RESEARCH ARTICLE

Sleep monitoring with the Apple Watch: comparison to a clinically validated actigraph [version 1; peer review: 2 approved with reservations, 1 not approved]

Sirinthip Roomkham [1], Michael Hittle [2], Joseph Cheung[3,4], David Lovell [1], Emmanuel Mignot[3], Dimitri Perrin [1]

[1]School of Electrical Engineering and Computer Science, Queensland University of Technology, Brisbane, Australia
[2]Department of Health Research and Policy, Stanford University School of Medicine, Palo Alto, CA, USA
[3]Stanford Center for Sleep Sciences and Medicine, Stanford University, Palo Alto, CA, USA
[4]Division of Pulmonary Medicine, Mayo Clinic, Jacksonville, FL, USA

v1 **First published:** 29 May 2019, **8**:754 (https://doi.org/10.12688/f1000research.19020.1)

Latest published: 29 May 2019, **8**:754 (https://doi.org/10.12688/f1000research.19020.1)

Abstract

Background: We investigate the feasibility of using an Apple Watch for sleep monitoring by comparing its performance to the clinically validated Philips Actiwatch Spectrum Pro (the gold standard in this study), under free-living conditions.
Methods: We recorded 27 nights of sleep from 14 healthy adults (9 male, 5 female). We extracted activity counts from the Actiwatch and classified 15-second epochs into sleep/wake using the Actiware Software. We extracted triaxial acceleration data (at 50 Hz) from the Apple Watch, calculated Euclidean norm minus one (ENMO) for the same epochs, and classified them using a similar algorithm. We used a range of analyses, including Bland-Altman plots and linear correlation, to visualize and assess the agreement between Actiwatch and Apple Watch.
Results: The Apple Watch had high overall accuracy (97%) and sensitivity (99%) in detecting actigraphy-defined sleep, and adequate specificity (79%) in detecting actigraphy defined wakefulness. Over the 27 nights, total sleep time was strongly linearly correlated between the two devices (r=0.85). On average, the Apple Watch over-estimated total sleep time by 6.31 minutes and under-estimated Wake After Sleep Onset by 5.74 minutes. The performance of the Apple Watch compares favorably to the clinically validated Actiwatch in a normal environment.
Conclusions: This study suggests that the Apple Watch could be an acceptable alternative to the Philips Actiwatch for sleep monitoring, paving the way for larger-scale sleep studies using Apple's consumer-grade mobile device and publicly available sleep classification algorithms. Further study is needed to assess longer-term performance in natural conditions, and against polysomnography in clinical settings.

Keywords
Apple Watch, Sleep Study, Actigraphy

Open Peer Review

Reviewer Status ? ✗ ?

	Invited Reviewers		
	1	**2**	**3**
version 1 published 29 May 2019	? report	✗ report	? report

1 **Sean P. A. Drummond**, Monash University, Clayton, Australia

2 **Lillian Skeiky**, Washington State University (WSU), Spokane, USA
Devon A. Hansen, Washington State University (WSU), Spokane, USA

3 **Vincenzo Natale**, University of Bologna, Bologna, Italy

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Dimitri Perrin (dimitri.perrin@qut.edu.au)

Author roles: Roomkham S: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Hittle M**: Software, Writing – Review & Editing; **Cheung J**: Conceptualization, Formal Analysis, Methodology, Writing – Review & Editing; **Lovell D**: Conceptualization, Formal Analysis, Methodology, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Mignot E**: Conceptualization, Methodology, Writing – Review & Editing; **Perrin D**: Conceptualization, Data Curation, Formal Analysis, Methodology, Software, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: Devices were funded by internal funds from the Queensland University of Technology (equipment grant from the Science and Engineering Faculty for the Philips Actiwatch Spectrum Pro and Actiware Software, ECARD grant for the Apple Watch).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Roomkham S, Hittle M, Cheung J *et al.* **Sleep monitoring with the Apple Watch: comparison to a clinically validated actigraph [version 1; peer review: 2 approved with reservations, 1 not approved]** F1000Research 2019, **8**:754 (https://doi.org/10.12688/f1000research.19020.1)

First published: 29 May 2019, **8**:754 (https://doi.org/10.12688/f1000research.19020.1)

Introduction

Good sleep is vital for our health and wellbeing. Without it, our body and mind function poorly, with consequences that include risk of obesity[1], diabetes[2] and cardiovascular disease[3]. Polysomnography (PSG) is currently the gold standard method to monitor sleep. During PSG, subjects spend a night in a dedicated sleep lab, hooked up to a range of devices to measure physiological signals. However, information-rich PSG is expensive, laborious and intrusive. Because subjects are attached to many electrodes, sleep is disturbed by the recording, so that it is better at looking at qualitative abnormalities such as sleep apnea or narcolepsy.

Widespread adoption of smartphones, smartwatches and fitness devices opens up new opportunities for monitoring sleep and physical activity. Many of these consumer devices come with a built-in accelerometer, gyroscope, magnetometer, and in some cases, a heart rate sensor—more sensors than wrist-worn actigraphs, which only use an accelerometer to measure movement to infer sleep and wake states. Nevertheless, actigraphs are regarded as useful tools in clinical practice when measuring sleep and wakefulness in normal living environments, especially in the context of assessing regularity and overall diurnal fluctuations[4–7].

The affordability of consumer-grade mobile devices has driven their popularity and the abundance of applications (“apps”) available—including apps for health and sleep. These devices have the potential to be used in sleep studies and to inform clinical diagnosis. However, their performance needs to be properly validated in comparison to accepted methodologies such as actigraphy and PSG. This is difficult because the sleep classification methods (e.g., algorithms and analysis techniques) used in consumer-grade devices are proprietary, so the relationship between the underlying physiological measurements and the sleep state reported by an app is unclear.

We chose to investigate the feasibility of the Apple Watch as a sleep monitoring device because the manufacturer allows software developers to access the device’s triaxial accelerometer data. To the best of our knowledge, this is the first study to compare the Apple Watch against an actigraph.

Methods

Evaluation framework

Figure 1 sets out the framework for validating the Apple Watch against an actigraph. First, raw acceleration measurements are collected from each device and transformed into activity counts for the actigraph, and the “Euclidean Norm Minus One” (ENMO) for the Apple Watch. We then explore statistical relationships between activity counts and ENMO. Next, we determine a threshold to classify each epoch of ENMO values as “sleep” or “wake” and evaluate the sleep outputs of the two devices using Pearson correlation and Bland-Altman plots.

Participants

In total, 14 healthy participants (9 males, 5 females) were recruited by word of mouth and direct approach. Data were recorded from April to May 2018, participants wore the two devices for two consecutive nights at home. The inclusion criterion for participants was age of at least 18 years old. Exclusion criteria were a previously diagnosed sleep disorder, or any condition that would lead to difficulty/discomfort while wearing the devices. The Queensland University of Technology Human Research Ethics Committee (#1800000242) approved all procedures, and all participants gave their signed consent prior participating the study. They were asked to wear both the Apple Watch and Actiwatch on their non-dominant wrist for two consecutive nights and sleep as they normally would. All wearable devices were then returned for data extraction. One participant forgot to charge the Apple Watch, so we lost one night of data and were left with 27 nights from 14 participants.

