Assessment of Consumer-Grade Wearable Devices to Track Sleep in Healthy Individuals in Free-Living Conditions

Sanjay Rajput, Alexandra Jamieson, Nishi Chaturvedi, Alan Hughes, Michele Orini

MRC Unit for Lifelong Health and Aging, University College London, Institute of Cardiovascular Science, London, United Kingdom

Abstract

Novel consumer-grade devices provide the opportunity to track sleep, but their use in research is limited by a lack of validation. This study aimed to validate sleep tracking in free-living conditions by comparing total sleep duration measured by popular wearable devices with sleep diaries. Twenty-seven heathy volunteers of mean age 25 (± 9.49), 71% male, and with wide range of skin tones (Fitzpatrick scale from 1 to 6) wore 5 devices for 2 consecutive nights and provided sleep diaries. Devices included the Garmin Vivoactive 4 (GV4) and 4S (GV4S), Fitbit Sense (FS), Withings ScanWatch (WS), and the Oura Ring (OR), which measured sleep using both a standard and beta software. Agreement was assessed using the Spearman's correlation coefficients and Bland Altman plots. Correlation ranged from 0.41 for GV4 to 0.76 for FS, while limits of agreement ranged from (-125, 71) minutes for OR-β to (-115, 256) minutes for GV4. Pair-wise comparisons showed that the absolute percentage error was not significantly different in most cases, except for GV4 (larger than FS, WS, OR and OR-β) and for OR-β (lower than OR). No association was found between the absolute error and skin tone, body mass index or wrist circumference. This data shows moderate to good agreement between wearable-enabled sleep tracking and sleep diaries in free-living conditions.

1. Introduction

Sleep is a vital physiological process that is critical to maintaining physical and mental health. Variations in sleep durations have been implicated in the pathogenesis of multiple diseases, including obesity, Alzheimer disease and cardiovascular disease (CVD). Moreover, compared to normal sleep (6-9 hours), both short and long sleep durations have been positively associated with all-cause mortality (1,2). Polysomnography (PSG) is the gold standard for sleep tracking. It can accurately obtain a detailed picture of sleep including sleep-wake discrimination and sleep-stage analysis, using multiple probes and sensors (3). A less invasive option is wrist actigraphy, which uses accelerometry to detect movement. However, it is significantly less accurate than PSG (4). Alternatively, sleep diaries can be used to document perceived sleep onset and waking. Diaries are low-cost and convenient and widely used in epidemiological studies but are subjective and open to forgetfulness. Wearable devices track sleep through measurements such as heart rate (HR), skin temperature, oxygen saturation and movement. Compared to cumbersome PSG technology, wearables are more convenient and therefore scalable (Figure 1). Moreover, they are extremely prevalent across the globe, with more than 10% of people in the USA, UK and Australia owning a smartwatch (5). The high prevalence of wearables could potentially enable researchers to collect millions of hours of sleep-data, thus transforming the landscape of longitudinal sleep studies. The limiting factor to such change is the known about the accuracy of these devices.

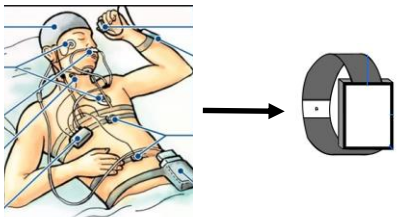

Figure 1: Polysomnography vs wearable device. Adapted from Crivello et al (6).

The technology used to track sleep primarily relies on accelerometry, to measure movement, and photoplethysmography (PPG), which uses light to measure HR and HR variability (7,8). Some devices use other body measurements to further discriminate between sleep-stages, including oximetry, and surface temperature.

This data is analyzed by proprietary algorithms and compared to a known reference range of sleep-data to categorize sleep/wake periods and sleep stages. The exact algorithms used by companies is non-standardized and undisclosed, leading to vastly different results.

Due to PPG's dependence on light absorption, studies have suggested that it performs less accurately in darker skin tones due to the influence of melanin on light

penetrance. In a meta-analysis of ten studies, including 469 participants, six suggested significantly reduced accuracy of HR monitoring in darker-skinned individuals (9). The translation of this to sleep-tracking is yet to be adequately addressed. This study aims to provide an independent report of the accuracy of wearable devices in free-living conditions:

1. Assess the accuracy of four consumer grade wearable devices and one beta software at assessing total sleep duration (TSD) when compared to participant's sleep diary.
2. Assess the influence of factors, such as skin tone, wrist circumference and body mass index (BMI) on TSD accuracy.

