Combining Numerical and Visual Approaches in Validating Sleep Data Quality of Consumer Wearable Wristbands

Zilu Liang
Graduate School of Engineering, The University of Tokyo
School of Engineering, Kyoto University of Advanced Science
Tokyo/Kyoto, Japan
z.liang@cnl.t.u-tokyo.ac.jp

Mario Alberto Chapa Martell
Advanced Technology Division
CAC Corporation
Tokyo, Japan
mchapam0300@gmail.com

***Abstract*—The recent rise of the Quantified Self movement has witnessed a significant increase in the adoption of consumer wearable wristbands for sleep tracking. Nevertheless, data quality of these devices has been a main concern. This study aimed to validate a most popular consumer wristband, i.e. Fitbit Charge 2™, against medical devices. We proposed a new validation approach that combines numerical technique with visual aid for epoch-by-epoch comparison on sleep stages. We found that Fitbit Charge 2™ had low accuracy in detecting wake and reasonable accuracy in detecting light, deep, and REM sleep stages. The visual aid of scatter plots showed that Fitbit was more accurate in detecting deep sleep stage in the first half of a night and more accurate in detecting REM sleep stage in the second half of a night. Our results indicate that consumer wearable wristbands are not able to produce high quality data of sleep stages in ecological settings. Future studies should consider the effect of time on device accuracy and may resort to segmented modelling techniques to improve data quality.**

Keywords— wearable wristbands, data quality, sleep, validation, Fitbit, data visualization

I. INTRODUCTION

The flourish of the consumer wearable market and the rise of the Quantified Self movement have led to a sharp increase in the number of adopters who use consumer wearable devices to track various physiological and psychological metrics [1-3]. In the meanwhile, these devices are also becoming popular among researchers because they enable longitudinal and cost-efficient data collection in ecological settings [4-6]. Nevertheless, the data quality of these consumer devices has been a main concern for end users who aim to gain insights from their data and for researchers who plan to use these devices in scientific studies [5, 7-10]. To this end, data quality is key to the sustained and large-scale adoption of these technologies [9].

In this paper, we aim to validate a most popular and recent consumer wristband, i.e. Fitbit Charge 2™, in measuring sleep stages under free-living conditions. Many validation studies have endeavored to establish the discrepancy between consumer wristbands and clinical sleep monitors based on epoch-by-epoch (EBE) comparison. However, the majority of these studies only investigated device validity in wake/sleep classification based on sensitivity and specificity [11-13]. There has been solely one study investigating device accuracy in classifying all four sleep stages [13], as the feature of detecting light, deep and REM sleep has only become available very recently in the latest Fitbit models. Nevertheless, this study was conducted in a sleep laboratory. Previous validations studies on clinical actigraphy suggest that device accuracy under free-living conditions may deviate from that in laboratory settings. Hence, there is still need to examine the performance of Fitbit in classifying sleep stages in home settings.

In this study, we followed the common practice in the field and conducted epoch-wise comparison between Fitbit data and medical data. We calculated a confusion matrix to demonstrate the capability of Fitbit in classifying wake, light sleep, deep sleep, and REM sleep. Assuming that device validity relies on the characteristics of the underlying phenomenon being measured, we were also interested in understanding the effect of time on data quality throughout the course of one night. Human sleep demonstrates temporal patterns. The amount of deep sleep epochs decreases throughout night while the number of REM sleep epochs increases. Nevertheless, previous validation studies generally consider device accuracy as a static property. Since temporal information is lost in these validation studies, it remains unknown as to whether and how data quality depends on the time of the measuring. We approached the problem from an unconventional perspective, as traditional validation techniques such as Bland-Altman plots or t-test do not server our purpose of exploring the temporal pattern of information quality in sleep tracking [14, 15]. We drew inspiration from the latest advance in data visualization [16, 17] and used visual aid to uncover the temporal patterns of data quality. By converting the sleep hypnograms of a cohort into color spectrums and then segmenting the spectrum into sleep cycles, we obtained powerful visual cues to probe the accuracy of Fitbit in different sleep cycles.

The contribution of this study is two-fold. First, we proposed a new validation approach that uses data visualization to complement the numerical approach that has been routinely used in validation studies. This new approach allows us to observe the temporal patterns of device accuracy throughout the course of one night. Second, we offer new insights into the validity of Fitbit Charge 2™ compared to medical devices under free-living conditions. The rest of the paper is organized as follows. Section II provides a summary of related work on

XXX-X-XXXX-XXXX-X/XX/$XX.00©20XX IEEE
978-1-5386-9151-9/19/$31.00 ©2019 IEEE

sleep analysis in clinical settings and previous validation studies of consumer sleep tracking devices. Section III and IV present the proposed methodology and the corresponding results. In Section V, we interpret the numerical and visual results within the landscape of consumer sleep tracking devices. The whole paper is closed in the conclusion.

