In Vivo Comparison of Sleep Stage Scoring Between Commercialized Wearable Devices and Polysomnography System

Quyen N.T. Nguyen, Phuong N. Bui, Trung Q. Le, Hoang H. Nguyen, Cac T. Nguyen, and Loc X. Bui

Abstract

This study aims to evaluate the performance of 4 commercialized wearable devices in scoring sleep stages with the ground-truth from polysomnography (PSG) system. The comparisons were performed using data from 14 human subjects simultaneously wearing 4 wearable devices with sleep monitoring function monitored by polysomnography overnight at a Type 1 sleep lab. The compared features were categorized into 2 groups including (1) sleep–wake pattern and (2) sleep distribution. Wearable devices with sleep monitoring functions used in this study are from 4 different brandnames including Misfit, Garmin, Jawbone and Fitbit. These devicesare anonymously named as Device A, B, C and D. Using PSG system as benchmark, wearable devices earned good sensitivity, especially in detecting sleep onset and sleep period time in contrast to poor specificity, particularly in monitoring sleep stages. However, specificities in terms of wake–sleep transition features reported from wearable devices are low compared to those reported from polysomnography. Among 4 wearable devices, device C with the sensors to capture the heart rate, respiratory rate, body temperature, galvanic skin response as well as an accelerometer proved the best device in detecting not only sleep–wake transition but sleep stages as well. From the device perspective, we suggest that the usage of both actigraphy and heart rate sensors in the wearable devices and proper selection of sleep features can yield better agreement between wearable devices and the gold standards—PSG—in determining the sleep stages.

Keywords

Wearable device • Polysomnography • Sleep monitoring • Sleep stage scoring

1 Introduction

Actigraph-based methods have been used widely for the monitoring of sleep although its accuracy is still questionable. In 1995, Sadeh et al. [1] made the comparison between actigraphy and PSG and concluded that actigraphy did provide useful information and might be a promising method for detecting some sleep disorders. In 2003, there was a review of AASM (American Academy of Sleep Medicine) using 171 articles of actigraphy comparison and concluded that actigraph was valid and reliable in the normal population; it is still questionable in the population of sleep-related disorder patients. Also, while actigraph was good at detecting sleep, it was poor at wake detection [2]. Since that time, this technology has been developed rapidly both in hardware and algorithm; as a result, comparative research becomes necessary. In addition to actigraphy, wearable devices recently have been facilitated with more sensors for the recording of other physiological parameters during sleep

Q.N.T. Nguyen · P.N. Bui · T.Q. Le (✉)
Biomedical Engineering, International University—Vietnam National University, Quarter 6, Linh Trung, Thu Duc District, Ho Chi Minh City, Vietnam
e-mail: trung.le@hcmiu.edu.vn

H.H. Nguyen
Phoi Viet Clinics, Ho Chi Minh City, Vietnam

C.T. Nguyen · L.X. Bui
Fossil Group, Ho Chi Minh City, Vietnam

T. Vo Van et al. (eds.), *6th International Conference on the Development of Biomedical Engineering in Vietnam (BME6)*, IFMBE Proceedings 63, https://doi.org/10.1007/978-981-10-4361-1_135

time; hence, their performances have been improved considerably [3]. However when it comes to a detailed sleep stage classification that consists of all wake, light sleep, deep sleep, and REM stages, the current wearable devices do not show moderate agreement in comparison with PSG [4].

The drawbacks of the previous work are the lack of assessing methods to compare sleep architecture detected by actigraph-based devices and justifications to explain the differences between wearable devices and PSG in detailed structures of sleep under different health conditions. The primary goal of this study is to make the comparison between the gold standard—PSG system and 4 commercial wearable devices, which is separated into 2 types—the first ones scored sleep based on actigraphy while the second one derived the results from actigraphy combining with other bio-signals. The compared sleep parameters were classified in a wide range of features characterizing normal human sleep [4–6] and common sleep disorders [7]. Additionally, the secondary objective of this study is to provide reasonable explanations for the discrepancies between sleep monitoring system and the PSG system.