2. Methods

Twenty-seven participants completed the sleep analysis. All procedures were in accordance with the principles of Helsinki declaration. All participants provided written informed consent and the study was approved by UCL Research Ethics Committee (Project ID: 21787.001).

Two sets of devices were used to collect data. As shown in Figure 2, set 1 included: Garmin VivoActive 4 (GV4), Garmin VivoActive 4S (GV4S), Fitbit Sense (FS), Oura Ring Generation 3 (OR3), Withings ScanWatch (WS), Nokia G20 (Nokia, Espoo, Finland) and iPhone 14 Pro Max (Apple, California, United States). Set 2 included: GV4 (x2), FS, OR3 WS, Nokia G20 and iPhone X. The Oura Ring beta software (OR3β), accessed via the Oura app, used the same raw data from OR3.

The only small difference between the two sets of devices was that set 1 included one GV4 and GV4S, while set 2 contained two GV4. From here, set 1 will be referred to as 'GV4/4S set 1' and set 2 as 'GV4 set 2'. Despite the difference in models, Garmin quote that both device variants have the same Advanced Sleep Tracking functionality.

Figure 2: Devices.

2.1. Procedure

Each participant wore all devices for two consecutive nights, including at least two hours before and after getting into/out of bed. They slept in their home environment and filled out a sleep diary each morning documenting their perceived time of sleep onset, waking, wake events throughout the night and sleep rating.

Following each recording, the data was extracted from the device's native app, onto a spreadsheet. The following datapoints were extracted: time of sleep onset, time of waking, TSD, number of wake events throughout the night.

2.2. Statistical Analysis

TSD was compared to the sleep diaries (reference measure) using modified Bland-Altman (BA) plots. Both Pearson's (Pe.cc) and Spearman's (Sp.cc) correlation coefficients were calculated to assess the strength of relationship between TSD from the sleep diary and from each device. The absolute error for TSD across devices was assessed using the Wilcoxon signed-rank test, a non-parametric test for paired comparisons. The level of significance was adjusted for multiple comparisons, using the Bonferroni correction ($P<0.05/15$).

The impact of sources of inaccuracy (skin tone, wrist circumference and BMI) on absolute TSD error was assessed using linear regression models (10).

3. Results

3.1. Study Population

Table 1. Demographics of study population. BMI = body mass index, FPS = Fitzpatrick Scale, circ. = circumference.

Category	Number	Mean	Standard deviation
Male	19 (70%)	-	-
Female	8 (30%)	-	-
Age (years)	27	25	9.49
BMI (kg/m^2)	27	23.4	3.0
FPS	27	4.0	1.8
Wrist circ. (cm)	27	16.8	1.8

3.2. Connectivity

Despite continuous syncing, some devices reported connectivity issues, resulting in data loss. The device which missed the greatest number of nights was GV4/4S set 1, missing 27.8% of nights, and hence were the least reliable. Conversely, the WS did not drop any nights of data (0%), making it the most reliable.

Table 2. Data loss from connectivity issues

Device	Total nights worn	Nights lost	Total nights recorded

Garmin 4/4S set 1	54	15	39
Garmin 4 set 2	54	13	41
Fitbit Sense	54	3	51
Withings ScanWatch	54	0	54
Oura Ring Gen 3	52	1	51

3.3. Agreement with sleep diary

FS and OR3β demonstrated the greatest correlation with the sleep diary for TSD (Sp.cc = 0.76 and 0.75 respectively). Meanwhile, GV4 set 2 demonstrated the lowest correlation with the diary (Sp.cc=0.41).

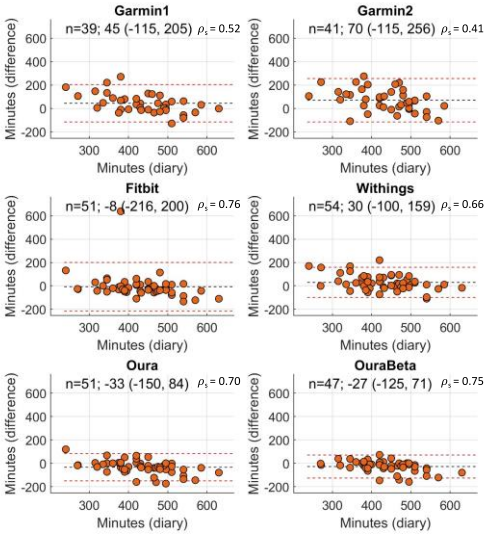

Figure 3: Bland Altman plots of total sleep duration
X-axis = total sleep duration documented in diary (minutes), Y axis = difference between device and diary (minutes). Documented as: n = number of nights recorded, bias, (lower limit of agreement, upper limit of agreement), ρs = Spearman's correlation coefficient. Garmin1 = Garmin Vivoactive 4/4S (set 1), Garmin2 = Garmin Vivoactive 4 (set 2), Fitbit = Fitbit Sense, Withings = Withings ScanWatch, Oura = Oura Ring Generation 3, Oura Beta = Oura Ring Generation 3 with beta software.