II. Related Work

A. Sleep Analysis in Clinical Settings

Human sleep can be measured both subjectively and objectively. Subjective sleep quality is usually measured using established questionnaires such as the Pittsburgh Sleep Quality Index (PSQI) [18] and Sleep Disorder Questionnaire (SDQ) [19]. Objectively sleep quality can be measured by polysomnography (PSG), portable EEG, or actigraphy. PSG has been considered the gold standard in measuring human sleep. A PSG test measures multiple channels of biosignals such as cortical EEG, muscle tone (EMG), and eye movement (EOG). These data need to be analyzed by a trained technician following well-defined protocol [20]. First, biosignals are divided into short intervals called epochs. The common practice is 30 seconds [20], though it is possible to choose other length depending on the purpose of the analysis. Second, the technician visually inspect all signals to infer sleep stage epoch by epoch. Third, the epoch-wise results are summed up to produce aggregated outputs including total sleep time (TST), sleep onset latency (SOL), wake after sleep onset (WASO), the number of awakenings (NAWK), sleep efficiency (SE), and the ratio of each sleep stage [21]. It is worth noting that sleep scoring involves certain degree of subjectivity and the average inter-scorer agreement is approximately 82.6% [22].

Despite of its accuracy, PSG requires the use of bulky devices and constant technical support. The test is also expensive and uncomfortable for patients. Hence alternative devices have been developed for measuring sleep in clinical settings. The most notable ones are portable EEG and actigraphy, which have been widely used in sleep studies and tests in daily life settings. A portable EEG device consists of a cubed device body and several gel-type electrodes to be attached to the head of a user. These devices are simplified version of the EEG device in PSG in that they usually have fewer channels. Nevertheless, these devices are much less intrusive and requires no special setups. In comparison, actigraphy is a wristband worn on the non-dominant arm of a user. While EEG relies on the measurement of brainwaves, actigraphy infers sleep from movement based on data collected by an embedded accelerometer and has been widely used to capture sleep patterns spanning over multiple days in ecological settings [23, 24]. Several studies have validated that portable EEG and actigraphy are reasonably accurate compared to PSG [23, 25-28].

B. Validation of Consumer Sleep Tracking Devices

Many studies have examined how users interact with consumer sleep tracking devices (including wearable wristbands and wearable EEG) and how they interpret sleep data from these devices [4, 5, 9, 29-32]. Fitbit devices have also been used in many scientific studies where the measuring of sleep was not the main focus. These studies found that data quality is a main issue that impede individual users from gaining insights into their sleep patterns [7-9, 31] and prevent researchers from drawing rigid conclusions from their studies [4, 5, 10].

Given that many consumer devices are not validated [33, 34], there is increasing interest in research community to understand the validity of these devices. Table I summarizes the main findings in previous validation studies of Fitbit devices. These studies all relied on numerical approaches, based on either aggregated comparison or epoch-by-epoch comparison between consumer devices and medical devices. These studies indicate that the accuracy of Fitbit devices has been greatly improved since its first model. Validations on aggregated sleep metrics (i.e. TST, WASO, SOL, SE) have shown that the latest model has good accuracy on NAWK and SE, but not on other metrics [6, 11, 35, 36]. Epoch-by-epoch validations demonstrate that Fitbit devices generally have high accuracy in detecting sleep epochs but low accuracy in detecting wake epochs [6, 13, 35-37]. This characteristic is consistent with clinical actigraphy [23, 38-42].