2 Methodology

2.1 Experimental Design and Data Collection

Fourteen volunteers (9 males, 5 females) in the age range of 20.2 ± 2.0 took part in this study for 17 overnight data acquisitions. During the experiment, the participants are required to wear four types of wearable devices on wrist along with the PSG system and asked to sleep comfortably for at least 8 h per night. After average of 8 h of data collection, PSG data and wearable devices' data will be collected and stored anonymously. Hypnograms—graphs that indicate the sleep stages—from the wearable devices as well as from the PSG system are collected for further sleep stage analysis. Type1 PSG system-Alice5 Philips™ system (Fig. 1) utilized in the research consists of 11 types of sensors including EEG, EOG, ECG, leg EMG, chin EMG, thermal flow and snore, respiratory inductance plethysmography bands at the chest and abdomen, position sensor, and pulse oximeter. The signals were presented graphically on a computer screen for the real-time monitoring by technicians. The PSG signals were used to categorize sleep into 5 stages-Wake, NREM (including N1, N2, and N3) and REM-using American Academy of Sleep Medicine (AASM) scoring manual.

Each of the commercial wearable devices with sleep monitoring functions will be named respectively as device A, device B, device C, and device D and compared anonymously. In detail, device A, B, C, and D use accelerometer sensors to detect the 3-axis motion of the subject's wrist over the night and associate it with the sleep stages. In addition to the accelerometers, Device B is equipped with an additional optical sensor for heart rate detection. Device C uses the bio-impedance sensor to measure heart rate, respiratory rate, body temperature, galvanic skin response as well as an accelerometer to detect stages of sleep. These devices stream data automatically into the server therefrom the algorithms classify sleep into different stages. In particular, devices A, B and D classify sleep into Wake, Light and Deep stages while device C divides sleep into Wake, REM, Light and Deep stages.

2.2 Sleep Feature Extraction

Features for the comparisons have been categorized into 2 groups namely wake–sleep analysis and sleep distribution. The first group is wake–sleep analysis including 3 groups of sleep features: sleep quality, sleep disturbance and wake–sleep transition. The sleep quality which includes sleep features having impacts on human health based on three components of health–mental, social, and physical, identified by World Health Organization (WHO) [8]. These features are sleep period time and sleep efficiency. Considering the AASM guideline, we define the sleep duration as the time from sleep onset to the last epoch of sleep and the sleep efficiency is the percentage of total sleep time per sleep

Fig. 1 PSG system and wearable device setups

period time. The other two groups are sleep disturbance and wake–sleep transition comprising 5 parameters which are the important diagnostic factors for several popular sleep disorders like insomnia, narcolepsy and depression [9]. These features include Wake After Sleep Onset, REM latency (time from sleep onset to first REM), sleep onset and wake after sleep. These four characteristic features are recommended to be evaluated between the benchmarked system and wearable devices [5]. These 3 groups of sleep features technically distinguish when subjects wake or sleep, statistical comparison of this group show how imprecise wearable devices can analyze between awake and sleep stages.

The second comparative group is sleep distribution features which characterize the relative of different sleep stages's durations over the sleeping time. Criteria related to sleep distribution have been used to evaluate a normal human sleep. Besides characterizing the normal human sleep, sleep architecture can help to understand sleep pathology. These features is also the gap of previous research, which mostly focused on Wake–sleep analysis features only while sleep distribution is necessary in medical diagnosis. Comparisons in sleep distribution help us to know how exactly wearable devices identify sleep stages.