This study was designed as a proof of concept for whether it is possible to use the Apple Watch for sleep monitoring. Power calculations are appropriate when the distribution of the underlying data is known (and ideally, normal). At that point in time, this was not applicable given that there was no prior study for us to confidently characterize the distribution of the differences between the measurements obtained using different platforms. Sample size was therefore determined pragmatically[8].

Wrist-worn devices

Table 1 shows specifications of the two wrist-worn devices used in this study: the Apple Watch Series 1 (Apple Inc., California, United States) and the Actiwatch Spectrum Pro (Philips, Bend OR). The Apple Watch has limited data storage, so data was downloaded to an iPhone via Bluetooth. We used the Core Motion Framework to develop an app to record triaxial accelerometer data at 50 Hz. The Actiwatch Spectrum Pro is a clinical-grade actigraph used for sleep and activity monitoring. It samples accelerometer data at 32 Hz and we set a 15-second epoch for processing this raw data. Processed data were downloaded using Philips’ Actiware Software (version 6.0.9). Outputs were activity counts and sleep/wake stage at each epoch.

Data processing and analysis

Accelerometer. Raw acceleration data from the Apple Watch was downloaded and processed using R statistical software. We calculated ENMO, the Euclidean Norm (magnitude) of the triaxial acceleration vector A = (A_x, A_y, A_z) minus 1 gravitational unit. ENMO is used widely in physical activity and sleep monitoring[9–13] and defined as:

$$\text{ENMO (A)} = \sqrt{A_x^2 + A_y^2 + A_z^2}\ \text{-}1$$

We compared the mean ENMO of each 15-second epoch against activity counts from the actiwatch. We ensured that the clocks of both devices were synchronized and compared recordings of vigorous movement applied simultaneously to both devices to check that each devices’ timestamps were in sync.

Sleep algorithm. Philips’ Actiware software computes the total activity counts at epoch *e* using a weighted sum:

$$\text{Total\_counts}(e) = 0.04\sum_{i=-8}^{i=-5} a_{e+i} + 0.2\sum_{i=-4}^{i=-1} a_{e+i} + 4a_e + 0.04\sum_{i=1}^{i=4} a_{e+i} + 0.2\sum_{i=5}^{i=8} a_{e+i}$$

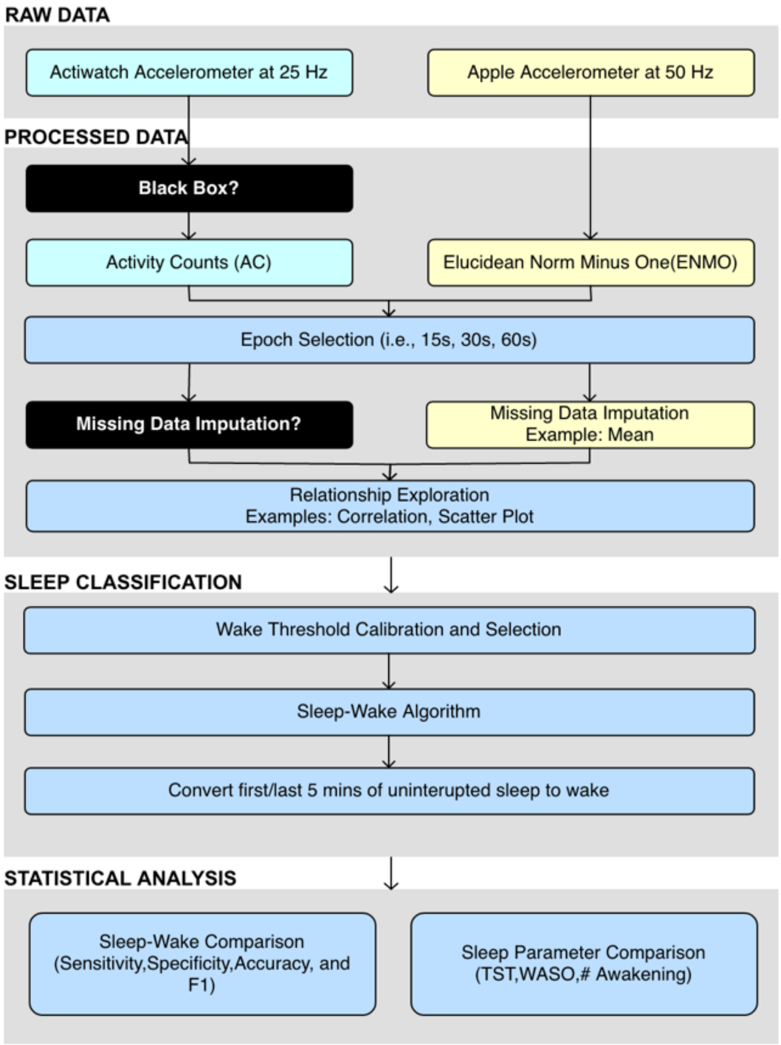
RAW DATA
Actiwatch Accelerometer at 25 Hz
Apple Accelerometer at 50 Hz
PROCESSED DATA
Black Box?
Activity Counts (AC)
Elucidean Norm Minus One(ENMO)
Epoch Selection (i.e., 15s, 30s, 60s)
Missing Data Imputation?
Missing Data Imputation
Example: Mean
Relationship Exploration
Examples: Correlation, Scatter Plot
SLEEP CLASSIFICATION
Wake Threshold Calibration and Selection
Sleep-Wake Algorithm
Convert first/last 5 mins of uninterupted sleep to wake
STATISTICAL ANALYSIS
Sleep-Wake Comparison
(Sensitivity,Specificity,Accuracy, and
F1)
Sleep Parameter Comparison
(TST,WASO,# Awakening)

Figure 1. Evaluation framework for wrist-worn accelerometer devices.

Table 1. Device specifications.

Specification	Apple Watch	Actiwatch
Model	Series 1	Spectrum Pro
Price	$AU 359-559	$AU 3470
Battery life	1 day	50 days
Sensor	Accelerometer, Gyroscope, Heart rate	Accelerometer, Light sensor
Accelerometer sampling rate	50 Hz	32 Hz
Recording Time (Accelerometer)	3 days	50 days

where a_e is the activity count of epoch e. We used the shortest (15-second) epoch so that data could be converted to longer epochs if required. Each epoch was classified as sleep if its total activity counts were less than or equal to a threshold; epochs with counts above the threshold were classified as wake. Our study used a low and medium threshold (20 and 40, respectively) derived from Actiware 6.0.9 software.

Statistical analysis. The R program (version 3.3.2) is used for statistical analysis and visualization. To measure agreement between Apple Watch and Actiwatch, we used a range of statistical methods including:

i Calculating the Pearson Correlation between total activity counts and ENMO.

ii Using a receiver operating characteristic (ROC) analysis, taking the Actiwatch as ground truth.

iii Using Bland-Altman plots to measure the agreement between two measurements by quantifying the mean bias and constructing an agreement interval.

iv Computing the overall accuracy and performance the devices' sleep/wake classifications using a confusion matrix, again taking the Actiwatch as ground truth.

The confusion matrix of sleep-wake classifications has four outcomes: true positive (TP), true negative (TN), false positive (FP), and false negative (FN), with sleep positive, and wake negative. Using these outcomes, we calculated accuracy, sensitivity, specificity, and F1 score as defined in Table 2[14].

