J Healthc Inform Res (2018) 2:152–178
https://doi.org/10.1007/s41666-018-0013-1

RESEARCH ARTICLE

Validity of Consumer Activity Wristbands and Wearable EEG for Measuring Overall Sleep Parameters and Sleep Structure in Free-Living Conditions

Zilu Liang[1] · Mario Alberto Chapa Martell[2]

Received: 15 June 2017 / Revised: 10 January 2018 / Accepted: 29 January 2018 /
Published online: 20 April 2018

Abstract Consumer sleep tracking technologies offer an unobtrusive and cost-efficient way to monitor sleep in free-living conditions. Technological advances in hardware and software have significantly improved the functionality of the new gadgets that recently appeared in the market. However, whether the latest gadgets can provide valid measurements on overall sleep parameters and sleep structure such as deep and REM sleep has not been examined. In this study, we aimed to investigate the validity of the latest consumer sleep tracking devices including an activity wristband Fitbit Charge 2 and a wearable EEG-based eye mask Neuroon in comparison to a medical sleep monitor. First, we confirmed that Fitbit Charge 2 can automatically detect the onset and offset of sleep with reasonable accuracy. Second, analysis found that both consumer devices produced comparable results in measuring total sleep duration and sleep efficiency compared to the medical device. In addition, Fitbit accurately measured the number of awakenings, while Neuroon with good signal quality had satisfactory performance on total awake time and sleep onset latency. However, measuring sleep structure including light, deep, and REM sleep remains to be challenging for both consumer devices. Third, greater discrepancies were observed between Neuroon and the medical device in nights with more disrupted sleep and when the signal quality was poor, but no trend was observed in Fitbit Charge 2. This study suggests that current consumer sleep tracking technologies may be immature for diagnosing sleep disorders, but they are reasonably satisfactory for general purpose and non-clinical use.

✉ Zilu Liang
z.liang@cnl.t.u-tokyo.ac.jp

✉ Mario Alberto Chapa Martell
mchapam0300@gmail.com

[1] The University of Tokyo, 7-3-1 Hongo, Bunkyo-Ku, Tokyo 113-8656, Japan

[2] CAC Corporation, 24-1, Hakozaki-cho, Nihonbashi, Chuo-Ku, Tokyo 103-0015, Japan

Keywords Sleep · Fitbit · Neuroon · Wearable devices · EEG · Validation · Self-tracking · mHealth

1 Introduction

With new gadgets entering the market every year, the field of consumer sleep tracking has been expanding rapidly under the influence of the Quantified Self movement [1]. Consumer sleep tracking devices enable the collection of longitudinal sleep data in non-clinical settings, which helps individuals understand sleep patterns and circadian rhythms in an unobtrusive and cost-efficient fashion [2]. Common sleep tracking devices, such as Fitbit and Jawbone, estimate sleep quality based on movement data measured by embedded accelerometers. Previous validation studies suggest that these devices tend to overestimate sleep and underestimate wake compared to clinical sleep monitors [3–5].

The findings from previous validation studies, however, may become out of date as new gadgets are release on the market. The manufacturers have been continuously improving the hardware and software of the wearable wristbands to reduce the discrepancy to clinical devices, and new functions such as sleep structure (light, deep, and REM sleep) detection have recently become available in new wristbands Fitbit Charge 2. On the other hand, new types of consumer sleep trackers based on electroencephalography (EEG) such as Neuroon and SleepShepherd also appeared in the consumer market. These EEG-based sleep trackers are claimed to be more accurate than accelerometer-based wristbands, but no study has validated this claim.

Therefore, the research community is calling for more validation studies on these devices especially in free living conditions [6]. In response to such need, this paper aimed to provide an update to the validity of the consumer sleep tracking technologies. We investigated the validity of two most up-to-date consumer wearable sleep tracking gadgets, i.e., Fitbit Charge 2 and Neuroon EEG eye mask. They were selected as the representatives of accelerometer-based devices and EEG-based devices, respectively, because they are popular, affordable and are available to be purchased online. In addition, Neuroon EEG eye mask is the only wearable EEG device that can be used concurrently with clinical devices due to sensors placement. In this study, we investigated the validity of consumer sleep trackers in measuring five overall sleep parameters including total sleep time (TST), wake after sleep onset (WASO), number of awakenings (NAWK), sleep onset latency (SOL), and sleep efficiency (SE), and sleep structure including light, deep, and REM sleep. These dimensions are the most measurable characteristics of human sleep and are closely related to health and well-being [7, 8]. The measurements of a clinical portable sleep monitor named SLEEP SCOPE (Sleep Well Co., Osaka, Japan) were used as the ground truth to compare with.

Analysis found that both Fitbit and Neuroon (when the signal quality of the electrodes was good) produced accurate results on TST and SE when compared to the medical device. In addition, good agreement was found on NAWK for Fitbit and on WASO and SOL for Neuroon, respectively. However, measuring sleep structure determined by sleep stage transitions was challenging for both devices. Regardless of the differences in hardware and software, both Fitbit and Neuroon (with good signal) underestimated light sleep while overestimated deep sleep in comparison to the medical device. Moreover, Fitbit underestimated REM and overestimated SOL, whereas Neuroon underestimated NAWK and overestimated REM. These results indicate that new mechanism and algorithms have improved the accuracy of consumer sleep tracking devices in measuring some of the overall sleep parameters, but there is still large room for improvement especially on sleep stage detection.

In what follows, we first summarize related work in the domain of human sleep and provide a literature review on the validation of consumer sleep-tracking technologies. We then describe the method that was used to validate the latest consumer devices in this study. In Sect. 4, we present the results of statistical analysis. Section 5 discusses how the analysis results updated the findings of previous studies. We also highlighted three directions for future research. By clarifying the strength and weakness of consumer sleep tracking technologies, this study can help both end-users and researchers select devices that best suit their needs.

2 Related Work

2.1 Fundamentals of Human Sleep

Human sleep can be measured along multiple dimensions such as quantity, continuity, and timing [9, 10]. Focusing on the most measurable characteristics of sleep that are closely related to physical and mental well-being, five dimensions are used to quantify sleep health [7]: sleep duration, sleep continuity or efficiency, timing, alertness/sleepiness, and satisfaction/quality.

The multiple dimensions of human sleep can be measured both objectively and subjectively [7]. Subjective methods include the Pittsburgh Sleep Quality Index (PSQI) [11] and sleep dairy [12]. The PSQI questionnaire is widely used in clinical settings for rough evaluation of sleep quality over the past 1 month. Nevertheless, PSQI may fail to capture the inter-night sleep variations. In contrast, sleep diary can be used to collect longitudinal sleep data for understanding long-term trends and patterns of sleep [12].

The objective methods for measuring human sleep focus on analyzing a set of sleep parameters [13], including total sleep time (TST), wake after sleep onset (WASO), number of awakenings (NAWK), sleep onset latency (SOL), sleep efficiency index (SE), REM sleep latency, total time in each sleep stage, and sleep stage ratio [14]. Human sleep that satisfies the following range is generally considered as abnormal: SOL $\geq$ 46 min, WASO $\geq$ 41 min, NAWK (for awakenings longer than 5 min) $\geq$ 4, SE $<$ 75%, REM ratio $\geq$ 41%, Stage 1 sleep $\geq$ 21%, Stage 2 sleep $\geq$ 81%, and Stage 3 + 4 sleep $<$ 10% [15–17]. However, since people's sleep needs varies significantly, there may be a wide spectrum of acceptable sleep structures in addition to the recommended standards. Polysomnography (PSG) and actigraphy are widely used tools for objectively measuring sleep in clinical settings. A PSG test measures a whole set of sleep parameters as well as many other physiological signals and is mainly used for diagnosing sleep diseases [14]. On the other hand, actigraphy is a wristband-like device that is widely used for diagnosing circadian-related disorders [18, 19]. A common problem of actigraphy is the underestimation of wake and overestimation of sleep [20].

Despite of measuring the same phenomenon, subjective sleep quality and objective sleep quality characterize different aspects of human sleep and are only modestly correlated [7, 8, 21]. Subjective sleep quality may not reliably reflect sleep patterns in some populations as people may have distorted impression on their sleep quality [22–26]. Therefore, we assessed the validity of consumer devices in comparison to objective sleep quality measured by a portable medical sleep monitor.

2.2 Validity of Consumer Sleep Tracking Technologies

Consumer sleep-tracking technologies help individuals monitor and reflect on sleep in home settings and enable researchers to conduct large-scale longitudinal studies at low cost. This field is expanding very rapidly, and new devices are entering the market every year. Based on the mechanisms of the technologies, consumer sleep tracking tools can be divided into two categories: accelerometer-based devices (e.g. mobile apps, activity wristbands, smart mattress) and EEG-based devices (e.g. ZEO headband, Neuroon eye mask, Sleep Shepherd headband). Comprehensive reviews on recent developments in home sleep-tracking devices and mobile apps can be found in [27–31].

A number of studies have quantitatively and qualitatively evaluated the validity of consumer sleep tracking devices. Quantitative validation studies compared the data obtained from consumer sleep tracking devices to measurements by clinical devices or instruments including PSG, actigraphy, and PSQI [27]. These studies mainly validated previous models of popular activity trackers for home sleep tracking, including Fitbit Tracker [32], Fitbit Ultra [4], and Jawbone [5, 33]. A few studies also investigated the accuracy of some mobile apps such as Sleep Time [34, 35]. These studies found that activity wristbands had the common problem of underestimating sleep disruptions and overestimating total sleep time and sleep efficiency in healthy adults. A few studies also analyzed the ability of consumer sleep trackers to measure sleep stages and found no correlations between these devices and PSG [5, 36].

Researchers in human-computer interaction have also investigated the validity of consumer sleep tracking technologies from user’s perspectives. The long-term impact of sleep-tracking was studied in [37], and measurement accuracy was highlighted as one of the main obstacles for improving sleep health using consumer sleep trackers. Following the same line, the authors of [38] investigated the sources of measurement errors and proposed countermeasures.

This study was designed to assess the validity of the latest consumer sleep trackers in measuring sleep structure as well as overall sleep parameters in free-living conditions, aiming at helping both individual users and researchers make informed decisions when adopting consumer sleep tracking devices for personal use and for scientific studies. This study provided new insights on the validity of activity wristbands and wearable EEG for home sleep tracking, and the results offered rich implications for future studies in this field. Specially, we were interested in the following three questions:

1. Do new consumer sleep tracking devices expand the ability of previous models in measuring sleep parameters including sleep structure in free-living conditions?
2. Are wearable EEG devices more accurate than activity wristbands for measuring sleep?
3. What are the limitations of these devices?

3 Methods

3.1 Sleep Parameters

Human sleep can be measured along multiple dimensions such as quantity, continuity, and timing [9, 39]. The outcome variables in this study were overall sleep parameters including total sleep time (TST), wake after sleep onset (WASO), sleep onset latency (SOL), number of awakenings (NAWK), sleep efficiency (SE), and sleep structure including light, deep, and REM sleep. These parameters are the most measureable characteristics of sleep that are closely related to physical and mental well-being [7], and some of them such as SOL and WASO are important indicators of sleep disorders [39–42]. The sleep parameters and their definitions are summarized in Table 1.

3.2 Devices

3.2.1 Fitbit

The Fitbit Charge 2 is a wearable activity wristband that tracks the frequency and intensity of a user's movements with an embedded triaxial accelerometer. It tracks sleep in addition to physical activity, workout, and calorie consumption. The normal sleep-recording mode was used during the data collection process, which accounts "significant movements (such as rolling over) as being awake, and is appropriate for most users" according to the manufacturer's website [43].