TABLE I. Summary of Previous Validation Studies on Fitbit

Device Model	Aggregated Validation	Epoch-by-Epoch Validation
Fitbit Flex™	Overestimation of TST by 46min and SE by 8.1%. Underestimation of WASO by 44min [35].	High sensitivity and accuracy above 0.80, with low specificity lower than 0.40 for both healthy people and people with chronic disorders [35, 37].
Fitbit Ultra™	Overestimation of TST by 41 min and SE by 8%. Underestimation of WASO by 32min [36].	Good overall accuracy of 0.84 and sensitivity of 0.86, but poor specificity of 0.52 [36].
Fitbit Charge HR™	Overestimation of TST by 8min and SE by 1.8%. Underestimation of WASO by 5.6min [6].	High overall accuracy 0.91, high sensitivity of 0.97, and low specificity of 0.42 based on 1min-epoch analysis in lab settings [6].
Fitbit Charge 2™	No significant difference on NAWK and SE. Overestimation of WASO by 24.5min and deep sleep by 39.8min. Underestimation of TST by 12.3min, SOL by 11.1min, light sleep by 42.4min, and REM by 11.6min [11].	High sensitivity of 0.96 and reasonable specificity of 0.61 based on 30s-epoch analysis in lab settings. Accuracy for "light sleep", "deep sleep" and "REM" sleep were 0.81, 0.49, and 0.74 respectively [13].

III. Methodology

A. Numerical Approach

As illustrated in Fig.1, we conducted epoch-by-epoch comparison between Fitbit sleep data and medical data due to the epoch-wise nature in standard sleep scoring process in clinical settings [20]. A Fitbit Charge 2™ and a medical device need to be used concurrently to measure sleep from a cohort of participants. To enable direct comparison between the two devices, we first unified a mapping scheme between Fitbit

sleep data and medical data. The "light sleep" in Fitbit data corresponded to "stage N1" and "stage N2" in the medical data, "deep sleep" corresponded to "stage N3", "REM sleep" corresponded to "stage R", and "wake" corresponded to "stage W". Following the common practice in validation studies [13], a 4×4 confusion matrix was calculated for each subject and then averaged over the whole cohort. The confusion matrix represents a cross-tabular of 4 rows containing each sleep stage classified by the medical device versus 4 column containing the corresponding sleep stage classified by Fitbit Charge 2™. The diagonal values represent device accuracy in classifying each sleep stage. Sensitivity and specificity were also calculated to indicate the ability of Fitbit in correctly detecting sleep epochs and wake epochs respectively.

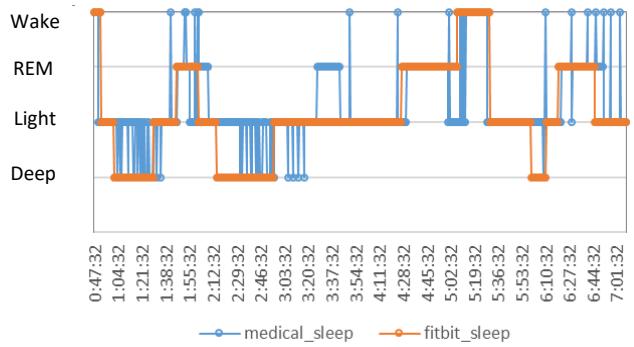

Fig. 1. Epch -by-epoch comparison between Fitbit data and medical data.

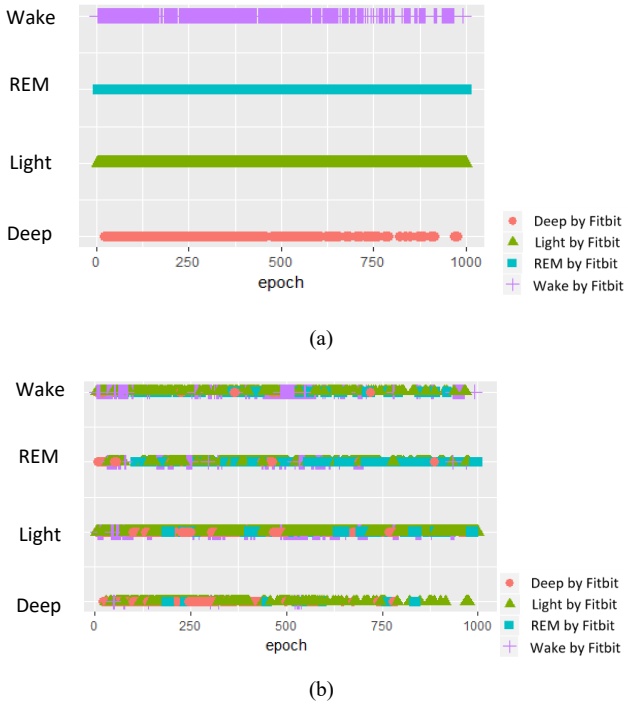

Fig. 2. Scatter plots showing agreement between Fitbit and medical device through a night. The *x*-coordinate indicates the elapse of time by epoch and the *y*-coordinate represents the ground truth measured by medical device. (a) An ideal scenario where Fitbit matches medical device perfectly; (b) a realistic scenario where Fitbit frequently misclassifies sleep stages.