Table 2. Classification performance statistics.

Measure	Formula
Accuracy	(TP +TN)/(TP+TN+FN+FP)
Sensitivity or Recall	TP/(TP+FN)
Specificity	TN/(TN+FP)
Precision	TP/(TP+FP)
F1 Score	2*Precision*Recall/ (Precision+Recall)

We also calculated measures of sleep quality of interest in sleep studies: total sleep time (TST), wake after sleep onset (WASO), and number of awakenings. TST is the total duration of epochs classified as sleep; WASO is the total duration of wake epochs. Number of awakenings is the number of wake events of at least 30 seconds duration[15].

Results

Sleep-wake agreement

Figure 2 displays measurements over the first night of randomly selected participant: overall patterns of Actiwatch and Apple Watch are very similar with clearly aligned periods of movement. For example, around 22:10, both activity counts and ENMO were quite active with high peaks, then both signals gradually declined. During a sleep period from 02:00–02:30, both features remained steady with no obvious movements. Raw measurements are available as *Underlying data*[16].

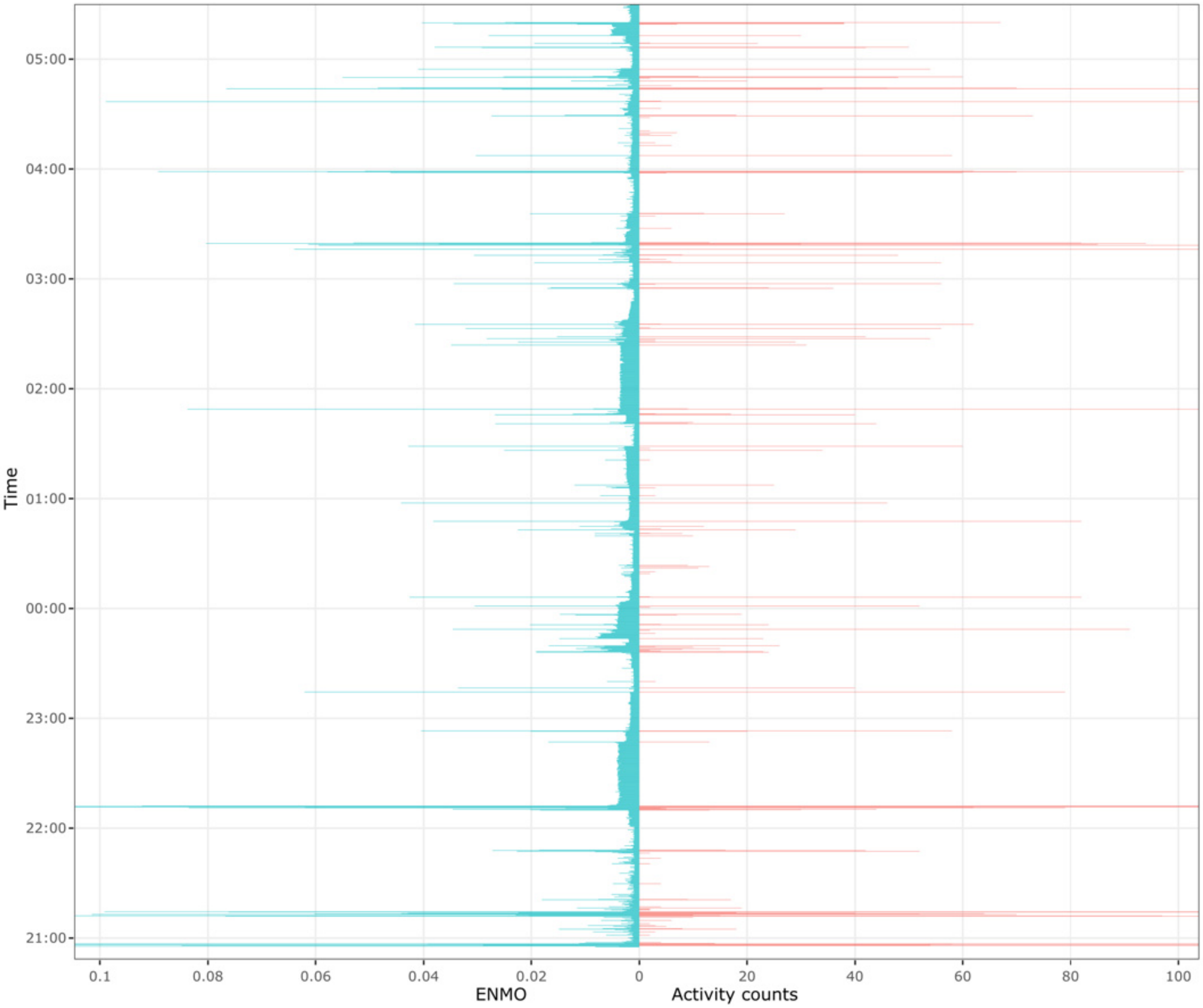

Figure 2. One night of measurements at 15 second epochs. ENMO is shown in turquoise (left) and Activity Counts in pink (right).

Pearson correlation was computed to assess the relationship between activity counts and ENMO shown in Figure 3. Overall, there was strong, positive correlation between activity counts and ENMO (r = 0.85, nights = 27, p<0.001). There were similar patterns across 15-second, 30-second, and 60-second epochs. We used 15-second-epochs in the remainder of this analysis.

Optimal wake threshold

Unlike the Actiwatch, there is no pre-established wake threshold for the Apple Watch ENMO data. To identify an optimal threshold, we isolated 11 nights without any missing data, and created all the possible training sets containing *k* nights, where *k* varies from 1 to 10. For instance, for *k* = 1 or *k* = 10 there are 11 such training sets, while for *k* = 5 or *k* = 6 there are 462 sets. In total, we considered 2046 distinct training sets. For each of these sets, we identified the threshold that maximized the F1 score on the nights not used for training. Figure 4 shows how this threshold converges to 0.0608 when comparing against the medium Actiwatch threshold. This 'optimal' threshold was used for the rest of our analysis. A similar pattern was observed for the low Actiwatch threshold, with the Apple Watch threshold converging to 0.0523. We also measured the impact of the threshold choice through a ROC curve for the medium Actiwatch setting (Figure 5).

Overall, there is good agreement between the Apple Watch and Actiwatch for both medium and low setting thresholds. Significantly, the ability to detect sleep (sensitivity) are higher than 98%. However, the ability to detect awake ranges from 60% to 79% for both thresholds. The overall accuracy and F1 score were consistent for both settings. The results are summarized in Table 3.

Bland-Altman plots

Figure 6 plots the difference versus the mean of TST, WASO and numbers of awakenings for Apple Watch and Actiwatch. We used a 15-second epoch and medium threshold for our main analysis. For TST, most nights were within the limits of agreement and close to perfect agreement (the black line, Figure 6a)—three nights of TST were in perfect agreement. Only one night was outside the agreement intervals and was overestimated TST by 30 minutes. The overall bias of TST was 6.31 minutes.