A Fitbit Charge 2 can automatically detect the start of sleep if a user has not moved for approximately 1 h. The reliability of the automatic detection was investigated in this study, and the results are presented in Sect. 4.2. The movement data were collected in 1-min epochs by default. After being synchronized to the Fitbit database, these data are mapped to aggregated sleep parameters such as total minutes asleep and minutes awake

Table 1 Definition of sleep parameters

Sleep parameters	Definition
Total sleep time (TST)	Time in minutes from sleep onset to sleep offset less the wake time; equal to the total of all REM and NREM sleep
Wake after sleep onset (WASO)	Periods of wakefulness occurring after defined sleep onset
Light sleep	The first and second stages of NREM sleep during which the heart rate slows and body temperature decreases
Deep sleep	The third stage of NREM sleep during which more slow-wave is observed and individuals have high awakening threshold to nonsignificant stimuli
REM sleep	A stage of sleep characterized by rapid eye movements and associated with dreaming
Number of awakenings (NAWK)	The number of awakenings occurring after defined sleep onset
Sleep onset latency (SOL)	Time in minutes from "light out" to the first epoch scored as sleep
Sleep efficiency (SE)	Percentage of total time in bed spent in sleep

using proprietary software and algorithms. The final results are then shown to the end users on the Fitbit dashboard. Since July 2017, Fitbit Charge 2 updated its software and started to use new algorithms that integrate heart rate data with movement data for sleep staging, which enables the device to detect sleep structure including light sleep, deep sleep, and REM sleep. Two screenshots of the Fitbit dashboard are shown in Fig. 1. The left one shows sleep readings inferred only from movement data, and the right one shows sleep stages inferred from a combination of heart rate data and movement data.

3.2.2 Neuroon

Different from Fitbit Charge 2, Neuroon is a wearable EEG eye mask that estimates a user's sleep based on the measurement of brainwave. This is done by using an embedded single channel EEG sensor. Neuroon was developed exclusively for sleep tracking, and it is claimed to be the first consumer sleep-tracking device that uses the same mechanism as portable clinical EEG sleep monitors. According to the manufacturer, Neuroon can also track other sleep-concurrent physiological parameters such as heart rate, eye ball movements, body temperature, and body movements. All these data were used to estimate overall sleep parameters and sleep stages by the company's proprietary algorithms. According to the manufacturer's website [44], the accuracy of Neuroon reached up to 94% compared to PSG. Figure 2 shows two screenshots of the Neuroon dashboard.

3.2.3 Sleep Scope

As shown in Fig. 3, Sleep Scope is a portable clinical 1-channel EEG device developed by the Sleep Well Company based in Osaka, Japan. Sleep Scope has been previously validated against PSG and achieved 86.9% agreement (average Cohen's Kappa value =

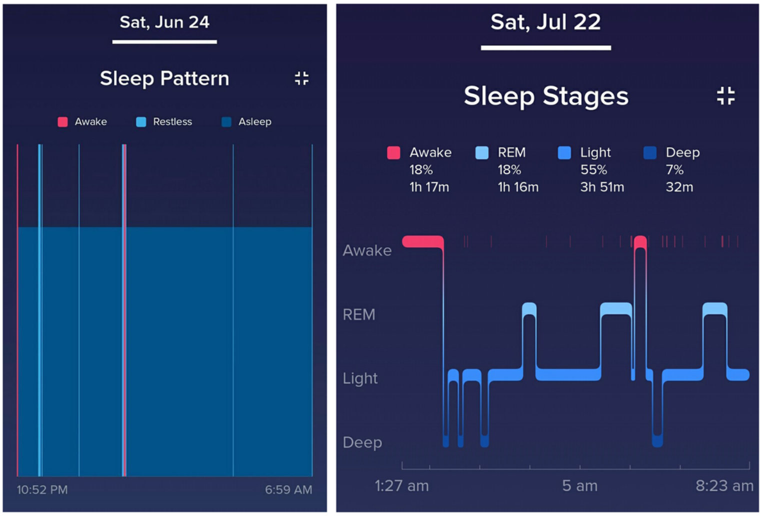
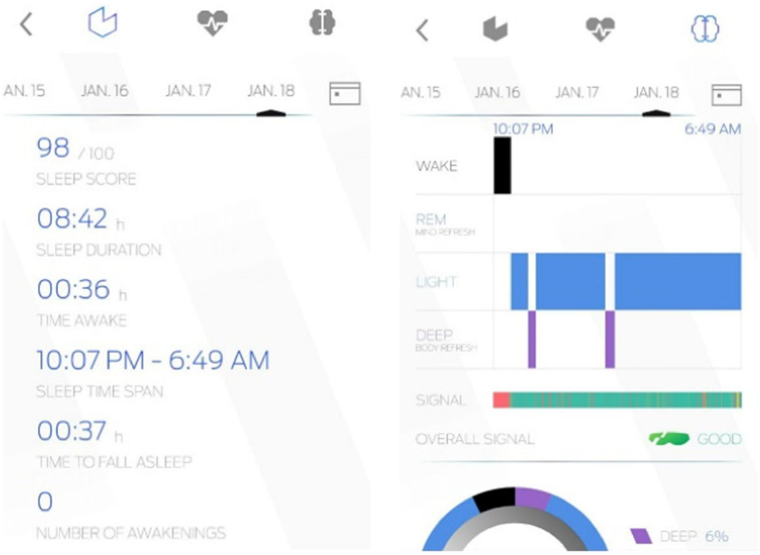


Fig. 1 Screenshots of sleep data on Fitbit dashboard. Sleep readings inferred only from movement data (left) and sleep stages inferred from a combination of heart rate data and movement data (right)



AN. 15 JAN. 16 JAN. 17 JAN. 18

98 /100
SLEEP SCORE

08:42 h
SLEEP DURATION

00:36 h
TIME AWAKE

10:07 PM - 6:49 AM
SLEEP TIME SPAN

00:37 h
TIME TO FALL ASLEEP

0
NUMBER OF AWAKENINGS

Fig. 2 Screenshots of Neuroon dashboard. Overall sleep parameters (left) and sleep structure (right)

0.753) [45]. Given that the average inter-scorer reliability on sleep stage scoring is 82.6% [46], Sleep Scope agrees well to PSG. We chose Sleep Scope as an alternative of PSG because the purpose of this study was not to diagnose sleep problems. Sleep Scope is more portable and less obtrusive in comparison to PSG, and it can be used to monitor sleep at a user's home, though still not as convenient as wearable devices. A Sleep Scope device uses two electrodes to be placed on the forehead and behind an ear. In this study, we used gel-type electrodes to improve the accuracy of EEG measurement. The data (i.e. raw EEG signals) from Sleep Scope need to be sent to the company for analysis. A sleep report will be generated based on the analysis and a sample report is shown in Fig. 4.

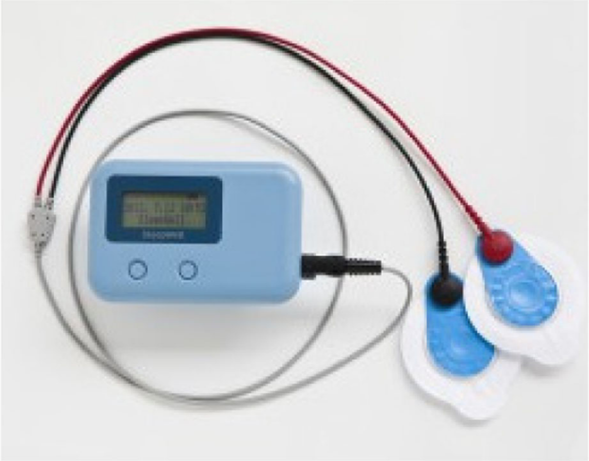
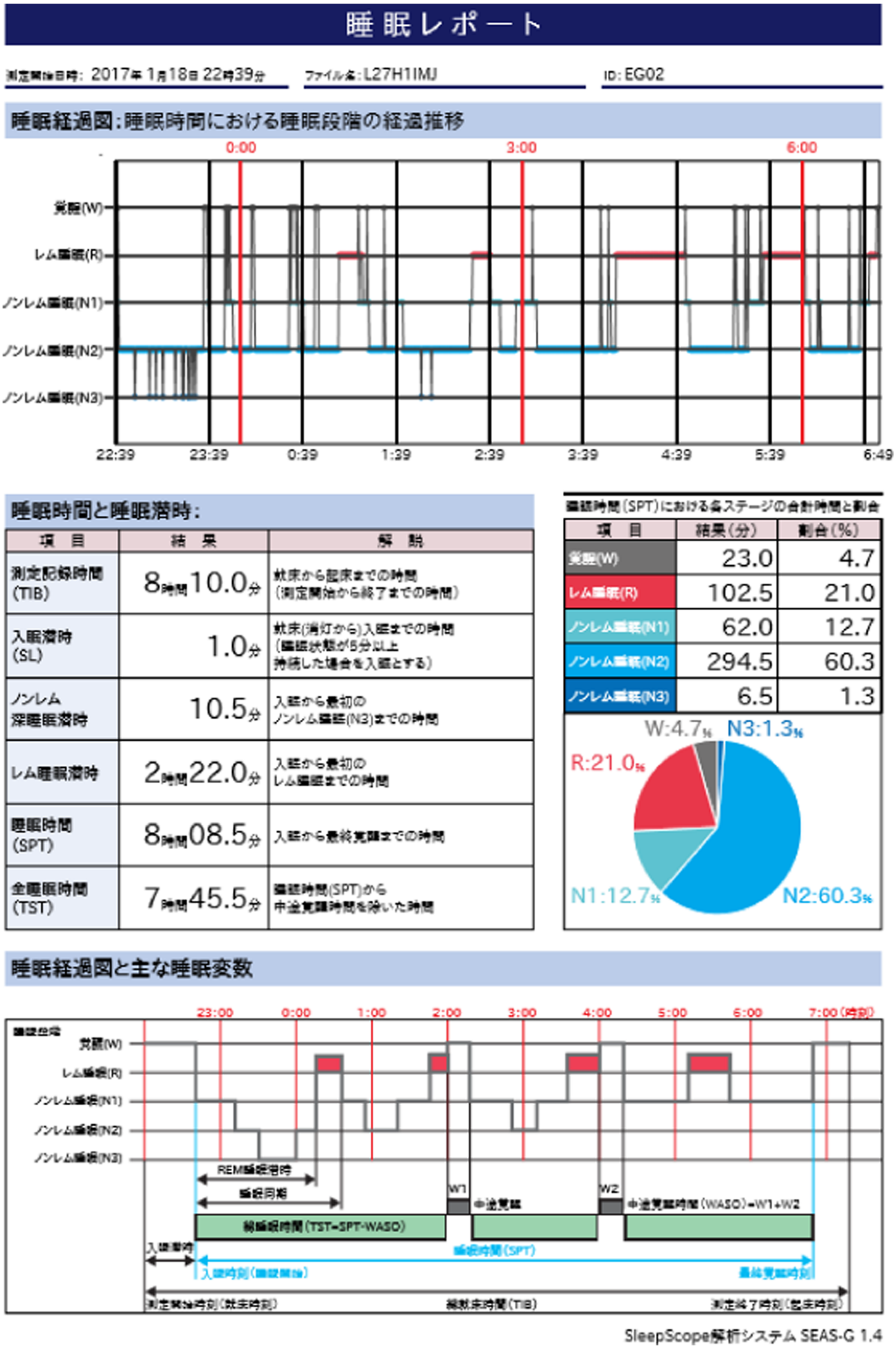


Fig. 3 Single-channel EEG clinical sleep monitor Sleep Scope (picture credits: https://sleepwell.co.jp/)