B. *Visual Approach*

We used scatter plots to visually inspect the accuracy of a consumer device in measuring sleep stages throughout a night. As depicted in Fig. 2, the *x*-axis is the number of epochs elapsed, and the *y*-axis represents the ground truth measured by a medical device. The sleep stages detected by Fitbit Charge 2™ were color-coded. As is shown in Fig.2, red dots, green triangle, blue square and purple cross were used to indicate deep sleep, light sleep, REM sleep and wake respectively. Fig.2 (a) illustrates ideal situation where Fitbit agrees perfectly with the medical device. The color spectrums of wake, light sleep, deep sleep and REM sleep are in single color of purple, green, pink and blue. In contrast, Fig.2 (b) demonstrates a practical scenario where Fitbit misclassifies sleep stages. The misclassification of Fitbit produced colorful spectrums.

IV. Results

A. *Dataset Preparation*

We collected sleep data from 22 participants using a Fitbit Charge 2™ and a medical-grade portable 1-channel EEG device concurrently. Participants measured their sleep for one night using both devices in their own homes. We conducted data collection in daily life settings to ensure that our validation results reflect the performance of the wristband in natural living environment. We retrieved intra-day sleep data at the resolution of 1s from Fitbit through their partner API using a Chrome extension called Postman Interceptor. This data and the medical data were synchronized to make sure that the start time was aligned. Subsequently we developed a C# script program to aggregate the Fitbit data into 30s epochs (i.e. averaged every 30s). The medical data was first analyzed at 30s-epoch by a validated proprietary sleep scoring software. A sleep expert then visually inspected the results and added necessary modification following established sleep scoring standard [20]. The total sleep time of the cohort ranges from 4 hours (480 epochs) to 9.8 hours (1176 epochs). The whole dataset contains 18759 records in total. Each records contains the following three fields: epoch number, sleep stage measured by Fitbit, sleep stage measured by medical device.

B. *Sensitivity and Specificity*

TABLE II. Confusion Matrix

		Fitbit Charge 2™			
		Wake	Light	Deep	REM
Medical Device	Wake	0. 38 ± 0. 20[a]	0.48 ± 0. 19	0.06 ± 0.11	0.08 ± 0.07
	Light	0.03 ± 0. 02	0.69 ± 0. 08	0.22 ± 0. 70	0.06 ± 0. 06
	Deep	0.03 ± 0.11	0.30 ± 0. 24	0.64 ± 0. 30	0.03 ± 0. 08
	REM	0.05 ± 0.06	0.32 ± 0. 20	0.03 ± 0. 09	0.60 ± 0. 25

[a] The results are presented in the form of "average ± standard deviation".