In terms of WASO, differences fall mostly within the levels of agreements (Figure 6b)—two nights were in perfect agreement. One night stood outside the agreements which an underestimation of WASO by 30 minutes. This night was the same night that we found in TST. The overall bias of estimating WASO was -5.74 minutes.

For the total number of awakenings, the overall bias was -4.56. Only one night was in perfect agreement but all nights lie within the upper and lower agreement levels (Figure 6c).

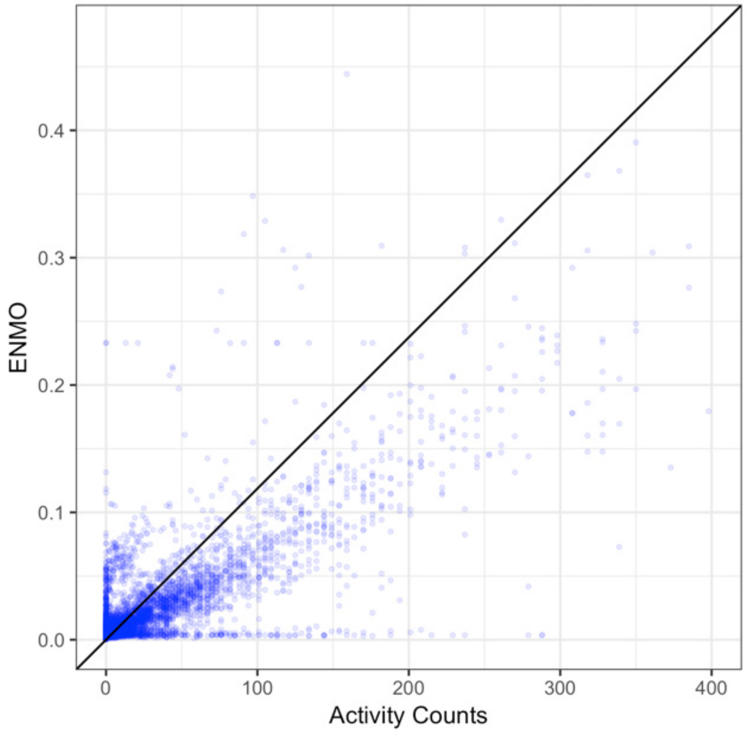

Figure 3. Apple Watch mean ENMO versus Actiwatch activity counts for all 15s-epochs over 27 nights of data. The diagonal line is the standardized major axis fit[17] of activity counts and mean ENMO, constrained to pass through the origin.

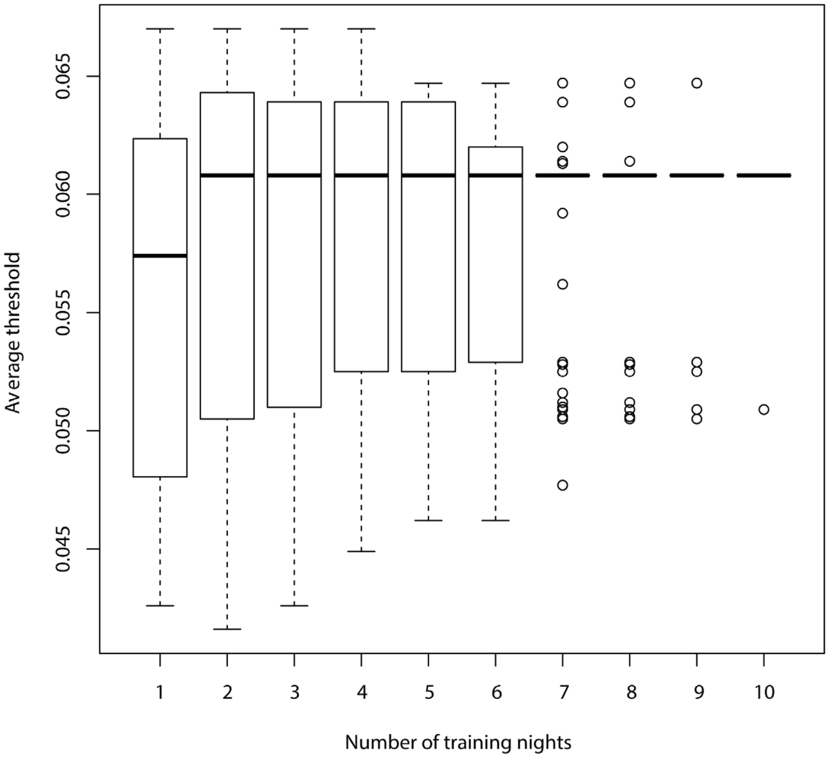

Figure 4. Box plot of the distributions of 'optimal' thresholds based on 1–10 nights of training data. As the number of training nights increase, the distributions converge around a value of 0.0608.

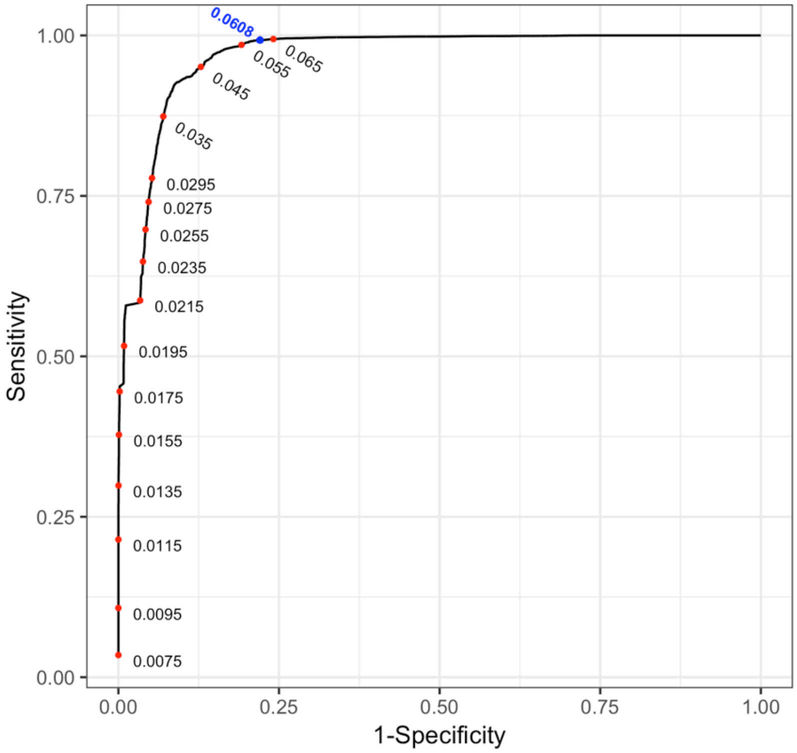

Figure 5. ROC curve for varying threshold, Apple Watch threshold taking the Actiwatch at a 15 s-epoch medium threshold as truth.

Discussion

This study compares measurements of sleep and wake obtained from Apple Watch, a consumer-grade device, and Philips Actiwatch Spectrum Pro, a clinically validated actigraphy device, in a set of healthy adults. We found that ENMO and activity counts were highly correlated. By using a combination of ten-night training data and ROC plot, we identified an optimal threshold for Apple Watch ENMO data in comparison to the Actiwatch for both medium and low settings (Table 3).

Table 3. Overall performance in accuracy, sensitivity, specificity and F1 in comparison with the Actiwatch in medium and low setting mode.