睡眠レポート

測定開始日時: 2017年 1月18日 22時39分　ファイル名: L27H1IMJ　ID: EG02

睡眠経過図：睡眠時間における睡眠段階の経過推移

0:00
3:00
6:00
覚醒(W)
レム睡眠(R)
ノンレム睡眠(N1)
ノンレム睡眠(N2)
ノンレム睡眠(N3)
22:39
23:39
0:39
1:39
2:39
3:39
4:39
5:39
6:49

睡眠時間と睡眠潜時：

項　目	結　果	解　説
測定記録時間 (TIB)	8時間10.0分	就床から起床までの時間 （測定開始から終了までの時間）
入眠潜時 (SL)	1.0分	就床(消灯から)入眠までの時間 （睡眠状態が5分以上持続した場合を入眠とする）
ノンレム 深睡眠潜時	10.5分	入眠から最初の ノンレム睡眠(N3)までの時間
レム睡眠潜時	2時間22.0分	入眠から最初の レム睡眠までの時間
睡眠時間 (SPT)	8時間08.5分	入眠から最終覚醒までの時間
全睡眠時間 (TST)	7時間45.5分	睡眠時間(SPT)から 中途覚醒時間を除いた時間

睡眠時間（SPT）における各ステージの合計時間と割合

項　目	結果（分）	割合（%）
覚醒(W)	23.0	4.7
レム睡眠(R)	102.5	21.0
ノンレム睡眠(N1)	62.0	12.7
ノンレム睡眠(N2)	294.5	60.3
ノンレム睡眠(N3)	6.5	1.3

睡眠経過図と主な睡眠変数

SleepScope解析システム SEAS-G 1.4

Fig. 4 A sample report of Sleep Scope (in Japanese only)

3.3 Study Procedure

3.3.1 Participants

We recruited 25 participants by distributing posters around the campus of the University of Tokyo. Screening conditions included no chronic conditions, no severe sleep problems or mental diseases, and being able to attend a briefing in person. There was no requirement on age, gender, and nationality. Participants filled in a PSQI (Pittsburg Sleep Quality Index) [11] 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 1 month, and a PSQI score equivalent to 5 or higher is indicator of poor sleep. The demographic information of the participants is summarized in Table 2. Ethics approval was obtained from the Ethic Committee of the University of Tokyo (Ethics ID: KE16–83). All participants provided informed consent.

3.3.2 Data Collection Procedure

Before the start of the self-tracking experiment, we held a briefing with each participant individually. In the briefing, we installed the Fitbit and Neuroon apps on participants' smartphones and explained how to use all the devices. We also explained the purpose of this study and informed the participants that they were free to stop the experiment if they noticed any discomfort. After the briefing, each participant was given the following items: a Fitbit Charge 2, a Neuroon, a medical device Sleep Scope, and necessary accessories such as chargers and batteries.

Thereafter, each participant tracked their sleep for three consecutive nights using all three devices concurrently. The self-tracking experiment took place in participants' homes. Fitbit Charge 2 and Neuroon were worn on the non-dominant wrist and on the lower forehead, respectively. The two electrodes of Sleep Scope were attached to the upper forehead and behind an ear, while the main body of the device was placed beside the participant's pillow. During the experiment, the participants did not need to charge the Fitbit as one charge lasts for more than 1 week. On the other hand, the battery of Neuroon only lasts for a couple of days, and we asked the participants to charge their Neuroon devices every morning after waking up to avoid device failures due to power-off. When participants completed the self-tracking experiment, they returned all the devices to us and received a coupon ($55) as appreciation for their participation.

Fitbit data were retrieved through Fitbit public API using a web application named SleepExplorer [47] which we developed in our previous study. Neuroon data were manually retrieved from the dashboard as there was no public API available. The EEG

Table 2 Demographic information of participants

	Age (years)	PSQI
All (*n* = 25)	24.8 ± 4.4	4.4 ± 2.3
Women (*n* = 10)	25.1 ± 3.6	3.8 ± 1.4
Men (*n* = 15)	24.7 ± 2.9	4.8 ± 2.8

data from the medical monitor were extracted from the SD card of the device and were forwarded to the Sleep Well Company for analysis. At the Sleep Well Company, raw EEG data were routinely analyzed at 30-s epoch, and the sleep stages were determined using proprietary automatic scoring system. The validity of the sleep staging was then visually assessed epoch-by-epoch by specialists according to sleep scoring standards [48], and corrections were added when necessary. We analyzed the second night for each participant to remove “the first night effect” [49], which is a common practice in sleep research [50, 51]. If any of the devices produced obviously wrong data on the second night, the third night was analyzed; the first night was analyzed only when the data of both the second and the third night were unreliable. Eventually, we obtained a dataset with 25 data entries.

3.4 Data Analysis

3.4.1 Data Preprocessing

Since consumer sleep trackers have their own naming scheme of the sleep parameters, the first step in data processing was to clarify the definitions of the sleep parameters in all devices and to ensure that the parameters measured the same underlying phenomenon across devices. As is summarized in Table 3, the sleep parameters measured by Fitbit and Neuroon were mapped to well-defined terminology in sleep science. For example, the length of sleep was described using *Minutes Asleep* by Fitbit and *Sleep Duration* by Neuroon. However, these two measures were correspondent to different terminology in sleep science; the former refers to total sleep time (TST), while the latter refers to time in bed (TIB). Similarly, the *Sleep Score* calculated by Neuroon did not have clinical meaning, and we had to calculate Sleep Efficiency according to its definition in sleep science, which is the ratio of the total time spent sleep (TST) compared to the total amount of time spent in bed (TIB). In addition, Neuroon only provided the ratio of each sleep stage, and there was no information on the duration of each sleep stage. Assuming that participants stopped the device immediately upon

Table 3 Mapping sleep parameters measured by consumer sleep trackers to clinical terminology

Clinical terminology	Fitbit	Neuroon
Total sleep time (TST)	Minutes asleep	Sleep duration − time awake
Wake after sleep onset (WASO)	Minutes awake	Time awake
Stage N1 + N2 (light sleep)	Light	Light ratio × sleep duration
Stage N3 (deep sleep)	Deep	Deep ratio × sleep duration
REM sleep	REM	REM ratio × sleep duration
Number of awakenings (NAWK)	Number of awakenings + number of restlessness	Number of awakenings
Sleep onset latency (SOL)	Time to fall asleep	Time to fall asleep
Sleep efficiency (SE)	Sleep efficiency	$1-\frac{\text{Time awake}}{\text{Sleep duration}}$

waking up, we calculated the duration of each sleep stage for Neuroon by multiplying the ratio of each sleep stage by *Sleep Duration*.

3.4.2 Statistical Analysis

The validity of a consumer sleep tracking device refers to how well this device actually measures the underlying sleep phenomenon in comparison to the ground truth [52, 53]. To gain a sense of the overall performance of Fitbit Charge 2 and Neuroon, we compared them to the medical device using the following statistical techniques.

- Box-and-whisker plot [54] was used to create intuitive display of the distribution of sleep data measured by three devices based on minimum, first quartile, median, third quartile, and maximum.
- Wilcoxon Signed-Rank test [55] was used to quantitatively compare the overall distribution of the measurements by the consumer trackers and that of the medical device. It is worth noting that WSR test was chosen over *t* test given the relatively small sample size and thus the potential non-normality of the dataset [56, 57].
- Bland-Altman plot [58] was used to examine the level of agreement between the consumer devices and the medical device. In clinical settings, if the within-mean differences equal ± 1.96 SD (standard deviation) are not clinically important, then the two devices are equivalent and may be used interchangeably [53].
- Pearson correlation coefficient was used to assess the linear relationship between the consumer devices and the medical device.

Similar to the criteria used in [59, 60], we defined the acceptable error range as ≤ 30 min for TST and < 5% for SE. For other sleep parameters and sleep structure, we determined an error rate of 5% ($p = 0.05$) to be within acceptable limits since this approximates a widely acceptable standard for statistical significance in health sciences research [61]. The analysis results are described in detail in the next section.

4 Results

4.1 Descriptive Statistics

Compared to the medical device, Fitbit Charge 2 showed lower values for TST (Fitbit 338.2 ± 94.1 min; medical device 350.5 ± 94.7 min, $z = 2.10$, $p = 0.036$), SOL (Fitbit 3.5 ± 4.1 min; medical device 14.6 ± 18 min, $z = 4.26$, $p < 0.0000$), light sleep (Fitbit 201.9 ± 42.8 min; medical device 244.3 ± 74.1, $z = 5.25$, $p = 0.0006$) and REM sleep (Fitbit 72.6 ± 27.7 min; medical device 84.2 ± 32.7 min, $z = 2.66$, $p = 0.0006$), and higher values for WASO (Fitbit 41.6 ± 18.8 min; medical device 17.1± 12.9 min, $z = -4.09$, $p < 0.0000$) and deep sleep (Fitbit 61.8 ± 19.7 min; medical device 22.0 ± 30.1 min, $z = -3.60$, $p < 0.0000$). No statistically significant difference was found between the two devices on NAWK (Fitbit 19.9 ± 7.9 counts; medical device 16.8 ± 8.1 counts, $z = -1.69$, $p = 0.093$) and SE (Fitbit 88.4 ± 3.6%; medical device 89.9 ± 8.8%, $z = 1.84$, $p = 0.067$). As for the proportion of individual sleep stages, Fitbit overestimated the ratio of wake (Fitbit 11.1 ± 3.6%; medical device 4.6 ± 3.2%, $z = -$

4.05, $p < 0.0000$), light sleep (Fitbit 66.6 ± 8.3%; medical device 52.8 ± 4.1%, $z = 3.95$, $p < 0.0000$), and REM sleep (Fitbit 22.7 ± 5.9%; medical device 18.1 ± 4.9%, $z = 3.04$, $p = 0.001$), while underestimated the ratio of deep sleep (Fitbit 6.1 ± 7.3%; medical device 16.3 ± 4.2%, $z = -3.56$, $p = 0.0001$).

On the other hand, Neuroon showed lower values for TST (Neuroon 194.0 ± 131.0 min; medical device 350.5 ± 94.7 min, $z = 3.69$, $p < 0.0000$), NAWK (Neuroon 3.2 ± 3.2 count; medical device 16.8 ± 8.1 min, $z = 4.38$, $p < 0.0000$), SE (Neuroon 53.6 ± 34.3%; medical device 89.9 ± 8.8%, $z = 3.43$, $p = 0.0002$), light sleep (Neuroon 79.1 ± 86.3 min; medical device 244.3 ± 74.1, $z = 4.17$, $p < 0.0000$), and high values for SOL (Neuroon 67.2 ± 115.8 min; medical device 14.6 ± 18.0, $z = -2.87$, $p =$ 0.0030), WASO (Neuroon 189.7 ± 159.5 min; medical device 17.1 ± 12.9 min, $z = -$ 4.16, $p < 0.0000$), and deep sleep (Neuroon 42.4 ± 44.3 min; medical device 22.0 ± 30.1 min, $z = -2.06$, $p = 0.039$). The results indicated no significant difference between Neuroon and the medical device on REM sleep (Neuroon 64.6 ± 57.4 min; medical device 84.2 ± 32.7 min, $z = 1.29$, $p = 0.208$). As for the proportion of individual sleep stages, no statistically significant difference was found on REM (Neuroon 18.3 ± 14.7%; medical device 18.1 ± 4.9%, $z = 1.23$, $p = 0.225$). However, Neuroon overestimated the ratio of wake (Neuroon 44.1 ± 33.5%; medical device 4.6 ± 3.2%, $z = -3.97$, $p < 0.0000$) and underestimated the ratio of light sleep (Neuroon 22.7 ± 22.2%; medical device 0.528 ± 0.041%, $z = 4.17$, $p < 0.0000$) and deep sleep (Neuroon 13.1 ± 14.7%; medical device 16.3 ± 4.2%, $z = -2.4332$, $p = 0.013$).