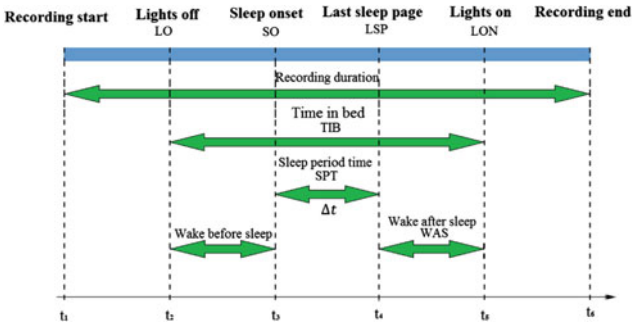

Fig. 2 Sleep recording

The description and estimation of each parameter were shown in Table 1; Fig. 2.

2.3 Statistics Tests

To investigate the accuracy of the wearable devices in scoring sleep with the references of PSG system we performed pairwise statistical analysis over sleep features aforementioned. In this study, we used paired sample t-test to consider

Table 1 Description of two groups of sleep feature

Category	Features	Description	Formula	Unit
Sleep distribution	Percentage of light sleep (PLS)	The ratio of total light sleep time (TLST) during total sleep time (TST)	$PLS = \frac{TLST}{TST}$	%
	Percentage of deep sleep (PDS)	The ratio of total deep sleep time (TDST) during total sleep time (TST)	$PDS = \frac{TDST}{TST}$	%
	Percentage of REM sleep (PRS)	The ratio of total REM sleep time (TRST) during total sleep time (TST)	$PRS = \frac{TRST}{TST}$	%
Minor features	Total sleep time (TST)	The time subject spends on sleep during sleep period time	$TST = TLST + TDST + TRST$	Min
Sleep disturbance	Sleep period time (SPT)	The elapsed time from sleep onset (SO) to last epoch of sleep (LSP)	$SPT = LSP - SO$	Min
	Sleep efficiency (SE)	The ratio of total sleep time (TST) and sleep period time (SPT)	$SE = TST/SPT$	%
	Wakefulness after sleep onset (WASO)	Minutes of awake time (AWT) during sleep onset (SO) to light on (Azagra Calero et al.)	$WASO = SPT - TST + WAS$	Min
	Sleep onset (SO)	The point of time when the subject undergoes a transition from wakefulness into sleep	$1st\ N1/N2/N3/REM$	hh: mm: ss

the statistical differences between the sleep features from PSG system and those from wearable devices. Besides, we also utilized the cohen's d effect size value to characterize the differences between the features.

3 Results

We reported our results regarding the pairwise comparisons of the 2 groups of aforementioned features. In this session, we summarized the comparisons in 2 groups namely wake–sleep analysis, sleep distribution and reported features from the investigated wearable devices that are significantly different from those from PSG system. A full list of all feature comparison and the statistical significant values were reported finally.

3.1 Wake–Sleep Analysis

The Wake–sleep analysis comparisons compare between wearable devices and polysomnography system in 4 sleep features: sleep period time (SPT), sleep efficiency (SE), wakefulness after sleep onset (WASO), and sleep onset (SO).

Figure 3 demonstrates how accurately 4 wearable devices scored sleep period time. The results were compared to PSG system. In case number 1 and 3, 3 over 4 devices detect a wrong SPT which is very different from PSG result. The reason for this mistake is subject took off these devices without turning off so they cannot detect any motion and give out results that subject still sleeping while he/she is not. These two cases make the range of different value (error) large. It is evident from the graphs that apart from device D, other devices performed quite well and device C are best fit to the PSG with the lowest error besides outlier cases.