Variable	Medium threshold	Low threshold
Accuracy	97.11% ± 0.53%	93.53% ± 0.86%
Sensitivity	99.28% ± 0.21%	98.97% ± 0.35%
Specificity	78.94% ± 1.95%	63.50% ± 2.40%
F1 Score	98.35% ± 0.32%	96.16% ± 0.57%

In the medium threshold setting, the sleep parameters of TST, WASO, and number of awakenings were comparable to that of the Actiwatch with no significant differences. The discrepancy between the two measurements appeared to be clinically acceptable as a difference of TST and WASO did not exceed 30 minutes[6,18]. The Apple Watch performs best in comparison to the Actiwatch at medium threshold, consistent with Quante's recommendation[19].

To the best of our knowledge, this is the first study to evaluate the Apple Watch, a popular-consumer grade device, against a clinical-grade actigraph device for sleep monitoring. We have compared the two devices at high resolution (i.e., 15-second epochs) and low-resolution sleep parameters (i.e., TST, WASO, and number of awakenings). A similar study compared a consumer fitness tracking device (Fitbit charge HR) against Philips' Actiwatch 2: the accuracy of sleep parameters was good[20].

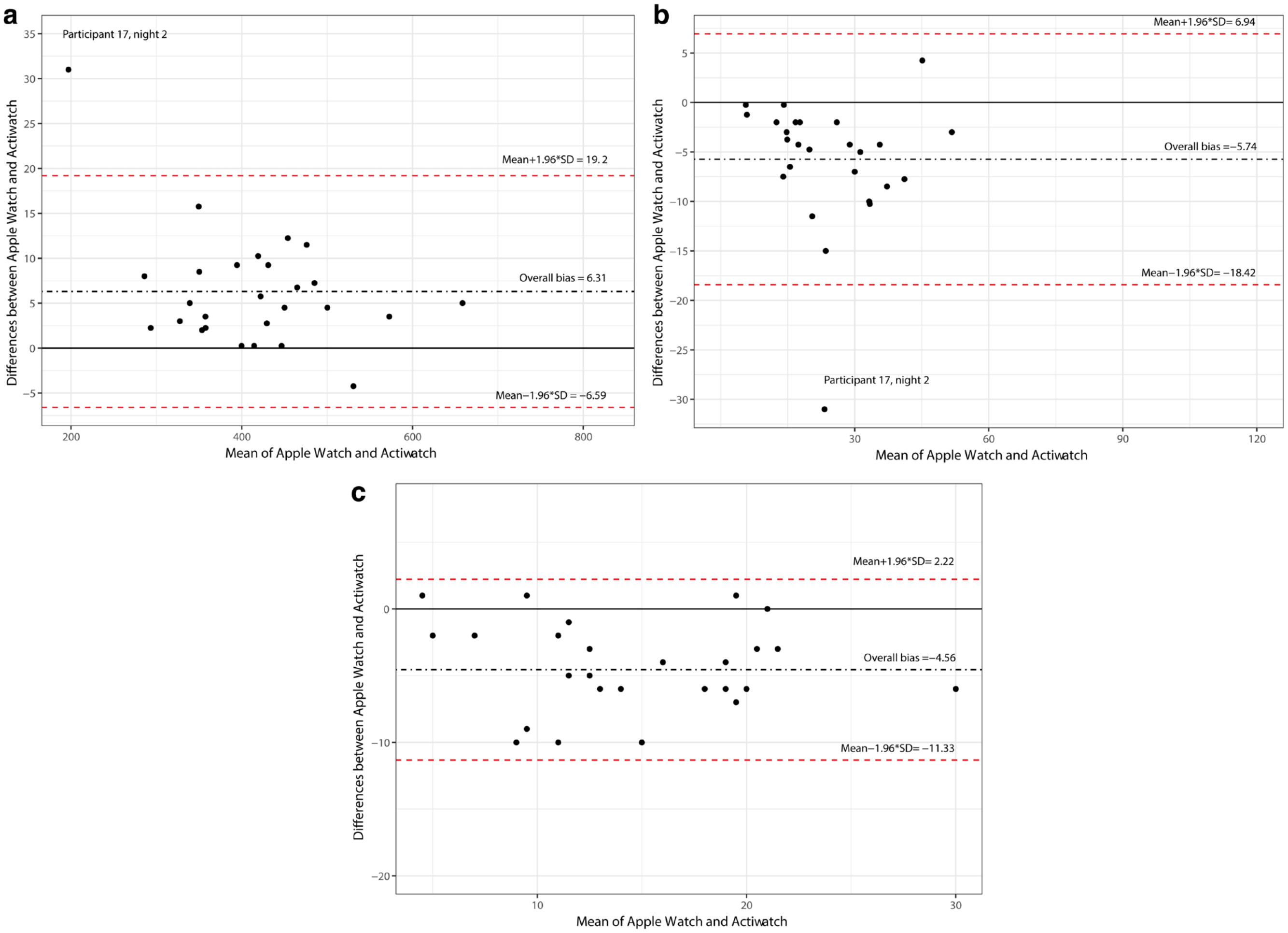

Figure 6. Bland-Altman plots show the agreement between Apple Watch and Actiwatch in medium threshold. The red dashed lines represent in the upper and lower agreement limits (95% confidence intervals). (**a**) Total sleep time (minutes). (**b**) Wake after sleep onset (minutes). (**c**) Number of awakenings.

However, Montgomery-Downs *et al.* suggested that both Fitbit and Actiwatch tended to have limited specificity[21]. Therefore, care must be paid in validation. Furthermore, both studies compared only low-resolution sleep parameters due to the limited information of the type of feature, and proprietary sleep algorithm[22]. Our study provides additional support for high-resolution data (i.e., ENMO), which could be further used in a sleep-wake algorithm validated against gold standard PSG.

We note that ENMO and activity counts are dominated by very small values (i.e., points near or at (0, 0)), as shown in Figure 3. This is a consequence of doing a study in which participants are moving very little for relatively long periods. While this is reasonable in assessing whether the Apple Watch produces comparable results to the Actiwatch for sleep monitoring, this study cannot draw conclusions about whether the Apple Watch is comparable to the Actiwatch across a broader range of activity levels (e.g., during exercise). The main benefit of assessing sleep using a well-known consumer-grade device lies in increasing the opportunity for longitudinal studies in a wider population. Over 30 million Apple Watches have been sold[23]. The availability of advanced sensor technology in smart watches (e.g., heart-rate sensors) opens up possibilities for improved sleep monitoring with consumer wearable devices.

We faced some practical challenges in implementing sleep monitoring on the Apple Watch, including constraints of power, memory management, data transfer, and sensor capabilities. In total, 16 nights had missing data from the Apple Watch. In each of these nights, data was missing in one contiguous block of 20–60 minutes in duration. We needed to impute data based on the average of previous and next available data. At this stage, the cause of the missing data is still to be determined.

Recording was limited to two consecutive nights per participant due to memory and data transfer constraints. With these limitations in mind, we suggest future studies are needed to carefully monitor and deal with data loss where more nights of recording are investigated. Lastly, our study assessed only one sleep-wake algorithm based on the Cole-Kripke algorithm[17]; future studies could assess other sleep-wake algorithms that use a combination of weighted sum activity with the cut-off wake threshold to classify sleep or wake stages[15].