The box-and-whisker plots of the sleep parameters measured by Fitbit, Neuroon, and the medical device are shown in Fig. 5. Thick lines indicate the median, box edges represent the 25–75% quartile range (interquartile range), and the whiskers indicate the overall range, while circles indicate potential outliers. The plots showed that Neuroon data had larger dispersion in comparison to Fitbit Charge 2 and the medical device on all sleep parameters except on TST, light sleep, and NAWK.

4.2 Agreement Between Consumer and Medical Device

4.2.1 Fitbit for Automatic Detection on Sleep Onset and Offset

We defined two metrics, sleep start delay Δt_s and sleep end delay Δt_e, to quantify the lag in sleep onset and offset detected by Fitbit in comparison to those measured by the medical device. The two metrics can be calculated using Eqs. (1) and (2). For sleep start delay, $\Delta t_s < 0$ may be explained either by long sleep onset latency or by user' habits such as reading in bed or watching television in coach. In case of $\Delta t_s > 0$, a short delay may be reasonable due to the lag in physiological change during sleep onset. However, long positive delay may indicate a measurement error due to misclassification of other sleep stages as SOL [38]. For sleep end delay, $\Delta t_e < 0$ may be reasonable because users needed to manually turn off the medical device and such action always happened after the actual sleep offset. Contrarily, $\Delta t_e > 0$ may indicate measurement errors due to misclassification by Fitbit.

$$\Delta t_s = t_{s,Fitbit} - t_{s,Medical} \tag{1}$$

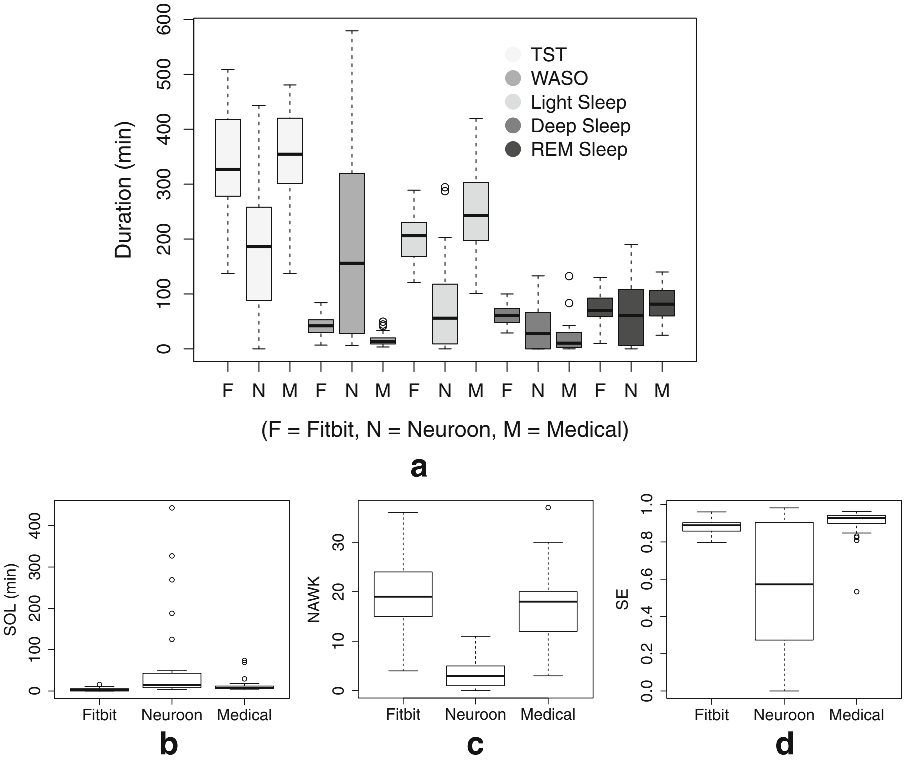


Fig. 5 Box-and-whiskers plot depicting different distributions for sleep parameters measured by three devices. Thick lines indicate the median, box edges represent the 25–75% quartile range, and the whiskers indicate the overall range. **a** TST, WASO, light, deep, and REM. **b** SOL. **c** NAWK. **d** SE

$$\Delta t_e = t_{e,Fitbit} - t_{e,Medical} \qquad (2)$$

The distribution of Δt_s and Δt_e is plotted in Fig. 6. In most cases, Δt_s is within the range of [0, 20 min] and Δt_e is within the range of [− 20 min, 20 min]. We noticed overlong Δt_s (> 20 min) for six participants. Four out of the six cases could be explained by overlong sleep onset latency indicated either by medical data or by participants' complaints. However, the rest two cases can only be attributed to measurement errors. As for Δt_e, the outlier at the right side was due to a human error. The participant turned off the medical device after sleep offset but later returned to bed and slept for another 1.5 h.

4.2.2 Agreement of Consumer Devices to Medical Device

The Bland-Altman plots of consumer devices versus the medical device are shown in Figs. 7 (Fitbit Charge 2) and 8 (Neuroon). A few measurements were situated out of the range between the lower limit of agreement (LLA) and the upper limit of agreement (ULA). According to the clinically satisfactory ranges defined in [59, 60], i.e., mean difference on TST ≤ 30 min and on SE < 5%, both Fitbit Charge 2 and Neuroon produced comparable results as the medical device in measuring TST and SE.

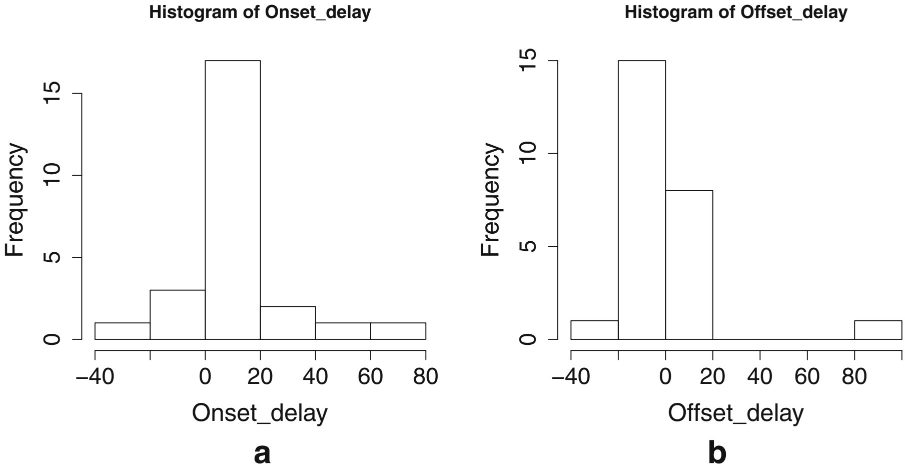


Fig. 6 Histogram of delay in **a** sleep onset and **b** sleep offset measured by Fitbit in comparison to medical device. The results show that Fitbit can automatically detect the start and end of sleep with reasonable accuracy in most cases

In general, Fitbit Charge 2 agreed well to the medical device on TST, NAWK, and SE (Fig. 7a, g, h). However, sleep parameters characterizing sleep-wake transitions (SOL, WASO) or sleep stage transitions (light, deep, REM) showed poor agreement between Fitbit and the medical device. In comparison, measurements by Neuroon markedly deviated from those by the medical device on all sleep parameters as shown in Fig. 8. The Bland-Altman plots for Neuroon and the medical device also demonstrated trends in device difference as a function of the sleep parameters. In general, Neuroon deviated more from the medical device with more disrupted sleep that was characterized by longer awakenings, higher frequency of awakenings, longer sleep onset, or lower sleep efficiency. No trend was observed for Fitbit.

The Pearson correlation coefficients between the consumer devices and the medical device are summarized in Table 4. As for Fitbit, very strong correlation on TST, strong correlations on light sleep and REM sleep, and moderate correlation on NAWK were found. However, only moderate correlation on deep sleep was identified between Neuroon and the medical device.

4.3 Impact of Signal Quality on Neuroon

Previous studies found that signal quality may affect the accuracy of EEG devices [62, 63]. Neuroon shows the *Overall Signal* of the three electrodes on the dashboard to help users estimate how accurate the sleep readings may be. The signal quality was either good (= 2, green), average (= 1, yellow), or poor (= 0, red), as is shown in Fig. 2. Several participants mentioned that the Neuroon mask moved off their face during the course of the night and the corresponding *Overall Signal* was poor.

Out of the 25 entries in the dataset, there were 14 entries of poor signal, 4 entries of average signal, and 7 entries of good signal. We therefore divided the dataset into two subsets, i.e., a subset with samples above average signal quality (= 1 or 2) and a subset with samples of poor signal quality, and then investigated how the validity of Neuroon differed between the two subsets.

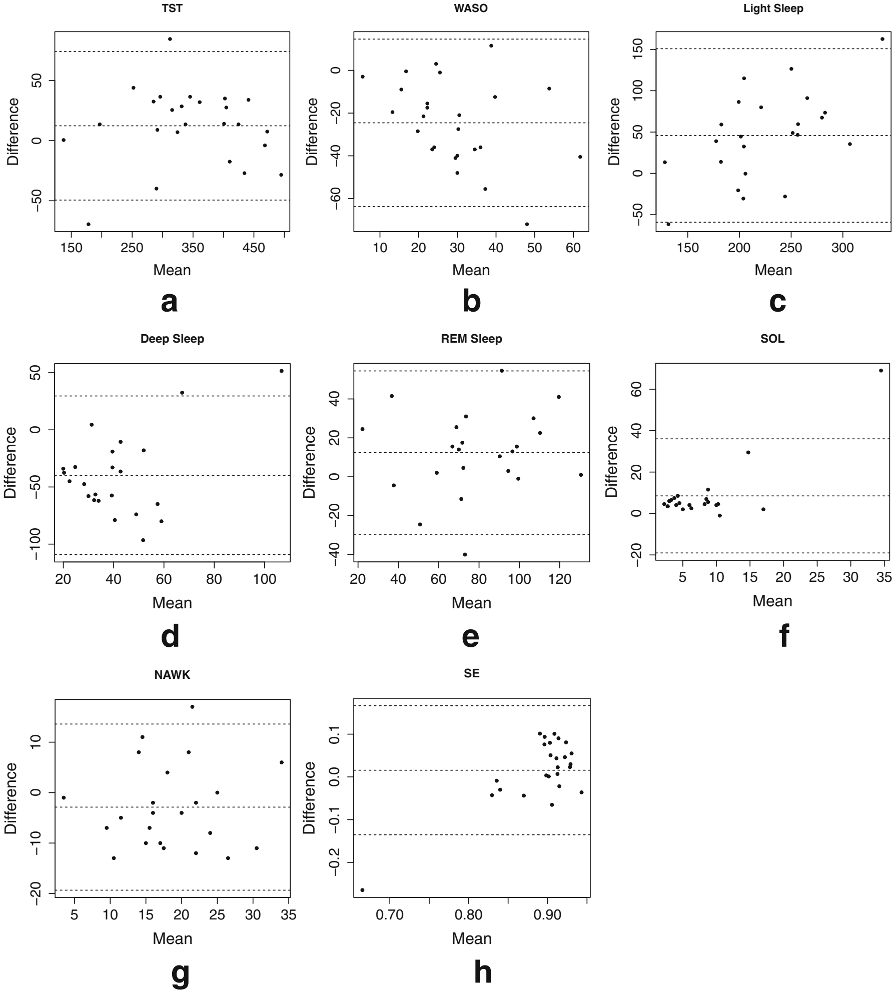


Fig. 7 Altman plots for Fitbit versus medical device. The dashed lines represent the mean difference (bias), − 1.96 standard deviation (LLA), and + 1.96 standard deviations (ULA). Fitbit had good agreement to the medical device on TST, NAWK, and SE/. **a** Total sleep time, **b** wake time after sleep onset, **c** light sleep, **d** deep sleep, **e** REM sleep, **f** sleep onset latency, **g** number of awakenings, **h** sleep efficiency

The box-and-whiskers plots of the two subsets are shown in Fig. 9 (for TST, WASO, light, deep, REM) and Fig. 10 (for SOL, NAWK, SE). As expected, signal quality had strong impact on the performance of Neuroon, especially for TST, WASO, and light sleep. These plots show that the distribution of Neuroon data markedly deviated from that of the medical device when the signal quality was poor, characterized by significant underestimation of sleep (and thus all sleep stages) and overestimation of awake. Using the medical measurements as baselines, we summarized the mean bias of Neuroon with varied signal quality in Table 5 (the overall performance of Fitbit Charge 2 and Neuroon was also shown as references). In order to show the upper and lower bound of Neuroon's performance, we only included data entries with good signal quality (= 2) and poor signal quality (= 0) in the analysis, and the four data entries with average signal quality (= 1) were excluded.