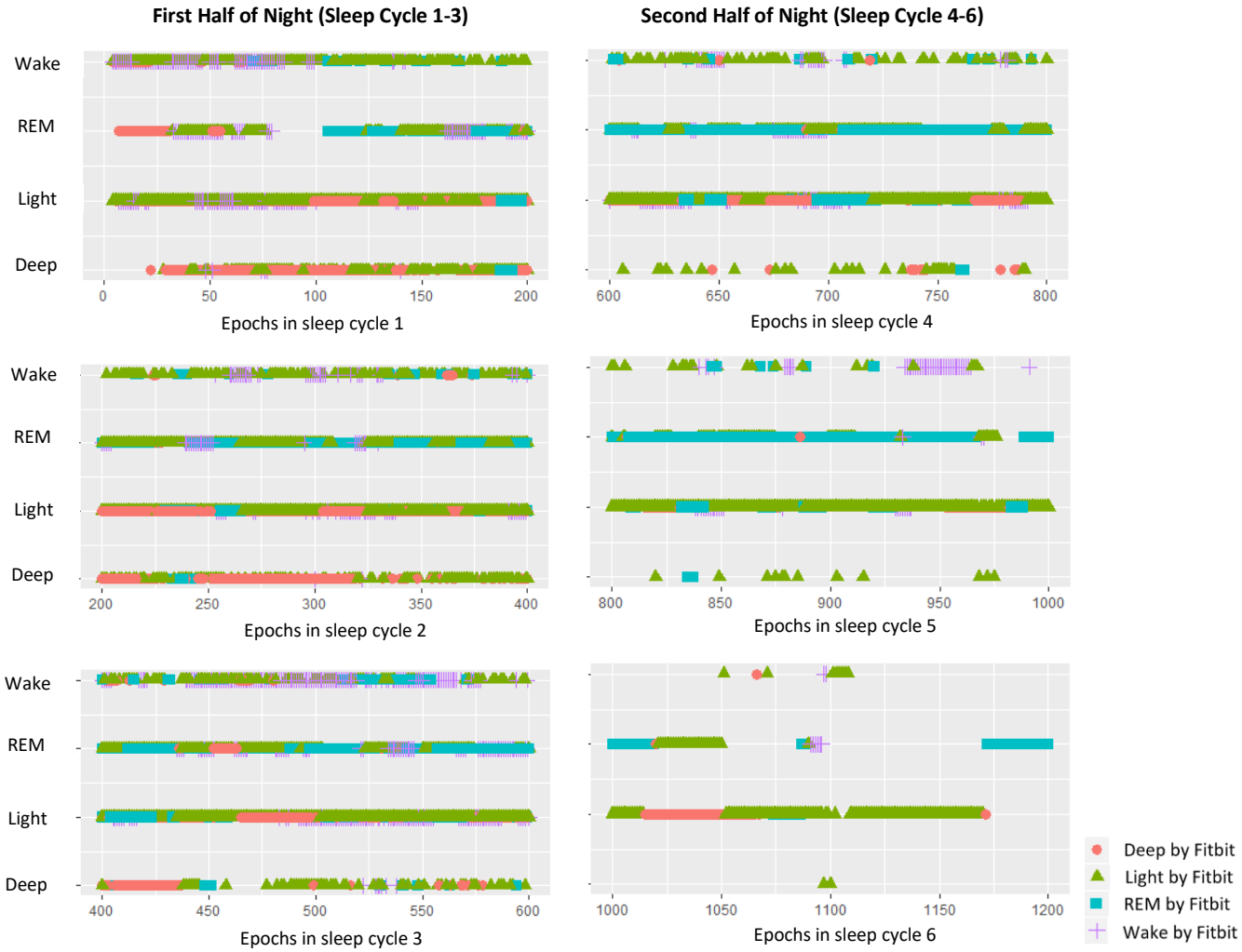

Fig. 3. Scatter plots of data during sleep cycle 1-6. The *x*-coordinate indicates the elapse of time by epoch and the *y*-coordinate represents the ground truth measured by a medical device. Left: first half of a night; Right: second half of a night.

We conducted epoch-wise comparison to calculate the 4×4 confusion matrix and the results are shown in Table II. The sensitivity of detecting sleep epochs and the specificity of detecting wake epochs are 0.96 and 0.38 respectively.

C. Temporal Pattern of Data Quality

We created scatter plots using a data visualization package named ggplot2 in R [16, 17, 43]. We divided the dataset into segments of 200 epochs, which approximates an average sleep cycle of 90 minutes [44, 45]. This allowed us to observe the plots at higher granularity. Consequently, we obtained six plots (Fig.3) demonstrating the agreement of Fitbit to medical device during sleep cycle 1~6.

V. Discussion

We have presented the numerical results and the visualization of comparisons between Fitbit Charge 2™ and the medical device. In what follows, we discuss the interpretation of these results and the limitations of this study.

A. Interpretation of Numerical Results

The confusion matrix in Table II shows that Fitbit Charge 2™ achieved mediocre accuracy in classifying all sleep stages. Compared to the medical device, 48% of wake epochs were misclassified as light sleep by Fitbit, which was even higher than the ratio of wake epochs corrected classified (38%). This echoes the low specificity of older Fitbit models identified in previous validation studies [6, 35, 37], and the result fits within the specificity range of 0.30-0.67 among studies validating standard actigraphy in health people [24]. Nevertheless, this result contradicts a recent validation study on Fitbit Charge 2™ by De Zambotti and colleagues [13]. The authors found higher specificity of Fitbit Charge 2™ (=0.61). The result was consistent between good sleepers and people with sleep problems. The discrepancy between our finding and this previous work could be explained by the differences in cohort characteristics and data collection protocol, and our findings complement this previous study by offering new insights into device accuracy in free-living conditions.

Fitbit achieved better performance in detecting sleep stages, with classification accuracy all above 0.60 for light, deep, and REM sleep epochs. Still, 22% of light sleep epochs were classified as deep sleep, and 30% of deep sleep epochs and 32% of REM sleep epochs were misclassified as light sleep. It seems to be easier to distinguish between deep and REM stages than to distinguish these two stages from light stage. These findings are largely consistent with the results in [13]. The function of Fitbit in detecting sleep stages has only become available very recently due to the integration of multi-stream biosignals from accelerometer and infrared heart rate sensor. Despite of the convenience for getting sleep stage information from Fitbit, our study suggests that such data may not be reliable and thus should be used with caution. We recommend future studies to focus on enhancing the accuracy of Fitbit in classifying sleep stages (including wake stage).