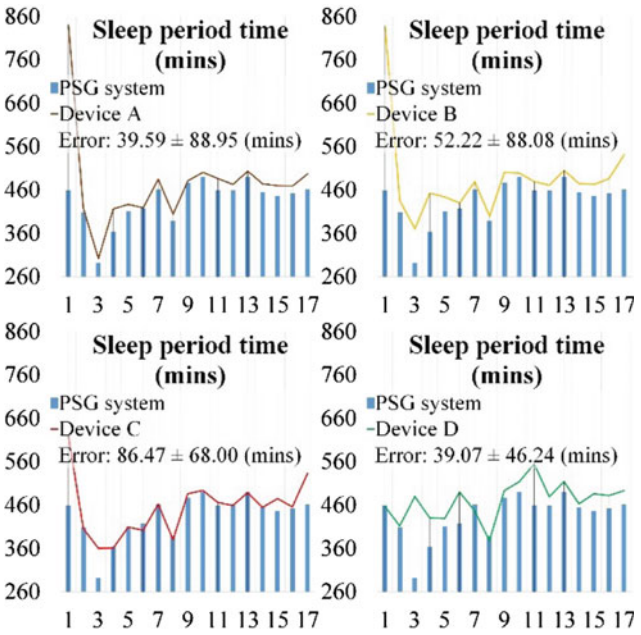

Fig. 3 Sleep period time of wearable devices versus PSG. Data series on horizontal axis does not represent for object, only represent for order

Figures 4 and 5 describes the accuracy of 4 wearable devices in scoring sleep efficiency (SE) and wake after sleep onset (WASO) compared to PSG. These two features stand for the ability of detecting wake after subject go to sleep of wearable devices. From the formula, SE and WASO are inversely propotional. Results from device A, B and D shows that in most of the case, SE nearly 100% while WASO almost 0 min. In general, device C showed the best result and it was followed by device D. In contrast, other devices seemed to fail in detecting this sleep feature. In the other words, after analyse that subject goes to sleep stage, wearable devices fail to detect wake during his/her sleeping time.

Figure 6 provides information about how well 4 wearable devices scored sleep onset in comparison with PSG. Overall, in most of the cases, Device A, B and D detect SO before subject actually sleep (base on PSG result). The reason is that when subjects did not sleep but lying still, devices considered they slept while they did not, which leads to the large range of error. Results also shows that there was a considerable accuracy of device C over other devices in detecting sleep onset. Particularly, the average error of device C is 7.15 $\pm$ 7.10 min which is 2 times lower than the others'.

Table 2 summarizes the overall t-test statistics and the cohen's d values.

It is evident from the table that apart from device A and B, device C and D performed quite well in wake–sleep analysis and device C are best fit to the PSG with the lowest error. Results also give out 2 problem of wearable devices in detecting wake and sleep stage. Firstly, when patients lie still, these devices will give the wrong diagnoses. Secondly, wearable devices are erratic in recognizing whether patients are wearing them or taking off. In this case, actigraphy will identify no motion, that cause of wrong analysis.