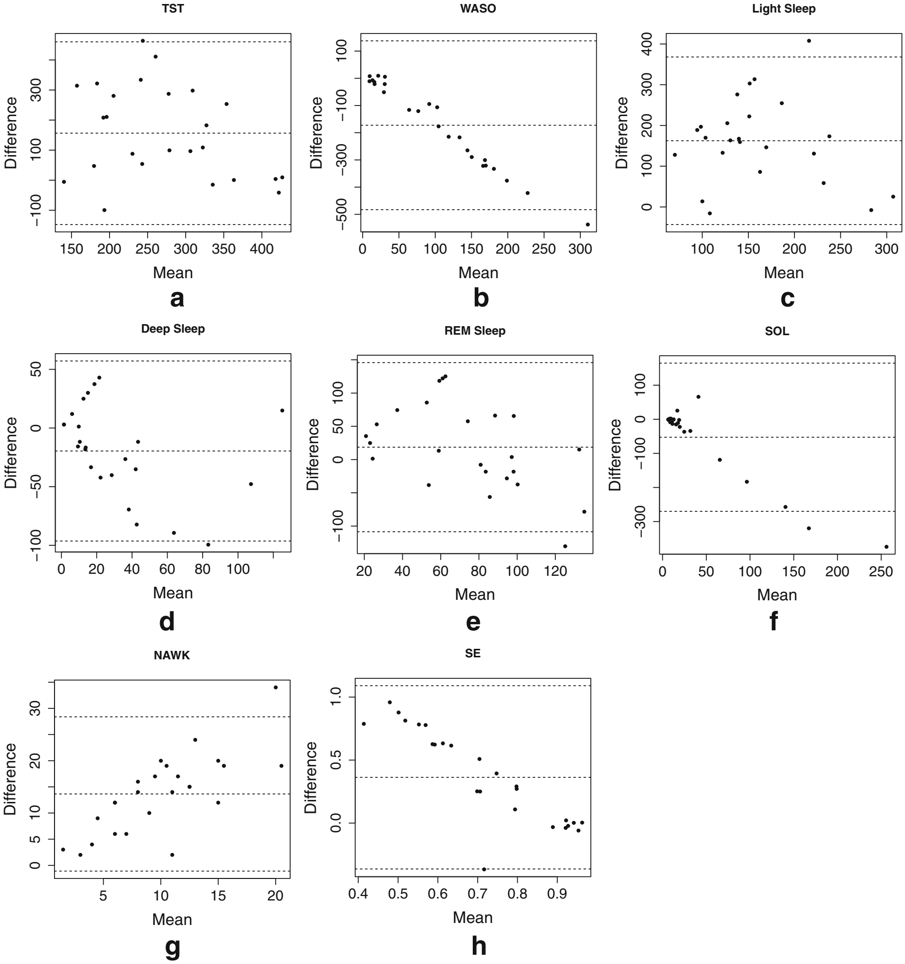


Fig. 8 Bland-Altman plots for Neuroon versus medical devices. The dashed lines represent bias, LLA, and ULA. Neuroon significantly deviated from the medical device on all sleep parameters. **a** Total sleep time, **b** wake time after sleep onset, **c** light sleep, **d** deep sleep, **e** REM sleep, **f** sleep onset latency, **g** number of awakenings, **h** sleep efficiency

The results revealed that when the signal quality of Neuroon was good, no significant differences were found between Neuroon and the medical device on TST, WASO, SOL, and SE. However, when the signal quality was poor, Neuroon

Table 4 Pearson correlation coefficients of sleep parameters between consumer and medical devices

	TST	WASO	Light sleep	Deep sleep	REM sleep	NAWK	SOL	SE
Fitbit vs medical device	**.94**[a]****	.25	**.65*** **	.08	**.73**** **	**.46***	−.04	**.50***
Neuroon vs medical device	.08	.13	.15	**.50***	.05	.37	.34	−.20

[a] Bold indicates a significant correlation (*$p \le 0.05$, **$p \le 0.01$, ***$p \le 0.001$, ****$p < 0.000$)

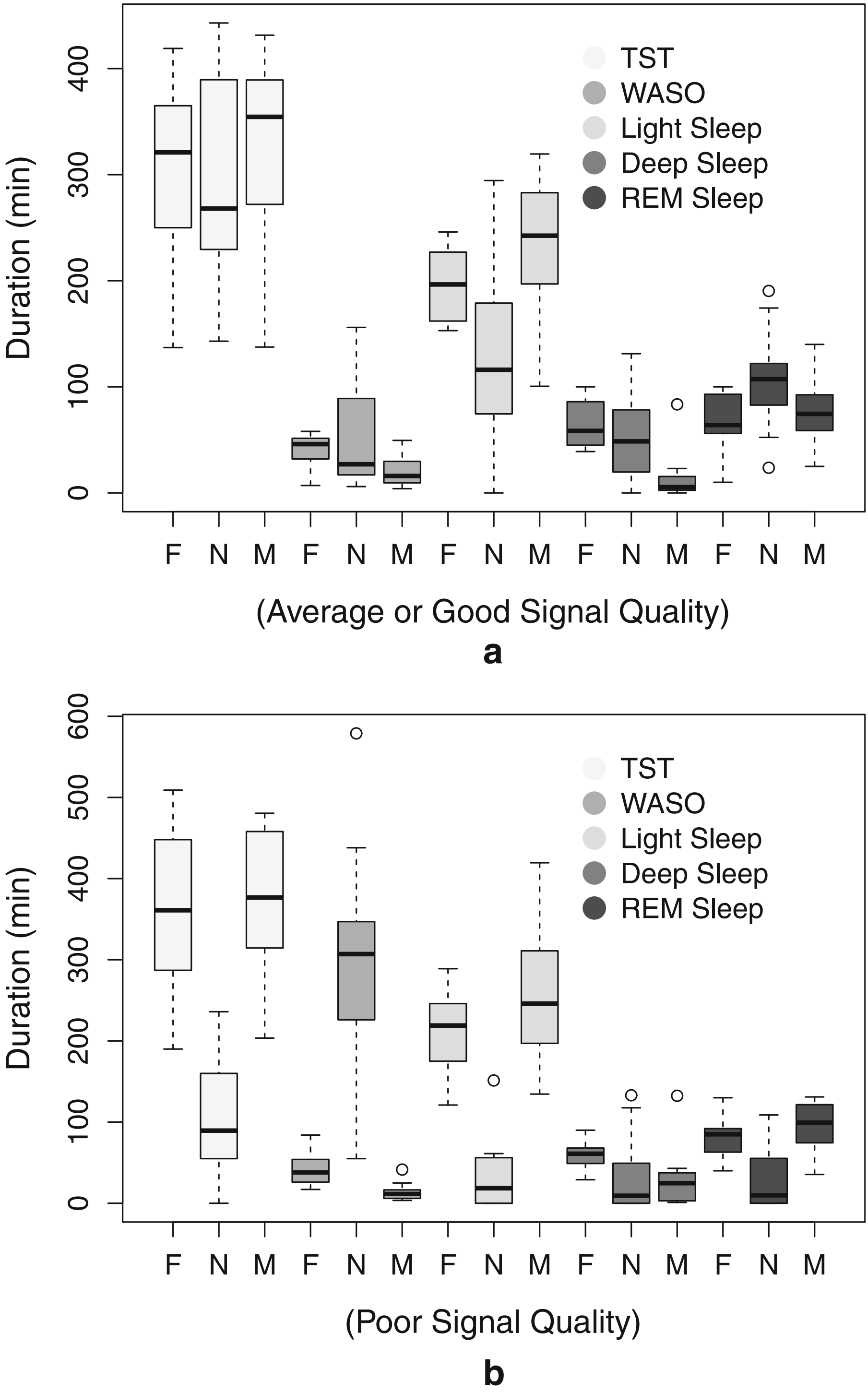


Fig. 9 Box-and-whiskers plots depicting different distributions for TST, WASO, and sleep stages when the signal quality of Neuroon was **a** above average and **b** poor

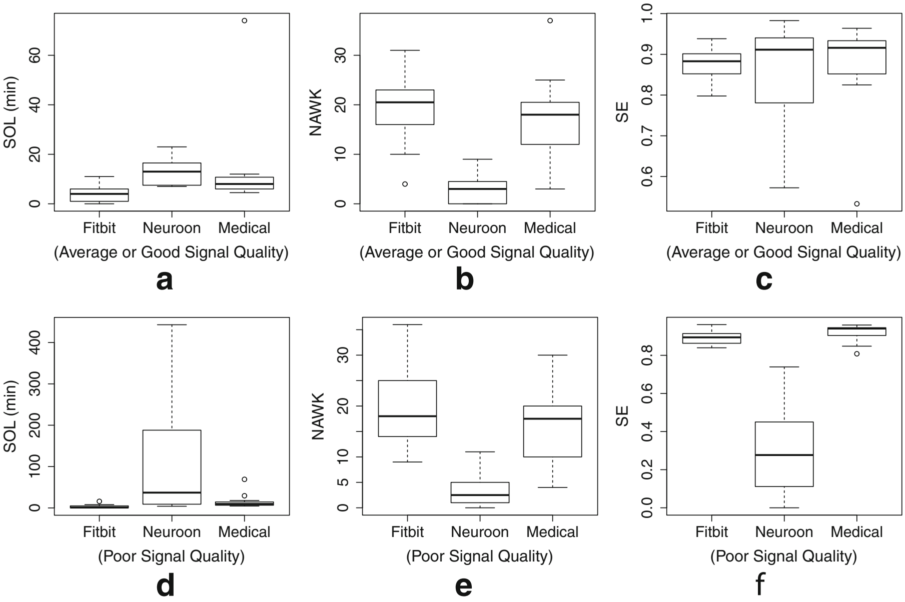


Fig. 10 Box-and-whiskers plots depicting different distributions for SOL, NAWK, and SE when the signal quality of Neuroon was **a~c** above average and **d~f** poor

underestimated TST and SE and overestimated WASO and SOL. We also found that Neuroon constantly underestimated NAWK and light sleep. As for deep sleep, Neuroon overestimated it when the signal quality was good but produced similar results as the medical device when the signal quality was poor. In addition, Neuroon overestimated REM when the signal quality was good and underestimated it when the signal quality was poor.

The Bland-Altman plots that characterize the agreement between Neuroon and the medical devices for the two data subsets are depicted in Figs. 11 (for TST, WASO, SOL, NAWK, and SE) and 12 (for light, deep, and REM). Poor signal quality led to increased discrepancy between Neuroon and the medical device on all sleep parameters except on NAWK and deep sleep, indicating that the poor agreement on NAWK could not be attributed to signal quality, and the ability to measure deep sleep regardless of signal quality may be a strength of Neuroon.