B. Interpretation of Visualization

Visual aid has been used in the assessment of sleep-staging algorithms in [46]. However, our study is the first attempt to leverage visual aid in the validation of consumer sleep trackers. The color spectrums in Fig.3 demonstrate the temporal patterns of Fitbit accuracy in measuring sleep stages. We did not find significant difference among six sleep cycles with respect to the accuracy of Fitbit in detecting wake and light sleep epochs. The difference mainly lies in the accuracy of deep and REM sleep. It is interesting to see that deep sleep epochs were more likely to be correctly detected during sleep cycle 1~2 (i.e. significantly higher portion of pink in the spectrum of "Deep"), whereas REM sleep epochs were more likely to be corrected detected during sleep cycle 4~6 (i.e. significantly higher portion of blue in the spectrum of "REM"). In other words, Fitbit Charge 2™ had better accuracy in detecting deep sleep during the first half of the night, while the accuracy in detecting REM sleep is better during the second half of the night. These temporal patterns of classification accuracy suggest that sleep staging algorithms should count in the effect of time and may even consider using segmented modelling techniques [47-49].

C. Limitations

This study has the following limitations. First, the data was collected only from a cohort of healthy young adults. The results thus may not be generalized to children or teenagers, older population, and people with chronic conditions. Second, the study design precluded the assessment of longitudinal performance of the device. Future studies are encouraged to establish evidence on the validity of Fitbit devices in measuring the sleep of various populations spanning over multiple days.

VI. CONCLUSION

We have presented our approach that combines both numerical techniques and data visualization in epoch-by-epoch validation of Fitbit Charge 2™. We compared Fitbit sleep data with medical data collected concurrently from 22 healthy young participants. The numerical results suggested that Fitbit device produced reasonably good accuracy (>0.60) in classifying light, deep and REM sleep stages. Distinguishing between deep and REM stages was relatively easier than distinguishing these two stages from light stage. Similar to previous models, Fitbit Charge 2™ has poor specificity of 0.38 in detecting wake. The challenge lies in differentiating wake from light stage, as 48% of wake epochs were misclassified as light sleep. As for the temporal patterns of device accuracy, data visualization using scatter plots revealed that Fitbit had better performance in detecting deep sleep in the first half of night, while the accuracy in detecting REM sleep was better in the second half of night. Our findings suggest that despite of using multiple sensing modalities, the ability of Fitbit in classifying sleep stages (especially wake stage) still remains limited. Therefore, users should count in these limitations when interpreting sleep data measured by these consumer devices.

ACKNOWLEDGMENT

This study was supported by the JSPS KAKENHI Grant-in-Aid for Research Activity Start-up (Grant Number 16H07469). The authors would like to thank the participants for their contributions to this study.

REFERENCES

[1] Z. Liang, and M. A. Chapa-Martell, "Framing self-quantification for individual-level preventive health care." In Proc. of HEALTHINF, pp. 336-343, 2015.

[2] M. Swan, "Sensor mania! The internet of things, wearable computing, objective metrics, and the quantified self 2.0," Journal of Sensors and Actuator Networks, vol. 1, no. 3, pp. 217-253, 2012.

[3] S. Fischer, "Sleep on it: sleep might just be the most important part of daily health‐and the biggest new target for biomedical engineering," IEEE Pulse, vol. 5, no. 5, pp. 8-13, 2014.

[4] Z. Liang, B. Ploderer, W. Liu, Y. Nagata, J. Bailey, L. Kulik, and Y. Li, "SleepExplorer: a visualization tool to make sense of correlations between personal sleep data and contextual factors," Personal Ubiquitous Comput., vol.20, no.6, 2016, pp. 985-1000.

[5] Z. Liang, B. Ploderer, M. A. Chapa-Martell, and T. Nishimura, "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, A. Martin-Gonzalez and V. Uc-Cetina, eds.: Springer, Cham, 2016.

[6] M. De Zambotti, F. Baker, C, and A. R. Willoughby, "Measures of sleep and cardiac functioning during sleep using a multi-sensory commercially-available wristband in adolescents," Physiological & Behavior, vol. 158, pp. 143-149, 2016.