Conclusion

Our study lays down a foundation for using accelerometer data of consumer-grade devices for sleep monitoring. Our experiments show that Apple Watch provides sleep measures comparable to the Philips Actiwatch, a clinical gold standard, with greatest similarity at the medium threshold of activity counts. These findings increase our confidence in using the consumer grade Apple Watch for sleep monitoring and open up possibilities for much larger-scale sleep studies. We also hope this work will serve as a basis for sleep clinicians in the use of data extracted from this device in their patients.

To further our research, we now intend to compare the Apple Watch against PSG and consider incorporating other types of physiological sensors from consumer-grade devices (e.g. heart-rate sensor). These findings add to a growing body of literature on the use of consumer-based accelerometer for sleep studies and will assist other researchers in establishing validation parameters for the use of other types of consumer devices.

Data availability

QUT Research Data Finder: Sleep Data. https://doi.org/10.25912/5cc28f62e81ad[16].

Underlying data are contained within SleepDataset.zip. There are 27 csv files in this archive, with each file corresponding to a single night of data. The files have four columns: timestamp, Actiwatch activity counts, Actiware classification (1 for wake, 0 for sleep), and ENMO value calculated from the Apple Watch data. Each row corresponds to the data for a 15-second epoch. Dates have been modified to preserve privacy, with times are unchanged.

Grant information

Devices were funded by internal funds from the Queensland University of Technology (equipment grant from the Science and Engineering Faculty for the Philips Actiwatch Spectrum Pro and Actiware Software, ECARD grant for the Apple Watch).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Watanabe M, Kikuchi H, Tanaka K, *et al.*: **Association of short sleep duration with weight gain and obesity at 1-year follow-up: a large-scale prospective study.** *Sleep.* 2010; **33**(2): 161–167.
PubMed Abstract | Publisher Full Text | Free Full Text

2. Spiegel K, Knutson K, Leproult R, *et al.*: **Sleep loss: a novel risk factor for insulin resistance and Type 2 diabetes.** *J Appl Physiol (1985).* 2005; **99**(5): 2008–2019.
PubMed Abstract | Publisher Full Text

3. Kasasbeh E, Chi DS, Krishnaswamy G: **Inflammatory aspects of sleep apnea and their cardiovascular consequences.** *South Med J.* 2006; **99**(1): 58–67; quiz 68-9, 81.
PubMed Abstract | Publisher Full Text

4. Natale V, Léger D, Martoni M, *et al.*: **The role of actigraphy in the assessment of primary insomnia: a retrospective study.** *Sleep Med.* 2014; **15**(1): 111–115.
PubMed Abstract | Publisher Full Text

5. Shin M, Swan P, Chow CM: **The validity of Actiwatch2 and SenseWear armband compared against polysomnography at different ambient temperature conditions.** *Sleep Sci.* 2015; **8**(1): 9–15.
PubMed Abstract | Publisher Full Text | Free Full Text

6. de Zambotti M, Baker FC, Colrain IM: **Validation of Sleep-Tracking Technology Compared with Polysomnography in Adolescents.** *Sleep.* 2015; **38**(9): 1461–1468.
PubMed Abstract | Publisher Full Text | Free Full Text

7. Sadeh A: **The role and validity of actigraphy in sleep medicine: an update.** *Sleep Med Rev.* 2011; **15**(4): 259–267.
PubMed Abstract | Publisher Full Text

8. Julious SA: **Sample size of 12 per group rule of thumb for a pilot study.** *Pharm Stat.* 2005; **4**(4): 287–291.
Publisher Full Text

9. Bakrania K, Yates T, Rowlands AV, *et al.*: **Intensity Thresholds on Raw Acceleration Data: Euclidean Norm Minus One (ENMO) and Mean Amplitude Deviation (MAD) Approaches.** *PLoS One.* 2016; **11**(10): e0164045.
PubMed Abstract | Publisher Full Text | Free Full Text

10. van Hees VT, Gorzelniak L, Dean León EC, *et al.*: **Separating movement and gravity components in an acceleration signal and implications for the assessment of human daily physical activity.** *PLoS One.* 2013; **8**(4): e61691.
PubMed Abstract | Publisher Full Text | Free Full Text

11. van Hees VT, Fang Z, Langford J, *et al.*: **Autocalibration of accelerometer data for free-living physical activity assessment using local gravity and temperature: an evaluation on four continents.** *J Appl Physiol (1985).* 2014; **117**(7): 738–744.
PubMed Abstract | Publisher Full Text | Free Full Text

12. Rowlands AV, Yates T, Davies M, *et al.*: **Raw Accelerometer Data Analysis with GGIR R-package: Does Accelerometer Brand Matter?** *Med Sci Sports Exerc.* 2016; **48**(10): 1935–1941.
PubMed Abstract | Publisher Full Text

13. Cheung J, Zeitzer JM, Lu H, *et al.*: **Validation of minute-to-minute scoring for sleep and wake periods in a consumer wearable device compared to an actigraphy device.** *Sleep Science Practice.* 2018; **2**(1): 11.
Publisher Full Text

14. Saito T, Rehmsmeier M: **The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets.** *PLoS One.* 2015; **10**(3): e0118432.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Galland BC, Kennedy GJ, Mitchell EA, *et al.*: **Algorithms for using an activity-based accelerometer for identification of infant sleep-wake states during nap studies.** *Sleep Med.* 2012; **13**(6): 743–751.
PubMed Abstract | Publisher Full Text

16. Roomkham S, Michael H, Joseph C, *et al.*: **Sleep study comparison using the Apple Watch and the Philips Actiwatch.** *Queensland University of Technology.* 2019.

17. Sadeh A, Lavie P, Scher A, *et al.*: **Actigraphic home-monitoring sleep-disturbed and control infants and young children: a new method for pediatric assessment of sleep-wake patterns.** *Pediatrics.* 1991; **87**(4): 494–499.
PubMed Abstract

18. Werner H, Molinari L, Guyer C, *et al.*: **Agreement rates between actigraphy, diary, and questionnaire for children's sleep patterns.** *Arch Pediatr Adolesc Med.* 2008; **162**(4): 350–358.
PubMed Abstract | Publisher Full Text

19. Quante M, Kaplan ER, Cailler M, *et al.*: **Actigraphy-based sleep estimation in adolescents and adults: a comparison with polysomnography using two scoring algorithms.** *Nat Sci Sleep.* 2018; **10**: 13–20.
PubMed Abstract | Publisher Full Text | Free Full Text

20. Lee HA, Lee HJ, Moon JH, *et al.*: **Comparison of Wearable Activity Tracker with Actigraphy for Sleep Evaluation and Circadian Rest-Activity Rhythm Measurement in Healthy Young Adults.** *Psychiatry Investig.* 2017; **14**(2): 179–185.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Montgomery-Downs HE, Insana SP, Bond JA: **Movement toward a novel activity monitoring device.** *Sleep Breath.* 2012; **16**(3): 913–917.
PubMed Abstract | Publisher Full Text

22. Roomkham S, Lovell D, Cheung J, *et al.*: **Promises and Challenges in the Use of Consumer-Grade Devices for Sleep Monitoring.** *IEEE Rev Biomed Eng.* 2018; 11: 53–67.
PubMed Abstract | Publisher Full Text

23. Idc: **IDC Forecasts Shipments of Wearable Devices to Nearly Double by 2021 as Smart Watches and New Product Categories Gain Traction.** 2017.
Reference Source

Open Peer Review

Current Peer Review Status: ? ✗ ?