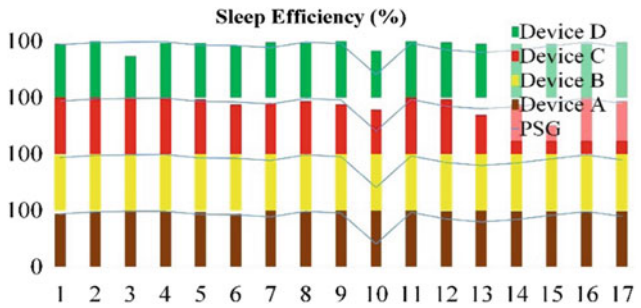

Fig. 4 Sleep efficiency of wearable devices versus PSG

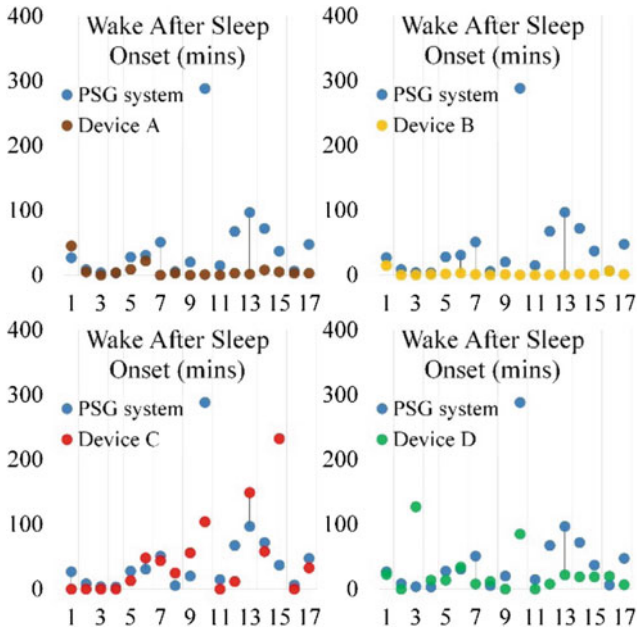

Fig. 5 Wake after sleep onset of wearable devices versus PSG

3.2 Sleep Distribution Features

Because four wearable devices define sleep stages differently (Wake–Light–Deep or Wake–REM–Light–Deep) and different from PSG's definition (Wake–REM–N1–N2–N3), especially while in medical field, REM sleep is not classified as light or deep sleep. Therefore, in order to find out the best corresponding definition to PSG's definition to use these devices for medical diagnosis, statistical test is done with 4 different ways of classification as shown in Table 1. Device C has REM sleep detection therefore there are only 2 possible definitions (group 1 and 2).

Table 4 summarizes statistical results systematically between 4 wearable devices in 4 groups of definition.

The result shows that Device C detects REM sleep quite accurately with PSG's result (There is no statistical difference between 2 results) but in both cases of definitions, it cannot detect Light and Deep sleep correctly. The best result is when classifying N1 and N2 as light sleep, N3 as deep sleep.

Also, levels of accuracy depend on wearable devices (sensors) and their sleep stage definition (algorithm). In general, the closest result belongs to Device A, which defines stage N1 and N2 as light sleep, N3 and REM as deep sleep; follow is Device B and D, which considers N1 as light sleep, and combine 3 other stages as deep sleep.

Overall, among the 4 ways of definition, device B gives the most incorrect detection while device D gives significant different result in all 4 groups. Besides, statistical tests show that results from group 3 and 4 definition are most inaccurate among 4 devices.

