleep parameters from actigraphy and polysomnography in older women: the SOF study. *Sleep* **31**(2):283-291.

[55] Renk, E., Collins, W., Rizzo, M., et al. (2005) Calibrating a triaxial accelerometer-magnetometer. *IEEE Control Systems Magazine* **25**(6):86-95.

[56] Masurier, G.L. (2004) Pedometer sensitivity and specificity. *Medicine and Science in Sports and Exercise* **36**(2): 346.

[57] Roylance, L., and Angell, J. (1979) A batch-fabricated silicon accelerometer. *IEEE Transactions on Electron Devices* **26**(12): 1911-1917.