Compared to the sleep diary, GV4/4S set 1, GV4 set 2 and WS overestimated TSD on average. As demonstrated by the BA plots (Figure), this was by 45 minutes (n=39), 70 minutes (n=41) and 30 minutes (n=54) respectively. Meanwhile, FS, OR3 and OR3β underestimated sleep. This was by 8 minutes (n=51), 27 minutes (n=51) and 33 minutes (n=47) respectively. FS showed the lowest overall bias, followed by OR3β, WS, OR3, GV4/4S set 1 and GV4 set 2. Despite its low bias, FS demonstrated the largest LOA (416 minutes, n=51). It is important to acknowledge that this is the influence of a single outlier, which drastically overestimated TSD by 639 minutes. This was followed by GV4 set 2 (371 minutes, n=41), GV4/4S set 1 (320 minutes, n=39), WS (259 minutes, n=54), OR3 (234 minutes, n=51) and OR3β (196 minutes, n=47).

Using Wilcoxon signed-rank test (after Bonferroni correction), to compare absolute differences in TSD (Figure), GV4 set 2 was significantly different from all other devices, excluding GV4/4S set 1. Further, OR3β was significantly different from the standard OR3. All other pairwise comparisons for TSD were statistically insignificant (p>0.003).

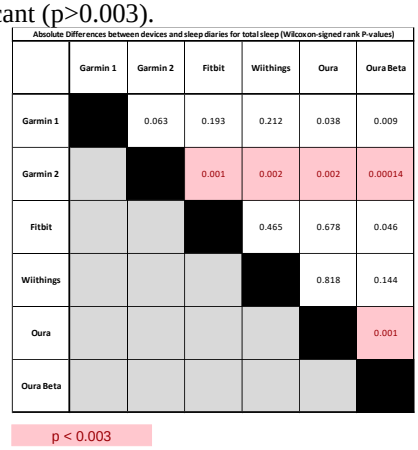

Figure 4: Heat map demonstrating p values from Wilcoxon signed-rank test to compare absolute differences in total sleep duration between each device pair. Level of significance has been adjusted for multiple comparisons using the Bonferroni coefficient (p<0.05/15). p<0.003 suggests statistically significant difference in absolute error between device pairings.

3.4. Sources of inaccuracies

Linear regression models broadly showed no associations (p>0.05) between absolute TSD error and skin tone, wrist circumference or BMI. For the WS, TSD error appeared to inversely correlate with BMI (p=0.026). However, this is likely to be a Type 1 error resulting from multiple analyses being ran on a relatively small sample (n=54).

Table 3. Linear regression of absolute error in total sleep duration and demographic variables. β = beta value, CI = confidence interval, FPS = Fitzpatrick scale, BMI = body mass index. P<0.05 highlighted in red.

Garmin V4/4S set 1	**β**	**95% CI**		**p value**
FPS	0.22	-0.11	0.54	0.184
Wrist circ.	-0.03	-0.45	0.39	0.888
BMI	-0.22	-0.58	0.14	0.214
Garmin V4 set 2				
FPS	0.27	-0.04	0.57	0.085
Wrist circ.	0.03	-0.33	0.39	0.852
BMI	-0.13	-0.46	0.20	0.420
Fitbit Sense				
FPS	0.14	-0.15	0.43	0.343

Wrist circ.	0.25	-0.07	0.56	0.124
BMI	0.11	-0.25	0.47	0.538
Withings ScanWatch				
FPS	0.21	-0.07	0.48	0.137
Wrist circ.	0.00	-0.30	0.30	0.975
BMI	-0.36	-0.67	-0.05	0.026
Oura Ring Gen 3				
FPS	-0.07	-0.36	0.21	0.616
Wrist circ.	-0.14	-0.45	0.18	0.394
BMI	-0.10	-0.38	0.18	0.482
Oura beta software				
FPS	-0.01	-0.31	0.28	0.936
Wrist circ.	-0.31	-0.69	0.08	0.117
BMI	-0.18	-0.46	0.10	0.211

4. Discussion

This study assessed agreement between total sleep duration measured in free-living conditions using consumer-grade wearable devices and total sleep duration reported in sleep diaries. Some of the better performing devices, such as OR3, FS and WS, demonstrated moderate to high correlation with sleep diaries for TSD and hence, would act as reliable replacements of diaries and surveys in field-based studies.