Table 5 Mean bias between consumer devices and medical devices

	TST (min)	WASO (min)	Light (min)	Deep (min)	REM (min)	SOL (min)	NAWK (count)	SE (%)
Fitbit	**− 12.3**[a]	24.5	− 42.4	39.8	− 11.6	− 11.1	**3.1**	**1.5**
Neuroon	− 156.5	172.6	− 165.2	20.4	− 19.6	52.6	− 13.6	− 36.3
Neuroon (signal quality = 2)	**− 6.9**	**4.4**	− 64.7	38.0	11.6	**4.5**	− 15.5	**1.7**
Neuroon (signal quality = 0)	− 265.1	281.6	− 223.6	5.8	− 62.1	94.7	− 12.9	− 62.7

[a] Bold indicates bias within acceptable range

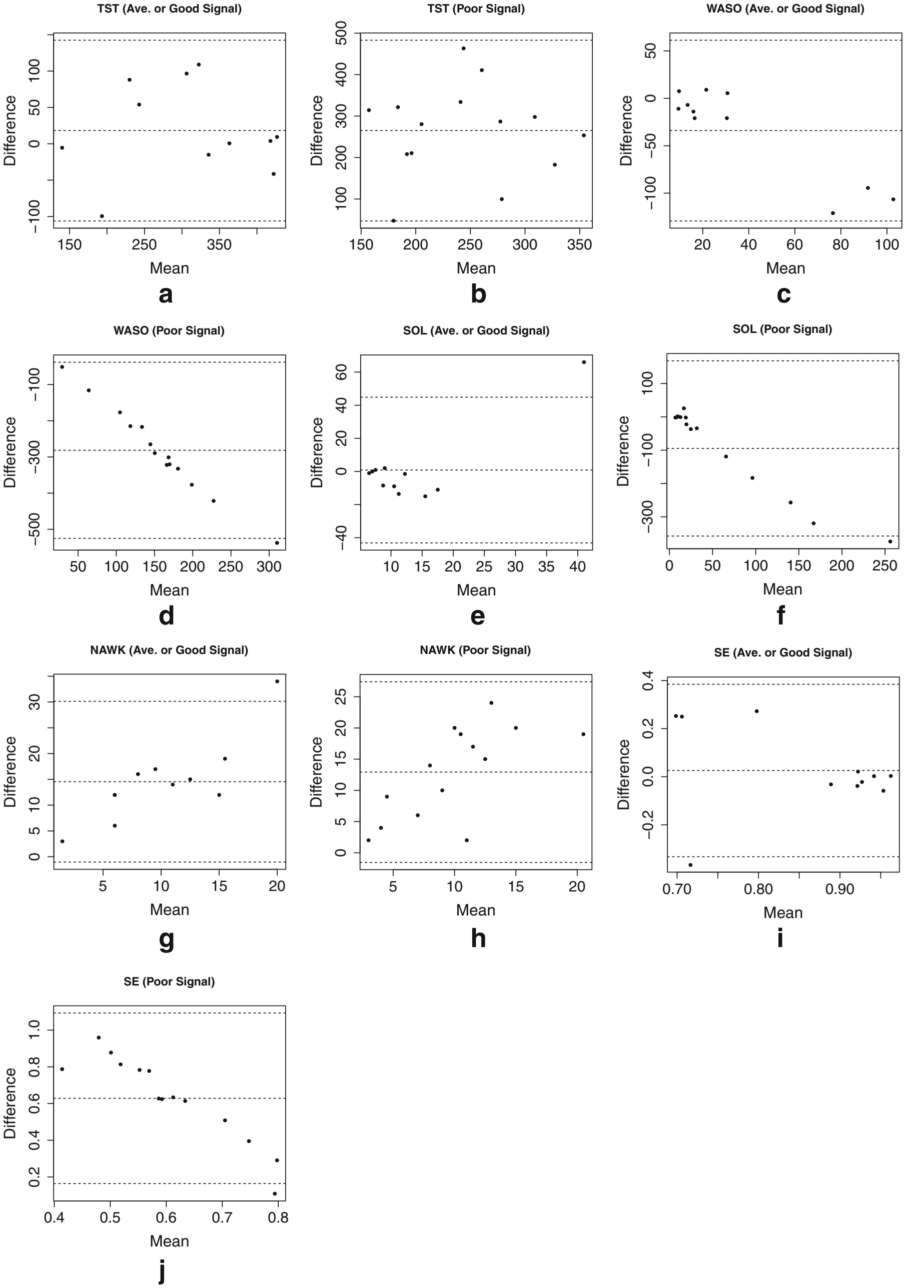


Fig. 11 Bland-Altman plots for Neuroon versus medical device on TST, WASO, SOL, NAWK, and SE when **a**, **c**, **e**, **g**, **i** signal quality was above average and **b**, **d**, **f**, **h**, **j** signal quality was poor

The impact of signal quality also manifested in the Pearson correlation coefficients shown in Table 6. Very strong correlation on TST, and strong correlations on WASO and deep sleep between Neuroon and the medical device were found when the signal quality was average or good. In comparison, strong correlation on WASO and moderate

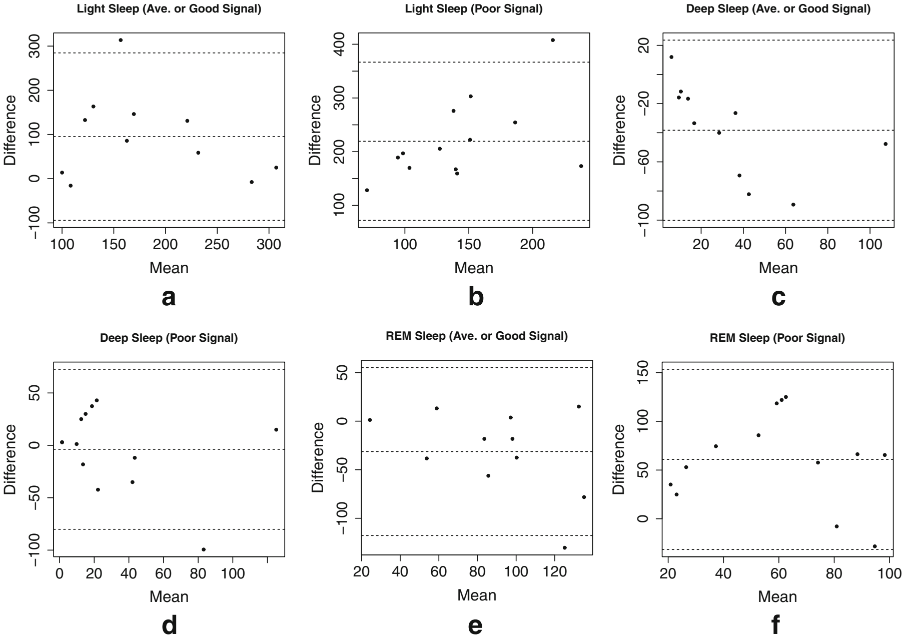


Fig. 12 Bland-Altman plots for Neuroon versus medical device on light sleep, deep sleep, and REM sleep when **a**, **c**, **e** signal quality was above average and **b**, **d**, **f** signal quality was poor

correlations on deep sleep stage and SOL were identified when the signal quality was poor. Interestingly, Neuroon was strongly correlated to the medical device on WASO in both subset ($r = .69$ when signal was above average and $r = .64$ when signal was poor, both statistically significant), but the correlation was cancelled out in the whole dataset ($r = .13$), which suggest that the relationship between the two devices may be non-linear on WASO so that such relationship was not captured by the Pearson correlation coefficient in the whole dataset.

5 Discussions

This study has shown a quantitative comparison between the latest consumer sleep tracking devices and a medical device in measuring overall sleep parameters and sleep

Table 6 Pearson correlation coefficients between Neuroon and medical device when Neuroon signal quality varied

	TST	WASO	Light sleep	Deep sleep	REM sleep	NAWK	SOL	SE
Good or average signal	**.81[a]****	**.69***	.38	**.66***	.46	.48	− .24	.15
Poor signal	.08	**.64***	.28	**.56***	.13	.31	**.53***	− .03
All	.08	.13	.15	**.50***	.05	.37	.34	− .20

[a] Bold indicates a significant correlation (*$p \le 0.05$, **$p \le 0.01$, ***$p \le 0.001$, ****$p < 0.000$)

structure. The impact of signal quality on the wearable EEG was also investigated. We will now discuss these results within the landscape of previous studies and highlight opportunities for future research.

5.1 Strength of Consumer Devices

Wearable activity wristbands were originally purported to measure and improve physical activity and were later expanded to also measure sleep duration and quality. Previous models of Fitbit and other activity trackers such as Jawbone require users to manually switch the device in and out of sleep tracking mode, which was one of the main reasons for missing data and measurement errors when these devices were used for sleep tracking [4, 38, 64, 65]. The feature of automatic sleep detection is no doubt an attractive and useful function of Fitbit Charge 2. This study showed that Fitbit Charge 2 was able to automatically detect the onset and offset of sleep with reasonable accuracy, though the accuracy could be hampered by the characteristics of sleep (e.g. long sleep onset latency) or by user behaviors (e.g. reading in bed).

Previous validation studies on consumer sleep tracking devices found that activity wristbands tend to overestimate time asleep and sleep efficiency due to the misclassification of inactive awake as sleep [4, 5, 32, 33]. Nevertheless, with better hardware and software, the latest models have overcome this limitation. Notably, our analysis showed that both Fitbit Charge 2 and Neuroon had good agreement to the medical device on TST (total sleep time) and SE (sleep efficiency). In addition, both consumer devices demonstrated the ability to measure other sleep parameters. Fitbit achieved good agreement to the medical device in measuring NAWK. It is worth noting that the NAWK (number of awakenings) of Fitbit counted in both long awakenings and brief restlessness. Very strong correlation was found on TST, and moderate correlation was found on NAWK and SE between Fitbit Charge 2 and the medical device. As expected, wearable EEG sleep trackers have the potential for accurately measuring more sleep parameters. Using brainwave data, Neuroon produced comparable results to the medical device on sleep onset latency and total awake time when signal quality was good.

Different from previous models of wristbands that dominantly rely on movement data, the latest devices build on new algorithms that use multimodal data for sleep scoring. Fitbit Charge 2 uses both movement data and heart rate signals [73], while Neuroon nurtures data from imbedded EEG, EOG, ECG, and accelerometer [66]. These results revealed that the multi-modal approach has significantly improved the accuracy of consumer sleep tracking devices in measuring overall sleep parameters.

5.2 Weakness of Consumer Devices

Although both consumer devices demonstrated the ability to accurately measure several sleep parameters especially total sleep duration and sleep efficiency, measuring sleep structure including light, deep, and REM sleep remains to be challenging for both devices. There has been no study validating the ability of Fitbit to measure sleep stages, as this function has only become available recently. However, another activity wristband Jawbone has been routinely distinguishing light sleep (corresponding to sleep Stage N1 and N2) and sound sleep (corresponding to sleep Stage N3 and REM). A few validation studies on Jawbone found that light and sound sleep measured by Jawbone did not agree with the light and deep sleep measured by medical devices [5]. Along the same line, our study found that both the latest consumer sleep trackers underestimated light sleep while overestimated deep sleep, regardless of their differences in hardware and software. As for REM, Fitbit Charge 2 overestimated REM by 11.6 min on average compared to the medical device, whereas Neuroon with good signal quality underestimated REM by 11.6 min on average. These results suggested that a combination of heart rate signal and movement data may be insufficient for sleep stage analysis in Fitbit Charge 2. On the other hand, Neuroon may need better sleep scoring algorithms to map brainwave signals to sleep stages. Better understanding on the patterns of measurement errors is required to design new algorithms for future wearable sleep tracking devices.