[7] Z. Liang, and B. C.-M. Ploderer, Mario Alberto, "Is fitbit fit for sleep-tracking?: sources of measurement errors and proposed countermeasures," in Proc. of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare, Barcelona, Spain, 2017, pp. 476-479.

[8] R. Yang, E. Shin, M. Newman, and M. Ackerman, "When fitness trackers don't 'fit': end-user difficulties in the assessment of personal tracking device accuracy," in Proc. of UbiComp, Osaka, Japan, 2015, pp. 623-634.

[9] Z. Liang, and B. Ploderer, "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, 2016, pp. 537-541.

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

[11] Z. Liang, and M. A. Chapa-Martell, "Validity of consumer activity wristbands and wearable EEG for measuring overall sleep parameters and sleep structure in free-living conditions " Journal of Healthcare Informatics Research, pp. 1-27, 2018.

[12] M. De Zambotti, S. Claudatos, S. Inkelis, I. Colrain, and F. Baker, "Evaluation of a consumer fitness-tracking device to assess sleep in adults," Chronobiology International, vol. 32, no. 7, pp. 1024-1028, 2015.

[13] M. De Zambotti, A. Goldstone, S. Claudatos, and e. al., "A validation study of Fitbit Charge 2 compared with polysomnography in adults," Chronobiology International, vol. 35, no. 4, pp. 465-476, 2017.

[14] D. Altman, and M. Bland, "Diagnostic tests 1: sensitivity and specificity," BMJ, vol. 308, no. 6943, pp. 1552, 1994.

[15] D. Altman, and M. Bland, "Diagnostic tests 2: predictive values," British Medical Journal, vol. 309, pp. 102, 1994.

[16] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, New York: Springer-Verlag, 2016.

[17] P. Teetor, R Cookbook, p.^pp. 223: O'Reilly, 2011.

[18] D. Buysse, C. Reynolds, T. Monk, and e. al., "The Pittsburgh sleep quality index: a new instrument for psychiatric practice and research," Psychiatry Res, vol. 28, no. 2, pp. 193-213, 1989.

[19] A. Douglass, R. Bornstein, G. Nino-Murcia, and e. al., "The sleep disorders questionnaire I. Creation and multivariate structure of SDQ," Sleep, vol. 17, pp. 160-167, 1994.

[20] I. Ancoli-Israel, A. Chesson, and S. Quan, "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, pp. Version 2.4, 2017.

[21] R. Berry, R. Brooks, C. Gamaldo, and e. al, The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications, Version 2.4 ed., Darien, IL: American Academy of Sleep Medicine, 2017.

[22] R. Rosenberg, and S. Van Hout, "The American Academy of Sleep Medicine inter-scorer reliability program: sleep stage scoring," J Clin Sleep Med, vol. 9, no. 1, pp. 81-87, 2013.

[23] V. Natale, D. Leger, M. Martoni, and e. al., "The role of actigraphy in the assessment of primary insomnia: a retrospective study," Sleep Medicine, vol. 15, no. 1, pp. 111-115, 2014.

[24] A. Van de Water, A. Holmes, and D. Hurley, "Objective measurements of sleep for non-laboratory settings as alternatives to polysomnography - a systematic review," J Sleep Res, vol. 20, pp. 183-200, 2011.

[25] B. Lucey, J. McLeland, C. Toedebusch, J. Boyd, and e. al, "Comparison of a single-channel EEG sleep study to polysomnography," J Sleep Res, vol. 25, no. 6, pp. 625-635, 2016.

[26] I. Fietze, T. Penzel, M. Partinen, J. Sauter, G. Kuchler, A. Suvoro, and H. Hein, "Actigraphy combined with EEG compared to polysomnography in sleep apnea patients," Physiol Meas, vol. 36, no. 3, pp. 385-396, 2015.

[27] M. Yoshida, H. Shinohara, and H. Kodama, "Assessment of nocturnal sleep architecture by actigraphy and one-channel electroencephalography in early infancy," Early Human Development, vol. 91, no. 9, pp. 519-526, 2015.

[28] M. Marino, Y. Li, M. Rueschman, and e. al., "Measuring sleep accuracy, sensitivity, and specificity of wrist actigraphy compared to polysomnography," Sleep, vol. 36, no. 11, pp. 1747-1755, 2013.

[29] E. K. Choe, S. Consolve, and N. F. Watson, "Opportunities for computing technologies to support healthy sleep behaviors," in Proc. of CHI, Vancouver, BC, Canada, 2011, pp. 3053-3062.