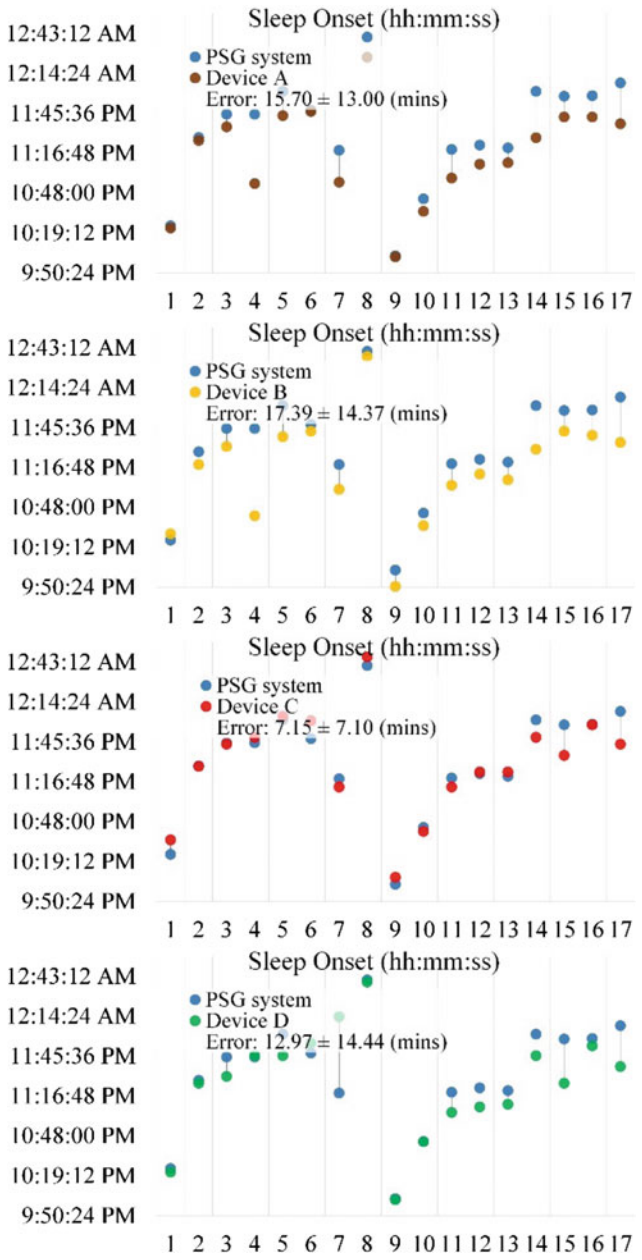

Fig. 6 Sleep onset detected from the wearable versus PSG system

4 Discussions

It is concluded that the wearable devices earned good sensitivity in detecting sleep onset and sleep period time in contrast to poor specificity, particularly in monitoring sleep stages. We proposed two possible reasons that affect the accuracy of wearable devices. The first reason came from how commercialized wearable devices classify sleep stages. The other factor affecting the result was the devices' sensors.

The first reason of the discrepancy comes from the way the devices group sleep stages as shown in Table 3. It has been shown that sleep stage N3 was extremely different from

Table 2 Summary of statistical comparisons of wake–sleep analysis

Features	Paired sample t-test			
	PSG versus Device A	PSG versus Device B	PSG versus Device C	PSG versus Device D
SPT	**No sig. dif.**	Sig. dif.	**No sig. dif.**	Sig. dif.
	$\rho = 0.085$	$\rho = 0.026$	$\rho = 0.095$	$\rho = 0.008$
	$\eta^2 = 0.474$	$\eta^2 = 0.663$	$\eta^2 = 0.323$	$\eta^2 = 0.770$
SE	Sig. dif.	Sig. dif.	**No sig. dif.**	Sig. dif.
	$\rho = 0.018$	$\rho = 0.010$	$\rho = 0.884$	$\rho = 0.125$
	$\eta^2 = 0.920$	$\eta^2 = 1.014$	$\eta^2 = 0.040$	$\eta^2 = 0.478$
WASO	Sig. dif.	Sig. dif.	**No sig. dif.**	**No sig. dif.**
	$\rho = 0.028$	$\rho = 0.014$	$\rho = 0.905$	$\rho = 0.147$
	$\eta^2 = 0.846$	$\eta^2 = 0.955$	$\eta^2 = 0.032$	$\eta^2 = 0.440$
SO	Sig. dif.	Sig. dif.	**No sig. dif.**	**No sig. dif.**
	$\rho < 0.001$	$\rho < 0.001$	$\rho = 0.494$	$\rho = 0.243$
	$\eta^2 = 0.047$	$\eta^2 = 0.051$	$\eta^2 = 0.005$	$\eta^2 = 0.017$

The shaded values indicate the sleep features that are significantly different while "No sig. dif." means "There is no significant difference between Device and PSG system", "Sig. dif." means "There is a significant difference between Device and PSG system", is significant value of t-test, the larger, the more accurate of the results, shows power of the test, the smaller, the closer between wearable devices and PSG detection

Table 3 Corresponding sleep stages definition for comparison

Devices	Sleep stages				
Group 1:					
PSG	N1	N2	N3	REM	Wake
Device C	Light		Deep	REM	Wake
Device A, B, D	Light		Deep		Wake
Group 2:					
PSG	N1	N2	N3	REM	Wake
Device C	Light	Deep		REM	Wake
Device A, B, D	Light	Deep			Wake
Group 3:					
PSG	N1	N2	N3	REM	Wake
Device A, B, D	Light			Deep	Wake
Group 4:					
PSG	N1	N2	REM	N3	Wake
Device A, B, D	Light			Deep	Wake