Devices showed wider LOA and more significant biases compared to previous studies conducted in sleep laboratories (11,12). Further, we have demonstrated significant variation in the reliability and accuracy of different models, emphasising the need for careful device selection when using sleep-monitoring wearables.

The difference between devices from the same manufacturer (GV4/4S set 1 vs GV4 set 2) was greater than expected, warranting further studies on reproducibility of results from the same device type.

Lastly, the improvement seen between the OR3β vs OR3 reemphasizes the importance of optimized software and large reference cohorts in sleep tracking.

All devices appeared to be broadly unaffected by potential sources of inaccuracy, including skin tone and wrist circumference, suggesting unbiased sleep tracking. The only significant finding (WS accuracy inversely correlated with BMI) was contrary to our original hypothesis. This may be the result of an unrepresentative reference cohort, or device fitting issues. However, it may also be a Type 1 error in running multiple analyses on a small cohort. Nonetheless, this finding warrants further exploration in a cohort with a greater BMI diversity.

Acknowledgments

MO is supported by BHF Accelerator Award AA/18/6/34223.

References

1. Itani O, Jike M, Watanabe N, Kaneita Y. Short sleep duration and health outcomes: a systematic review, meta-analysis, and meta-regression. Sleep Med. 2017 Apr 1;32:246–56.
2. Jike M, Itani O, Watanabe N, Buysse DJ, Kaneita Y. Long sleep duration and health outcomes: A systematic review, meta-analysis and meta-regression. Sleep Med Rev. 2018 Jun 1;39:25–36.
3. Lucey BP, McLeland JS, Toedebusch CD, Boyd J, Morris JC, Landsness EC, et al. Comparison of a single-channel EEG sleep study to polysomnography. J Sleep Res. 2016 Dec;25(6):625–35.
4. Grandner MA, Rosenberger ME. Chapter 12 - Actigraphic sleep tracking and wearables: Historical context, scientific applications and guidelines, limitations, and considerations for commercial sleep devices. In: Grandner MA, editor. Sleep and Health [Internet]. Academic Press; 2019 [cited 2023 Sep 15]. p. 147–57. Available from: https://www.sciencedirect.com/science/article/pii/B9780128153734000125
5. Smartwatch popularity continues as we enter the festive season [Internet]. [cited 2023 Mar 1]. Available from: https://www.kantar.com/inspiration/technology/smartwatch-popularity-continues-as-we-enter-the-festive-season
6. Crivello A, Barsocchi P, Girolami M, Palumbo F. The Meaning of Sleep Quality: A Survey of Available Technologies. IEEE Access. 2019 Nov 14;PP.
7. Gil E, Orini M, Bailón R, Vergara JM, Mainardi L, Laguna P. Photoplethysmography pulse rate variability as a surrogate measurement of heart rate variability during non-stationary conditions. Physiol Meas. 2010;31(9):1271–90.
8. Orini M, Guvensen G, Jamieson A, Chaturvedi N, Hughes AD. Movement , Sweating , and Contact Pressure as Sources of Heart Rate Inaccuracy in Wearable Devices. 2022 Comput Cardiol CinC. 2022;498:1–4.
9. Koerber D, Khan S, Shamsheri T, Kirubarajan A, Mehta S. The effect of skin tone on accuracy of heart rate measurement in wearable devices: a systematic review. J Am Coll Cardiol. 2022 Mar 8;79(9_Supplement):1990–1990.
10. Schober P, Boer C, Schwarte LA. Correlation Coefficients: Appropriate Use and Interpretation. Anesth Analg. 2018 May;126(5):1763.
11. Chinoy ED, Cuellar JA, Huwa KE, Jameson JT, Watson CH, Bessman SC, et al. Performance of seven consumer sleep-tracking devices compared with polysomnography. Sleep. 2021 May 14;44(5):zsaa291.
12. Miller DJ, Sargent C, Roach GD. A Validation of Six Wearable Devices for Estimating Sleep, Heart Rate and Heart Rate Variability in Healthy Adults. Sensors. 2022 Aug 22;22(16):6317.

Address for correspondence:
Sanjay Rajput
UCL Institute of Cardiovascular Science, 62 Huntley St, London WC1E 6DD
sanjay.rajput.20@ucl.ac.uk