Version 1

Reviewer Report 27 August 2019

https://doi.org/10.5256/f1000research.20845.r49255

? **Vincenzo Natale**
Department of Psychology, University of Bologna, Bologna, Italy

The aim of this work was to compare the measurements of sleep and wake obtained from Apple Watch and Philips Actiwatch Spectrum. To this aim, 14 participants (9 males) were asked to wear on non-dominant wrist both equipments for two consecutive nights. On the whole results show that Apple Watch performed well.

The manuscript is potentially interesting. However, there are several limitations and features that Authors should better explicit. The sample size is relatively small, but above all the number of nights is too small. Several researches show that actigraphy output reaches stability after at least seven nights.

I feel that it is still possible to improve the manuscript. For example, Authors did not show the mean age of the sample. Authors did not explain how they calculated sleep onset. Authors did not clearly explain why they adopted a sampling rate of 15 seconds when usually, to asses sleep by actigraph, is adopted a 1 minute epoch.

In Figure 1 there is a mistake. It is reported 25 Hz but should be 32 Hz.

Pearson correlations are not useful. Do Authors really think that when a wrist is moving one of two devices could record no movements?

The interpretations of Bland Altman plot are a little bit questionable. Six minutes for total sleep time are few, I agree, representing more or less a variation of around 2%. However, 30 minutes for wake after sleep onset are not few because they correspond more or less to a variation of 75%. In other words, Authors have developed an algorithm very good to assess sleep (sensitivity) but less performing to evaluate wake (specificity). About this, data reported in Table 3 are very unusual because changing threshold leads to an increase in both sensitivity and specificity. Usually when one increases, the other one decreases.

A last note is about the Cole-Kripke quotation at page 9. In my memory, such algorithm was implemented for another brand of actigraph (Ambulatory Monitoring). Are Authors sure about this?

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Sleep and individual differences in circadian rhtyhms

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 19 August 2019

https://doi.org/10.5256/f1000research.20845.r52113

✗ **Lillian Skeiky**
Sleep and Performance Research Center (SPRC), Elson S. Floyd College of Medicine, Washington State University (WSU), Spokane, WA, USA
Devon A. Hansen
Sleep and Performance Research Center (SPRC), Elson S. Floyd College of Medicine, Washington State University (WSU), Spokane, WA, USA

Over the last 10 years, a large number of commercially available wearable activity trackers have been offered to the public due to the growing request for individualized health monitoring. While these commercially available devices certainly have a number of advantages compared to traditional wrist actigraphy, such as accessibility and affordability, their general validity remains unclear. Roomkham and colleagues investigated the feasibility of sleep monitoring using the commercially available Apple Watch against the Philips Actiware Spectrum pro, a research grade wrist accelerometer arguing that the performance of the Apple Watch compared favorably to the Actiwatch. This line of research is promising and should absolutely be pursued. That said, we have serious reservations regarding the methodology and analyses conducted by Roomkham and colleagues which unfortunately limits comparison between devices.

Our specific comments are outlined below:

Abstract:

1. The authors should remove the line reading, “The performance of the Apple Watch compares favorably to the clinically validated Actiwatch in a normal environment”. This is not an accurate statement as it implies that the Apple Watch has been clinically validated.

2. The authors should remove the line reading, “This study suggests that the Apple Watch could be an acceptable alternative to the Philips Actiwatch for sleep monitoring…” for the same reason stated in the above point.

Introduction:

1. In the first paragraph of the introduction, polysomnography (PSG) is described as qualitative for examination of OSA and narcolepsy. While we agree that PSG has certain limitations, this statement should be removed as sleep disorders are quantifiably determined using American Academy of Sleep Medicine guidelines.

2. The authors state that the overall aim of this study is to compare the Apple Watch against an actigraph. In actuality, the authors are comparing a variable (ENMO) based off of raw acceleration measurements produced by the Apple Watch, which should be clarified by the authors. In addition, the authors should explicitly state that they are comparing a commercial device against a validated research grade device.

Methods:

Evaluation framework:

1. Further describe Euclidean Norm Minus One (ENMO) and how this measure is comparable to Philips’ sleep algorithm.

2. It does not appear that sleep diaries were used in this study. Sleep diaries are integral in this line of work and are necessary for determining a number of sleep/wake variables in field-based research. Without any other supplementary information, the only reliable variable that can be used is Time in Bed. Was any supplemental information (e.g., sleep diaries, call-in times) collected from study participants? If so, this information should be presented. If not, this should be included as a limitation.

Participants:

1. Additional information is needed about research participants. Participants are described as “healthy”, but no further details are provided; indicate how this was ascertained. Provide additional information on how subjects were screened in general (questionnaires, physical exam, etc.)?

2. The authors should justify the short recording period. Traditionally, devices are worn for days to weeks at a time. For reference, a recently published abstract also using the Apple Watch had participants wear watches for > 1 month[1].

Wrist-worn devices:

1. Did subjects have access to the sleep statistics from the Apple Watch? The Actiwatch does not provide any feedback. It is possible that the Apple Watch could have been an unintentional intervention if they received feedback. Please address.

2. Please explain the Core Motion Framework in greater detail and describe the app used to record accelerometer data.

3. Please describe how the Actiwatch data were processed and describe data cleaning methods, if any.

4. How were the devices worn on the wrist? Meaning, was device placement consistent for all participants? If not, this need to be addressed as a limitation.

Data processing and analysis:

1. See the point above regarding ENMO.

2. Further explain why 15-second epochs were used as 30-second epochs are standard in sleep scoring.

Statistical analysis:

1. Please describe TST and WASO as “actigraphically determined TST and WASO”. TST and WASO can only be accurately determined using PSG. In regards to field studies, a sleep diary is needed to calculate these variables.

2. Sleep Onset Latency provides useful information and should also be reported.

3. The citation used as the basis to quantify the number of wake events is from a study done in infants (Galland *et al*., 2012[2]). Please use a reference(s) in a more relevant study in healthy adults.

Results:

1. The authors imputed missing data from the Apple Watch using epochs prior to and after the missing period. Twenty to sixty minutes of a sleep period is a considerable amount of time during which significant changes may have occurred not reflected in the data used as a replacement. The authors should remove the imputed data and reanalyze using a technique designed to account for missing data.

Discussion:

1. The authors acknowledge that a major limitation of this study is the limited battery life and memory capacity of the Apple Watch. Due to these limitations, subjects wore the devices for just two consecutive nights, however this is not consistent with information from the manufacturer. Please clarify.