In addition to the above-mentioned common weakness, each of the consumer devices has their own limitation in estimating several overall sleep parameters. Measurement bias of WASO remains to be a problem for Fitbit. Previous validation studies showed that activity wristbands tend to underestimate awake due to the difficulty in detecting inactive awake [4, 5, 32, 33]. However, our results showed that Fitbit Charge 2 overestimated WASO by 24.5 min on average compared to the medical device. A possible reason is that the new sleep scoring algorithm that incorporates heart rate data tends to misclassify sleep as awake. Moreover, Fitbit Charge 2 cannot effectively measure SOL. The average SOL measured by Fitbit Charge 2 was 3.5 min, which was 11.1 min shorter than that measured by the medical device. Unlike previous models such as Fitbit Flex that requires manual switch into sleep mode, Fitbit Charge 2 highlights the feature of automatic detection on sleep onset. Whereas this feature may significantly improve usability of the device and help reduce missing data [4, 64, 65], it also became impossible to capture the time stamp of “lights off” when a user is ready to sleep. Without such information, it is theoretically not possible to calculate SOL, as the definition of SOL is the time between lights off and the first epoch of Stage N2. This limitation of Fitbit Charge 2 is due to the trade-off between usability and accuracy.

On the other hand, the fundamental reason for Neuroon’s problem in measuring nocturnal awakenings can be traced back to how this parameter was defined. According to our observation, Neuroon only captured long awakenings and ignored brief ones, yielding an underestimation of NAWK by 15 times on average in comparison to the medical device. Although NAWK captured by Neuroon may agree to users' subjective experience of sleep, inconsistency in the definition of the underlying phenomenon being measured led to disagreement between Neuroon and the medical device. Given the discrepancy between subjective and objective sleep quality [22–26], Fitbit’s strategy of distinguishing brief awakenings (restlessness) from long awakenings strikes a good balance between subjective experience and objective definition of awakenings.

5.3 Wristbands or Wearable EEG?

The manufacturers of wearable EEG sleep trackers such as Neuroon and Sleep Shepherd claimed that their devices could measure sleep with higher accuracy in comparison to wearable wristbands, because wearable EEG devices are based on similar mechanism as clinical sleep monitors. Indeed, our study revealed that Neuroon with good signal quality produced comparable measurements on total sleep time, sleep efficiency, total wake time, and sleep onset latency. Since these sleep parameters as well as their night-to-night variability are important indicators of sleep disorders including insomnia [39–42] and depression [67], Neuroon offers an opportunity for detecting potential sleep-related health problems in free-living conditions. However, poor signal quality could markedly deteriorate the performance of Neuroon and thus reduce the quality of measurement results. Our analysis showed that Neuroon significantly deviated from the medical device on all sleep parameters when the signal quality was poor, though moderate correlation was found on deep sleep between the two devices. Even worse, getting good signal from Neuroon was not easy, as only 7 out of the 25 participants had good signal in our study. Many participants complained that Neuroon eye mask moved off their face during night. Moreover, our study found greater discrepancy between Neuroon and the medical device in nights with more disrupted sleep. Similar trends were observed in previous validation studies on both clinical (e.g. single-channel EEG and actigraphy) and consumer (e.g. Fitbit Charge HR) sleep tracking devices [68–71]. Given Neuroon's sensitivity to poor signal quality and to disrupted sleep patterns, it is improper to solely rely on data from Neuroon for diagnosing sleep disorders.

In comparison, the limitations of Fitbit in measuring total awake time and sleep onset latency greatly reduced its potential of being used for the purpose of diagnosing sleep disorders. However, Fitbit achieved consistent and satisfactory performance on measuring total sleep time, sleep efficiency, and number of awakenings. Factoring in wearability and usability, Fitbit could be a good tool for individual and non-clinical use.

By clarifying the strength and weakness of consumer sleep tracking technologies, this study can help both individual end-users and researchers to select devices that best suit their needs. However, it is worth noting that this study has several limitations. First, the cohort of this study only consists of young healthy adults. The findings from this study thus may not be generalized to clinical and elderly populations. Second, we did not examine the epoch-by-epoch analysis of sensitivity, specificity, and AUC [72] because Neuroon only generated aggregated sleep metrics and epoch-by-epoch data was not available. In addition, it remains unknown as to whether the consumer devices could correctly classify individual sleep stages in each epoch. Future studies are needed to (1) investigate the validity of the latest consumer devices for measuring the sleep of clinical or elderly populations, (2) clarify the classification performance of the devices through epoch-by-epoch comparison to medical devices, and (3) improve the ability of consumer sleep trackers to detect sleep stage transitions.

6 Conclusions

Despite the popularity and convenience of consumer sleep trackers, the validity of these tools has not yet been thoroughly examined, especially for the devices that have just entered the market recently and especially for their ability to measure sleep stages. This study investigated the validity of two latest consumer sleep tracking devices, an activity wristband Fitbit Charge 2 and a wearable EEG eye mask Neuroon, in comparison to a medical portable sleep monitor. Results from this study advances the understanding on what consumer sleep tracking device can and cannot achieve. Our analysis found good agreement between consumer sleep trackers and the medical device in measuring total sleep duration and sleep efficiency. In addition, Fitbit Charge 2 agreed well to the medical device on the number of awakenings, while Neuroon with good signal quality produced comparable measurements on total awake time and sleep onset latency. However, classifying sleep stages remains challenging for both devices. Both devices underestimated light sleep and overestimated deep sleep. Poor agreement was found on REM as well, which was overestimated by Fitbit but underestimated by Neuroon. As expected, Neuroon was able to accurately measure more sleep parameters than Fitbit. Since some of these parameters are important indicators of sleep disorders, Neuroon has the potential to be used for sleep disorder diagnosis in free living conditions. However, the performance of Neuroon may be significantly deteriorated by poor signal quality and disrupted sleep. Counting in other factors such as wearability and usability, Fitbit Charge 2 could be a good option for general-purpose sleep monitoring and tracking in home. In the end, we highlighted three directions for future research: (1) to investigate the validity of the latest consumer devices for measuring the sleep of clinical or elderly populations, (2) to clarify the classification performance of the devices through epoch-by-epoch comparison to medical devices, and (3) to improve the ability of consumer sleep trackers to detect sleep stage transitions.

Acknowledgements This work was supported by JSPS KAKENHI Grant-in-Aid for Research Activity Start-up (Grant Number 16H07469) and an internal research grant from the National Institute of Advanced Industrial Science and Technology of Japan. The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper.

Compliance with Ethical Standards

Conflict of Interest The authors certify that there is no conflict of interest involved in this manuscript and this study. The opinions expressed in this paper are those of the authors and do not represent the views of the second author’s company.

Ethical Approval Ethics approval was obtained from the Ethic Committee of the University of Tokyo (Ethics ID: KE16–83). All participants provided informed consent.

References

1. Liang Z, Chapa-Martell MA (2015) Framing self-quantification for individual-level preventive health care. In: Proceedings of the International Conference on Health Informatics, pp 336–343
2. Dittmar A, Axisa F, Delhomme G, Gehin C (2004) New concepts and technologies in home care and ambulatory monitoring. Stud Health Technol Inform 108:9–35
3. Mantua J, Gravel N, Spencer R (2016) Reliability of sleep measures from four personal health monitoring device compared to research-based actigraphy and polysomnography. Sensors 16:646
4. Meltzer LJ, Hiruma LS, Avis K, Montgomery-Downs H, Valentin J (2015) Comparison of a commercial accelerometer with polysomnography and actigraphy in children and adolescents. Sleep 38(8):1323–1330. https://doi.org/10.5665/sleep.4918
5. De Zambotti M, Baker FC, Colrain IM (2015) Validation of sleep-tracking technology compared with polysomnography in adolescents. Sleep 38(9):1461–1468. https://doi.org/10.5665/sleep.4990
6. De Zambotti M, Godino JG, Baker FC et al (2016) The boom in wearable technology: cause for alarm or just what is needed to better understand sleep? Sleep 39(9):1761–1762. https://doi.org/10.5665/sleep.6108

7. Buysse DJ (2014) Sleep health: can we define it? Does it matter? Sleep 37(1):9–17. https://doi.org/10.5665/sleep.3298
8. Coates TJ, Killen JD, George J, Marchini E, Silverman S, Thoresen C (1982) Estimating sleep parameters: a multitrait-multimethod analysis. J Consult Clin Psychol 50(3):345–352. https://doi.org/10.1037/0022-006X.50.3.345
9. Carskadon MA, Dement WC (2015) Normal human sleep: an overview. In: Kryger MH, Roth T, Dement WC (eds) Principle and practice of sleep medicine, 4th edn. Elsevier Saunders, Philadelphia, pp 13–23
10. Hall M (2010) Behavioral medicine and sleep: concept, measures and methods. In: Steptoe A (ed) Handbook of behavioral medicine: methods and applications. Springer, New York, pp 749–765. https://doi.org/10.1007/978-0-387-09488-5_49
11. Buysse DJ, Reynolds CF, Monk TH, Berman SR, Kupfer DJ (1989) The Pittsburgh sleep quality index: a new instrument for psychiatric practice and research. Psychiatry Res 28(2):193–213. https://doi.org/10.1016/0165-1781(89)90047-4
12. Carney CE, Buysse DJ, Ancoli-Israel S, Edinger JD, Krystal AD, Lichstein KL, Morin CM (2012) The consensus sleep diary: standardizing prospective sleep self-monitoring. Sleep 35(2):287–302. https://doi.org/10.5665/sleep.1642
13. Hublin C, Partinen M, Koskenvuo M, Kaprio J (2007) Sleep and mortality: a population-based 22-year follow-up study. Sleep 30(10):1245–1253. https://doi.org/10.1093/sleep/30.10.1245
14. Kushida CA, Littner MR, Morgenthaler T, Alessi CA, Bailey D, Coleman J Jr, Friedman L, Hirshkowitz M, Kapen S, Kramer M, Lee-Chiong T, Loube DL, Owens J, Pancer JP, Wise M (2005) Practice parameters for the indications for polysomnography and related procedures: an update for 2005. Sleep 28(4):499–519. https://doi.org/10.1093/sleep/28.4.499
15. Ohayon M, Wickwire EM, Hirshkowitz M, Albert SM, Avidan A, Daly FJ, Dauvilliers Y, Ferri R, Fung C, Gozal D, Hazen N, Krystal A, Lichstein K, Mallampalli M, Plazzi G, Rawding R, Scheer FA, Somers V, Vitiello MV (2017) National sleep foundation's sleep quality recommendations: first report. Sleep Health 3(1):6–19. https://doi.org/10.1016/j.sleh.2016.11.006
16. Hirshkowitz M (2004) Normal human sleep: an overview. Med Clin North Am 88:51–65
17. Keenan SA (1999) Normal human sleep. Respir Care Clin N Am 5:319–331
18. Sadeh A (2011) The role and validity of actigraphy in sleep medicine: an update. Sleep Med Rev 15(4):259–267. https://doi.org/10.1016/j.smrv.2010.10.001
19. Ancoli-Israel S, Cole R, Alessi C, Chambers M, Moorcroft W, Pollak C (2003) The role of actigraphy in the study of sleep and circadian rhythms. American Academy of sleep medicine review paper. Sleep 26(3):342–392. https://doi.org/10.1093/sleep/26.3.342
20. Littner M, Kushida CA, Anderson WM et al (2002) Practice parameters for the role of actigraphy in the study of sleep and circadian rhythms: an update for 2002. Sleep 26(3):337–341
21. Vitiello MV, Larsen LH, Drolet G et al (2002) Gender differences in subjective-objective sleep relationships in non-complaining healthy older adults. Sleep 25S:A61
22. Van Ravesteyn LM, Tulen JH, Kamperman AM et al (2014) Perceived sleep quality is worse than objective parameters of sleep in pregnant women with a mental disorder. J Clin Sleep Med 10(10):1137–1141. https://doi.org/10.5664/jcsm.4118
23. Lund HG, Rybarczyk BD, Perrin PB, Leszczyszyn D, Stepanski E (2013) The discrepancy between subjective and objective measures of sleep in older adults receiving CBT for comorbid insomnia. J Clin Psychol 69(10):1108–1120. https://doi.org/10.1002/jclp.21938
24. Most EIS, Aboudan S, Scheltens P, van Someren EJW (2012) Discrepancy between subjective and objective sleep disturbances in early- and moderate-stage Alzheimer disease. Am J Geriatr Psychiatry 20(6):460–467. https://doi.org/10.1097/JGP.0b013e318252e3ff
25. O'Donnel D, Silva EJ, Munch M et al (2009) Comparison of subjective and objective assessments of sleep in healthy older subjectis without sleep complaints. J Sleep Res 18(2):254–263. https://doi.org/10.1111/j.1365-2869.2008.00719.x
26. Tsuchiyama K, Nagayama H, Kudo K, Kojima K, Yamada K (2003) Discrepancy between subjective and objective sleep in patients with depression. Psychiatry Clin Neurosci 57(3):259–264. https://doi.org/10.1046/j.1440-1819.2003.01114.x
27. Kolla BP, Mansukhani S, Mansukhani MP (2016) Consumer sleep tracking devices: a review of mechanisms, validity and utility. Expert Rev Med Devices 13(5):497–506. https://doi.org/10.1586/17434440.2016.1171708
28. Shelgikar AV, Anderson PF, Stephens MR (2016) Sleep tracking, wearable technology, and opportunities for research and clinical care. Chest 150(3):732–743. https://doi.org/10.1016/j.chest.2016.04.016
29. Ong AA, Gillespie MB (2016) Overview of smartphone applications for sleep analysis. World J Otorhinolaryngol Head Neck Surg 2(1):45–49. https://doi.org/10.1016/j.wjorl.2016.02.001