REM sleep both in brain activity and body movement [5]. To be specific, during stage N3, the brain undergoes the most restful status [6] and there is a minimum in muscle activity compared to other sleep stages [10]. In contrast, REM sleep shows a highly activated brain, especially in motor area of the cortex and despite the inhibition of voluntary motor system, the body motion level of this sleep stage only less than wake stage and stage N1 as shown the following sequence: Wake>N1>REM>N2>N3 [10]. In statistical tests showed above, group 4 is divided base on levels of body movement: (N1 + REM + N2) = Light sleep, N3 = Deep sleep and gives inaccuracy results among 4 devices. As the result a new way to classify sleep stage into Wake, REM and NREM should be considered to improve the accuracy of the sleep scored by wearable devices.

The second reason was related to the low sensitivity of the motion sensors in capturing the sleep stage transitions. Besides autonomic nervous system, sleep stages are manifested by the somatic nervous system [11]. In the current commercial wearable devices, sleep stages are characterized by various somatic signals of the peripheral nervous system that are captured by the motion. However due to the low muscle tone threshold in the wake stage right after a REM sleep, it is too hard for devices to detect this wake. In other words, REM sleep experiences the inhibition of the voluntary motor system or the skeletal muscles; therefore, the muscles require more energy to overcome the inhibition threshold to move. As a result, sometimes subjects could not move immediately although there were changes in EEG and EMG which were considered as wakefulness by PSG

Table 4 Summary of statistical comparisons of Sleep distribution

Features	Paired sample t-test			
	Group 1	Group 2	Group 3	Group 4
Device A: (Accelerometer sensor)				
PLS and PDS	**No sig. dif.**	Sig. dif.	Sig. dif.	Sig. dif.
	$\rho = 0.527$	$\rho < 0.001$	$\rho < 0.001$	$\rho = 0.002$
	$\eta^2 = 0.200$	$\eta^2 = 3.021$	$\eta^2 = 1.562$	$\eta^2 = 1.210$
Device B: (Accelerometer sensor, optical sensor)				
PLS and PDS	Sig. dif.	**No sig. dif.**	Sig. dif.	Sig. dif.
	$\rho < 0.001$	$\rho = 0.311$	$\rho < 0.001$	$\rho < 0.001$
	$\eta^2 = 7.291$	$\eta^2 = 0.343$	$\eta^2 = 15.460$	$\eta^2 = 11.910$
Device C: (Accelerometer sensor, bio-impedance sensor)				
PLS	Sig. dif.	Sig. dif.		
	$\rho = 0.005$	$\rho < 0.001$		
	$\eta^2 = 1.053$	$\eta^2 = 4.065$		
PDS	Sig. dif.	Sig. dif.		
	$\rho < 0.001$	$\rho < 0.001$		
	$\eta^2 = 1.678$	$\eta^2 = 7.129$		
PRS	**No sig. dif.**			
	$\rho = 0.722; \eta^2 = 0.124$			
Device D: (Accelerometer sensor)				
PLS and PDS	Sig. dif.	Sig. dif.	Sig. dif.	Sig. dif.
	$\rho < 0.001$	$\rho = 0.017$	$\rho < 0.001$	$\rho < 0.001$
	$\eta^2 = 4.062$	$\eta^2 = 0.959$	$\eta^2 = 7.451$	$\eta^2 = 6.516$

system. Hence, characteristic signals generated by the autonomic nervous system need to be accounted due to their stability in capturing the voluntary movements in scoring sleep stages [11, 12]. Furthermore, the classification training should include different factors affecting sleep structure. We recommend 5 groups of factors which may affect the sleep which are (1) lifespan, (2) daytime habits including inactive (TV/game addiction, working in office) and overactive ones (manual labors), (3) diseases like mental problems or limb movement disorders, (4) usage of medications (antianxiety) or beverages (caffeine, alcohol) before bedtime, and (5) sleep conditions like temperature, skin conductance.