[30] W. Liu, B. Ploderer, and T. Hoang, "In Bed with Technology: Challenges and Opportunities for Sleep Tracking," in Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction, Parkville, VIC, Australia, 2015, pp. 142-151.

[31] R. S. Ravichandran, Sang-Wha, S. N. Patel, and J. A. P. Kientz, Laura R., "Making Sense of Sleep Sensors: How Sleep Sensing Technologies Support and Undermine Sleep Health," in Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, Colorado, USA, 2017, pp. 6864-6875.

[32] N. Daskalova, B. Lee, J. Huang, C. Ni, and J. Lundin, "Investigating the effectiveness of cohort-based sleep recommendations," Proc ACM Interact Mob Wearable Ubiquitous Technol, vol. 2, no. 3, pp. Article 101, 2018.

[33] J. M. Peake, G. Kerr, and J. P. Sullivan, "A critical review of consumer wearables, mobile applications, and equipment for providing biofeedback, monitoring stress, and sleep in physically active populations," Frontiers in Physiology, vol. 9, pp. 743, 2018.

[34] K. Grifantini, "How's my sleep?: Personal sleep trackers are gaining in popularity, but their accuracy is still open to debate," IEEE Pulse, vol. 5, no. 5, pp. 14-18, 2014.

[35] J. Cook, M. Prairie, and D. Plante, "Utility of the Fitbit Flex to evaluate sleep in major depressive disorder: a comparison against polysomnography and wrist-worn actigraphy," J Affect Disorder, vol. 217, pp. 299-305, 2017.

[36] L. Meltzer, L. Hiruma, K. Avis, and e. al., "Comparison of a commercial accelerometer with polysomnography and actigraphy in children and adolescents," Sleep, vol. 38, no. 8, pp. 1323-1330, 2015.

[37] S.-G. Kang, J. M. Kang, K.-P. Ko, P. Seon-Cheol, S. Mariani, and J. Weng, "Validity of a commercial wearable sleep tracker in adult insomnia disorder patients and good sleepers," Journal of Psychosomatic Research, vol. 97, pp. 38-44, 2017.

[38] J. Mantua, N. Gravel, and R. Spencer, "Reliability of sleep measures from four personal health monitoring devices compared to research-based actigraphy and polysomnography," Sensors, vol. 16, no. 5, pp. 646, 2016.

[39] J. Martin, and A. Hakim, "wrist actigraphy," CHEST, vol. 139, no. 6, pp. 1514-1527, 2011.

[40] T. Blackwell, S. Ancoli-Israel, S. Redline, and K. Stone, "Factors that may influence the classification of sleep-wake by wrist actigraphy: the MrOS sleep study," J Clin Sleep Med, vol. 7, no. 4, pp. 357-367, 2011.

[41] T. Blackwell, S. Redline, S. Ancoli-Israel, J. Schneider, and e. al, "Comparison of sleep parameters from actigraphy and polysomnography in older women: the SOF study," Sleep, vol. 31, no. 2, pp. 283-291, 2008.

[42] V. Natale, G. Plazzi, and M. Martoni, "Actigraphy in the assessment of insomnia: a quantitative approach," Sleep, vol. 32, no. 6, pp. 767-771, 2009.

[43] R. D. C. Team, R: A language and environment for statistical computing (http://www.R-project.org), Vienna, Austria: R Foundation for Statistical Computing.

[44] V. Ibanez, J. Silva, and O. Cauli, "A survey on sleep assessment methods," PeerJ, vol. 6, pp. e4849, 2018.

[45] M. Ohayon, E. M. Wickwire, M. Hirshkowitz, and e. al., "National Sleep Foundation's sleep quality recommendations: first report," Sleep Health, vol. 3, no. 1, pp. 6-19, 2017.

[46] A. Supratak, H. Dong, C. Wu, and Y. Guo, "DeepSleepNet: a model for automatic sleep stage scoring based on raw single-channel EEG," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 25, no. 11, pp. 1998-2008, 2017.

[47] V. Muggeo, and G. Adelfio, "Efficient change point detection for genomic sequences of continuous measurements," Bioinformatics, vol. 27, no. 2, pp. 161-166, 2011.

[48] V. Muggeo, "Interval estimation for the breakpoint in segmented regression: a smoothed score-based approach," Aust N Z J Stat, vol. 59, no. 3, pp. 311-322, 2017.

[49] A. Arnold, L. Matheson, L. Harvey, and B. McNeil, "Temporally segmented modelling: a route to improved bioprocess monitoring using near infrared spectroscopy," Biotechnology Letters, vol. 23, pp. 143-147, 2001.