30. Ko PT, Kientz JA, Choe EK, Kay M, Landis CA, Watson NF (2015) Consumer sleep technologies: a review of the landscape. J Clin Sleep Med 11(12):1455–1461. https://doi.org/10.5664/jcsm.5288
31. Kelly JM, Strecker RE, Bianchi MT (2012) Recent developments in home sleep-monitoring devices. International Scholarly Research Network Neurology, Article ID 768794, 10 pages
32. Montgomery-Downs HE, Insana SP, Bond JA (2012) Movement toward a novel activity monitoring device. Sleep Breath 6(3):913–917
33. De Zambotti M, Claudatos S, Inkelis S et al (2015) Evaluation of a consumer fitness-tracking device to access sleep in adults. Chronobiol Int 32(7):1024–1028. https://doi.org/10.3109/07420528.2015.1054395
34. Bhat S, Ferraris A, Gupta D, Mozafarian M, DeBari VA, Gushway-Henry N, Gowda SP, Polos PG, Rubinstein M, Seidu H, Chokroverty S (2015) Is there a clinical role for smartphone sleep apps? Comparison of sleep cycle detection by a smartphone application to polysomnography. J Clin Sleep Med 11(7):709–715. https://doi.org/10.5664/jcsm.4840
35. Shirazi AS, Clawson J, Hassanpour Y, Tourian MJ et al (2013) Already up? Using mobile phones to track & share sleep behavior. Int J Hum Comput Stud 71(9):878–888. https://doi.org/10.1016/j.ijhcs.2013.03.001
36. Behar J, Roebuck A, Shahid M, Daly J, Hallack A, Palmius N, Stradling J, Clifford GD (2015) SleepAp: an automated obstructive sleep apnoea scorning application for smartphones. IEEE J Biomed Health Inform 19(1):325–331. https://doi.org/10.1109/JBHI.2014.2307913
37. Liang Z, Ploderer B (2016) Sleep tracking in the real world: a qualitative study into barriers for improving sleep. In: Proceedings of OZCHI 2016: 537–541
38. Liang Z, Ploderer B, Chapa-Martell MA (2017) Is Fitbit fit for sleep-tracking? Sources of measurement errors and proposed countermeasures. In: Proceedings of Pervasive Health 2017
39. Natale V, Leger D, Martoni M et al (2014) The role of actigraphy in the assessment of primary insomnia: a retrospective study. Sleep Med 15(1):111–115. https://doi.org/10.1016/j.sleep.2013.08.792
40. Suh S, Nowakowski S, Bernert RA, Ong JC, Siebern AT, Dowdle CL, Manber R (2012) Clinical significance of night-to-night sleep variability in insomnia. Sleep Med 13(5):469–475. https://doi.org/10.1016/j.sleep.2011.10.034
41. Buysse DJ, Cheng Y, Germain A, Moul DE, Franzen PL, Fletcher M, Monk TH (2010) Night-to-night sleep variability in older adults with and without chronic insomnia. Sleep Med 11(1):56–64. https://doi.org/10.1016/j.sleep.2009.02.010
42. Natale V, Plazzi G, Martoni M (2009) Actigraphy in the assessment of insomnia: a quantitative approach. Sleep 32(6):767–771. https://doi.org/10.1093/sleep/32.6.767
43. Fitbit Charge 2: How do I track my sleep? https://help.fitbit.com/articles/en_US/Help_article/1314/?l=en_US&c=Topics%3ASleep&p=charge_2&fs=Search&pn=1#Whatisthedifference
44. Neuroon sleep analytics. https://neuroon.jp/features/sleep-analytics/
45. Yoshida M, Shinohara H, Kodama H (2015) Assessment of nocturnal sleep architecture by actigraphy and one-channel electroencephalography in early infancy. Early Hum Dev 91(9):519–526. https://doi.org/10.1016/j.earlhumdev.2015.06.005
46. Rosenberg RS, Van Hout S (2013) The American Academy of Sleep Medicine inter-scorer reliability program: sleep stage scoring. J Clin Sleep Med 9(1):81–87. https://doi.org/10.5664/jcsm.2350
47. Liang Z, Ploderer B, Liu W, Nagata Y, Bailey J, Kulik L, Li X (2016) SleepExplorer: a visualization tool to make sense of correlations between personal sleep data and contextual factors. Pers Ubiquit Comput 20(6):985–1000. https://doi.org/10.1007/s00779-016-0960-6
48. Iber C, Ancoli-Israel S, Chesson A et al (2007) The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications, 1st edn. American Academy of Sleep Medicine, Westchester
49. McCall C, McCall WV (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
50. Ahmadi N, Shapiro GK, Chung SA, Shapiro CM (2009) Clinical diagnosis of sleep apnea based on single night of polysomnography vs. two nights of polysomnography. Sleep Breath 13(3):221–226. https://doi.org/10.1007/s11325-008-0234-2
51. Tworoger SS, Davis S, Vitiello MV, Lentz MJ, McTiernan A (2005) Factors associated with objective (actigraphic) and subjective sleep quality in young adult women. J Psychosom Res 59(1):11–19. https://doi.org/10.1016/j.jpsychores.2005.03.008
52. Evenson KR, Goto MM, Furberg RD (2015) Systematic review of the validity and reliability of consumer-wearable activity trackers. Int J Behav Nutr Phys Act 12(1):159. https://doi.org/10.1186/s12966-015-0314-1
53. Higgins PA, Straub AJ (2006) Understanding the error of our ways: mapping the concepts of validity and reliability. Nurs Outlook 54(1):23–29. https://doi.org/10.1016/j.outlook.2004.12.004

54. Benjamini Y (1988) Opening the box of a boxplot. Am Stat 42(4):257–262
55. Wilcoxon F (1945) Individual comparisons by ranking methods. Biom Bull 1(6):80–83. https://doi.org/10.2307/3001968
56. Fay MP, Proschan MA (2010) Wilcoson-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. Stat Surv 4(0):1–39. https://doi.org/10.1214/09-SS051
57. Bradley JV (1968) Distribution-free statistical tests. Prentice-Hall
58. Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1(8476):307–310
59. Meltzer L, Walsh C, Traylor J, Westin A (2012) Direct comparison of two new actigraphs and polysomnography in children and adolescents. Sleep 35(1):159–166. https://doi.org/10.5665/sleep.1608
60. Werner H, Molinari L, Guyer C, Jenni OG (2008) Agreement rates between actigraphy, diary, and questionnaire for children's sleep patterns. Arch Pediatr Adolesc Med 162(4):350–358. https://doi.org/10.1001/archpedi.162.4.350
61. Rosenberger ME, Buman MP, Haskell WL et al (2015) 24h or sleep, sedentary behavior, and physical activity with nine wearable devices. Med Sci Sports Exerc 48(3):457–465
62. Kappenman ES, Luck SJ (2010) The effects of electrode impedance on data quality and statistical significance in ERP recordings. Psychophysiology 47(5):888–904. https://doi.org/10.1111/j.1469-8986.2010.01009.x
63. Duun-Henriksen J, Kjaer TW, Looney D, et al. (2015) EEG signal quality of a subcutaneous recording system compared to standard surface electrodes. J Sensors 2015: Article 341208, 9 pages
64. Baroni A, Bruzzese JM, Di Bartolo CA, Shatkin JP (2016) Fitbit Flex: an unreliable device for longitudinal sleep measures in a non-clinical population. Sleep Breath 20(2):853–854. https://doi.org/10.1007/s11325-015-1271-2
65. Ferguson T, Rowlands AV, Olds T, Maher C (2015) The validity of consumer-level, activity monitors in healthy adults worn in free-living conditions: a cross-sectional study. Int J Behav Nutr Phys Act 12(1):42. https://doi.org/10.1186/s12966-015-0201-9
66. Neuroon Open product overview. http://community.neuroonopen.com/product
67. Hein M, Lanquart JP, Loas G, Hubain P, Linkowski P (2017) Similar polysomnographic pattern in primary insomnia and major depression with objective insomnia: a sign of common pathophysiology? BMC Psychiatry 17(1):273. https://doi.org/10.1186/s12888-017-1438-4
68. Lucey BP, McLeland JS, Toedebusch CD et al (2016) Comparison of a single-channel EEG sleep study to polysomnography. J Sleep Res 25(6):625–635. https://doi.org/10.1111/jsr.12417
69. De Zambotti M, Baker FC, Willoughby AR et al (2016) Measures of sleep and cardiac functioning during sleep using a multi-sensory commercially-available wristband in adolescents: wearable technology to measure sleep and cardiac functioning. Physiol Behav 158:143–149
70. Taibi DM, Landis CA, Vitiello MV (2013) Concordance of polysomnographic and actigraphic measurement of sleep and wake in older women with insomnia. J Clin Sleep Med 9(3):217–225. https://doi.org/10.5664/jcsm.2482
71. Blackwell T, Redline S, Ancoli-Isreal S et al (2008) Comparison of sleep parameters from actigraphy and polysomnography in older women: the SOF study. Sleep 31(2):283–291. https://doi.org/10.1093/sleep/31.2.283
72. Cellini N, Buman MP, McDevitt EA et al (2013) Direct comparison of two actigrapy devices with polysomnographically recorded naps in healthy young adults. Chronobiol Int 30(5):691–698. https://doi.org/10.3109/07420528.2013.782312
73. Kosecki D (2017) Your heart rate is the key to smarter sleep stages. Here's why. Fitbit News, https://blog.fitbit.com/heart-rate-during-sleep-stages/