5 Conclusions

In conclusion, the fast growth of polysomnographic alternatives for point-of-care applications has urged more standardized comparative research conducted to validate the sleep monitoring devices and improve the current design of wearable sleep monitoring system. In this paper, we proposed a list of sleep characteristic features and statistical comparison to specify the impact of gender on devices as well as the deviation of commercialized sleep monitoring products with the PSG ground truths. The results were that among 4 wearable devices, device C with the bio-impedance sensor was the best one which detect well not only sleep–wake patterns but REM sleep as well. Meanwhile, device D showed good results on sleep–wake patterns but inaccurate result in sleep distribution. In contrast, the remaining devices cannot be used for medical purposes because of their poor scoring in stages and sleep–wake patterns. For the device improvement, we came up with two possible causes which were sleep-stage synchronization, and data acquisition process. Besides, we also proposed two suggestions which were to classify training data and to consider the autonomic signals such as heart rate used in device C. Temporal localized correlation coefficients of the comparative features have been investigated to characterize quantitatively the discrepancies between wearable devices and PSG. We expect that from the results of this test, the causes of the poor detection will become more evident. It helps to increase the accuracy of the machine learning algorithms in to the current PSG system to enhance the diagnosis and prediction of the other sleep disorders [13, 14].

Acknowledgements The authors would like to thanks the technician team of the Clinical Sleep Lab at Biomedical Engineering Department-Ho Chi Minh City International University of Vietnam National University for their help in organizing, collecting and revising the data. We also thanks Ms. Nguyen Thi Thu Hang for her English edition of the manuscript.

References

1. Sadeh A et al (1995) The role of actigraphy in the evaluation of sleep disorders. Sleep 18(4):288–302
2. Sadeh A (2011) The role and validity of actigraphy in sleep medicine: an update. Sleep Med Rev 15(4):259–267

3. Le TQ, Cheng C, Sangasoongsong A, Wongdhamma W, Bukkapatnam ST (2013) Wireless wearable multisensory suite and real-time prediction of obstructive sleep apnea episodes. IEEE J Trans Eng Health Med 1:2700109
4. Hedner J et al (2011) Sleep staging based on autonomic signals: a multi-center validation study. J Clin Sleep Med 7(3):301–306
5. Lee-Chiong TL (2006) Sleep: a comprehensive handbook. Wiley Online Library
6. Hall JE (2015) Guyton and Hall textbook of medical physiology. Elsevier Health Sciences
7. Hudson JI et al (1992) Good sleep, bad sleep: a meta-analysis of polysomnographic measures in insomnia, depression, and narcolepsy. Biol Psychiat 32(11):958–975
8. Hirshkowitz M et al (2015) National Sleep Foundation's sleep time duration recommendations: methodology and results summary. Sleep Health 1(1):40–43
9. Hudson JI, Pope HG, Sullivan LE, Waternaux CM, Keck PE, Broughton RJ (1999) Good sleep, bad sleep: a meta-analysis of polysomnographic measures in insomnia, depression, and narcolepsy. Biol Psychiatry 32(11):958–975
10. Wilde-Frenz J et al (1983) Rate and distribution of body movements during sleep in humans. Percept Mot Skills 56 (1):275–283
11. Herscovici S et al (2006) Detecting REM sleep from the finger: an automatic REM sleep algorithm based on peripheral arterial tone (PAT) and actigraphy. Physiol Meas 28(2):129
12. Bresler M et al (2008) Differentiating between light and deep sleep stages using an ambulatory device based on peripheral arterial tonometry. Physiol Meas 29(5):571
13. Karandikar K, Le TQ, Sa-ngasoongsong A, Wongdhamma W, Bukkapatnam ST (2013) Detection of sleep apnea events via tracking nonlinear dynamic cardio-respiratory coupling from electrocardiogram signals. 6th International IEEE/EMBS Conference, pp 1358–1361
14. Le TQ, Bukkapatnam ST (2016) Nonlinear dynamics forecasting of obstructive sleep apnea onsets. PloS One 11(11):e0164406