References

1. Eyal S, Mizrahi M, Baharav A: 1089 Popular Wearables May Facilitate Affordable Long-term Reliable Actigraphy. *Sleep*. 2018; **41** (suppl_1). Publisher Full Text

2. Galland BC, Kennedy GJ, Mitchell EA, Taylor BJ: Algorithms for using an activity-based accelerometer for identification of infant sleep-wake states during nap studies.*Sleep Med*. 2012; **13** (6): 743-51 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
No

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Sleep, sleep deprivation, wearables

We confirm that we have read this submission and believe that we have an appropriate level of expertise to state that we do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Reviewer Report 10 July 2019

https://doi.org/10.5256/f1000research.20845.r50110

Sean P. A. Drummond
Turner Institute for Brain and Behaviour, School of Psychological Sciences, Monash University, Clayton, Vic, Australia

In this this paper, the authors attempt to optimize an algorithm to convert accelerometer data from the Apple Watch to match that provided by a research grade actigraph, the Actiwatch Spectrum Pro. This is a proof of concept study, and thus has a fairly small sample size. Overall, the analyses show the authors are able to convert the Apple Watch data into the same epoch-by-epoch sleep/wake data as the Atiwatch with impressive sensitivity and specificity. Not surprisingly, then, the Total Sleep Time and Wake After Sleep Onset values from the Apple Watch also match the Actiwatch quite well. While I believe this algorithm development is a valuable addition to the literature, there are several areas where the manuscript could be improved. In particular, I believe the authors oversell their results and over state conclusions. There are also some areas where the reader would benefit from more detail or clarity. Overall, I am extremely supportive of this line of work and this paper. However, given the scepticism of this line of work among many in the field, it is critical research on consumer sleep trackers is conducted

with strong methodology, is reported clearly, and is interpreted in line with the findings.
My specific comments, below, are written with that in mind:

1. My biggest concern about the paper is the authors focus on the claim the Apple Watch can score sleep/wake the same as the Actiwatch. However, fundamentally, this is not what they have shown. They have shown their algorithm can faithfully replicate the sleep/wake designation made by the Actiwatch on a 15-sec basis. While this would be expected to, and indeed does, produce largely similar all-night sleep measures, that is not the same thing as validating the all-night sleep measures produced by Apple Watch. That can only be done with a PSG validation study. I appreciate the author never explicitly claim this is a “validation” study. Nonetheless, the language they use throughout certainly implies such. I would encourage the authors to refocus the paper around the finding related to their algorithm being optimized to produce the same epoch-level decision as the Actiwatch makes. They can still show the TST and WASO data as evidence this likely translates into all-night sleep measures, but the latter should not be the take home message. This concern propagates through several of the following comments, as well.

2. Abstract: I would encourage the authors to focus the Results section on the algorithm success, not on TST and WASO. They should also remove the claim the Apple Watch “…the Apple Watch could be an acceptable alternative to the Philips Actiwatch for sleep monitoring…”, as this cannot be shown without a formal PSG validation study.

3. Introduction: While everyone agrees PSG has serious limitations, I would not categorize sleep apnea and narcolepsy as "qualitative" abnormalities. They are indeed quantified in a very specific way. It would be more fair to simply say PSG is better suited for use on a single night to detect sleep disorders.

4. Introduction: The authors should clearly state the Aims of the study at the end of the Intro. As stated above, I believe the primary Aim should be to determine if ENMO can provide the same epoch-level sleep/wake decision as that produced by Actiwatch. The TST and WASO assessments should be secondary or exploratory Aims. This would both reflect the supporting role they play to the main Aim, as well as the fact the study is underpowered to detect differences between the two devices in all-night sleep measures.

5. Methods: The reader would benefit from a better description of the sample. What was the age mean, standard deviation, and range? What was the sex ratio? How did the authors screen for sleep disorders? Were there any medical/psychiatric diagnoses? In my view, it may not matter what the exact answers are, as the main purpose here is to see if the Apple Watch can detect motion to the same extent as the Actiwatch. In that regard, it should not really matter who is producing that motion. One the other hand, most actigraphs and consumer wearables perform differently in different populations. Thus, generalizability may well be affected by the demographics of the sample studied here.

6. Methods: Participants wore both devices simultaneously. Did the authors counter-balance the order on the wrist across participants to control for possible effects on motion?

7. Methods: It would provide additional clarity if the authors would link each analysis to an Aim. It was not until I had finished the Results that I realised why analysis 1 and 2 were necessary. This

confusion also reflects the fact the paper is written as if validating the TST and WASO measures are the main Aim (for which the first two analyses are not necessary), when in fact it is not.

8. Methods: I am not able to comment thoroughly on the mathematical procedures used to optimize Apple Watch activity counts so they produce the same sleep/wake decision as Actiwatch. I would suggest the authors ask a modeller or other quantitative specialist to review that information.

9. Methods: The authors report in the Discussion that they impute missing date from Apple Watch. This information should be in the Methods. Moreover, they should very clearly state how many nights and how many unique participants had missing date, as well as the mean/standard deviation and range of the missing data length. Across papers examining consumer wearables, it is becoming increasingly clear they all have problems with not recording data for chunks of time. It is critical for users of these devices to be made aware of this. Finally, I would encourage the authors to redo their analyses dropping these missing data, rather than imputing them. A lot can happen in 20-60 minutes of the night, especially given some of the nights appeared to only have ~5 hours of sleep (per Bland Altman plot). It is a faulty assumption to think the epoch before and after an hour long blank period (or even a 20 minute blank period) can substitute for every epoch within that blank period.

10. Results: Did the authors examine night-to-night variability or test-retest reliability of the Apple Watch, given they had 2 nights for most people? If a consumer device is to be used for long periods of time in the field (by a consumer, researcher, or clinician), it is important to know if the device reliably produces the same accuracy from night to night.

11. Results: The Bland-Altman plots seem to suggest the Apple Watch shows proportional bias, at least for TST and WASO (i.e, the device’s detection of motion, and thus sleep, becomes less accurate with lower TST and greater WASO). The authors should report proportional bias analyses for each plot, and discuss implications of the findings.

12. Discussion: The authors compare their findings to a single paper using a different actigraph. It is not clear why this specific paper was chosen as particularly relevant. Could they authors please clarify that? While the literature on consumer wearables is not particularly large, one would have expected the authors to put their findings into a broader context. For example, the authors could reference a recent review by de Zambotti[1] to compare their findings more broadly.

13. It seems to me the biggest limitation of this study is the limited storage capacity and battery life of the Apple Watch. If I understand correctly, the device was only worn during the night, because it had to be charged during the day or data would be lost. Even if I misunderstand that point, the fact ”recording was limited to two consecutive nights per participant due to memory and data transfer constraints” has serious implications for the practical translation of this work: 1) This does not reflect how the typical consumer will use the Apple Watch; 2) It seem unlikely any large scale (or even small scale) study will adopt the Apple Watch (a stated goal of the authors), as it is impractical to expect participants in large cohort studies to only wear the device at night. This is especially true when there are other devices with much longer recording periods where one can also access the raw data (e.g., GeneActiv and even Fitbit with specialized software) While these points do not undermine the value of the work done here, the authors need to be more clear about these limitations.

References

1. DE Zambotti M, Cellini N, Goldstone A, Colrain IM, Baker FC: Wearable Sleep Technology in Clinical and Research Settings.*Med Sci Sports Exerc*. 2019; **51** (7): 1538-1557 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
I cannot comment. A qualified statistician is required.

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Sleep, insomnia, validation of wearables, sleep deprivation, cognitive performance, mental health

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.