SLEEPJ, 2021, 1–16

doi: 10.1093/sleep/zsaa291
Advance Access Publication Date: 30 December 2020
Original Article

ORIGINAL ARTICLE

Performance of seven consumer sleep-tracking devices compared with polysomnography

Evan D. Chinoy[1,2], Joseph A. Cuellar[1,2], Kirbie E. Huwa[1,2], Jason T. Jameson[1,2], Catherine H. Watson[1,3], Sara C. Bessman[1,4], Dale A. Hirsch[1], Adam D. Cooper[1,3], Sean P.A. Drummond[5] and Rachel R. Markwald[1,*]

[1]Sleep, Tactical Efficiency, and Endurance Laboratory, Warfighter Performance Department, Naval Health Research Center, San Diego, CA [2]Leidos, Inc., San Diego, CA [3]Innovative Employee Solutions, San Diego, CA [4]Eagle Applied Sciences, San Diego, CA [5]Turner Institute for Brain and Mental Health, Monash University, Melbourne, Victoria, Australia

*Corresponding authors. Rachel R. Markwald and Evan D. Chinoy, Sleep, Tactical Efficiency, and Endurance Laboratory, Warfighter Performance Department, Naval Health Research Center, 140 Sylvester Road, San Diego, CA 92106. Email: rachel.r.markwald.civ@mail.mil and evan.d.chinoy.ctr@mail.mil.

Abstract

Study Objectives: Consumer sleep-tracking devices are widely used and becoming more technologically advanced, creating strong interest from researchers and clinicians for their possible use as alternatives to standard actigraphy. We, therefore, tested the performance of many of the latest consumer sleep-tracking devices, alongside actigraphy, versus the gold-standard sleep assessment technique, polysomnography (PSG).

Methods: In total, 34 healthy young adults (22 women; 28.1 ± 3.9 years, mean ± *SD*) were tested on three consecutive nights (including a disrupted sleep condition) in a sleep laboratory with PSG, along with actigraphy (Philips Respironics Actiwatch 2) and a subset of consumer sleep-tracking devices. Altogether, four wearable (Fatigue Science Readiband, Fitbit Alta HR, Garmin Fenix 5S, Garmin Vivosmart 3) and three nonwearable (EarlySense Live, ResMed S+, SleepScore Max) devices were tested. Sleep/wake summary and epoch-by-epoch agreement measures were compared with PSG.

Results: Most devices (Fatigue Science Readiband, Fitbit Alta HR, EarlySense Live, ResMed S+, SleepScore Max) performed as well as or better than actigraphy on sleep/wake performance measures, while the Garmin devices performed worse. Overall, epoch-by-epoch sensitivity was high (all ≥0.93), specificity was low-to-medium (0.18–0.54), sleep stage comparisons were mixed, and devices tended to perform worse on nights with poorer/disrupted sleep.

Conclusions: Consumer sleep-tracking devices exhibited high performance in detecting sleep, and most performed equivalent to (or better than) actigraphy in detecting wake. Device sleep stage assessments were inconsistent. Findings indicate that many newer sleep-tracking devices demonstrate promising performance for tracking sleep and wake. Devices should be tested in different populations and settings to further examine their wider validity and utility.

Statement of Significance

Representing a fast-growing trend, hundreds of millions of people now use consumer devices to track sleep and other biometric data. Previous studies found that device performance is quite variable, although sleep-tracking in some recent devices has improved. If validated, such devices could be used to help maintain or improve sleep health and potentially be important tools in research and clinical practice. In this study, we rigorously tested the sleep-tracking claims of seven devices against the gold-standard, polysomnography, and found most performed as well or better than the mobile sleep assessment standard, actigraphy, on key performance metrics. The current findings demonstrate promise for many newer devices to serve as valid sleep-tracking alternatives to actigraphy, however, further performance testing is needed.

Key words: validation; actigraphy; sleep technology; wearables; nearables; sensors; noncontact; polysomnography

Submitted: 29 August, 2020; **Revised:** 2 December, 2020

Introduction

Sleep plays an integral role in our physical and mental health [1–4] and for achieving high levels of alertness and performance [5–7], thereby impacting everyday functioning. Therefore, how sleep is defined and measured is important, because understanding the role of sleep in critical areas of health and behavior relies on the precision of the sleep metrics obtained. In the objective measurement of sleep, polysomnography (PSG) provides the most direct assessment and thus has remained the gold-standard technique in research laboratories and sleep medicine clinics for over half a century. Aside from the measurement and diagnosis of sleep disorders, PSG is also used to determine sleep and wake states as well as individual sleep stages based on standard criteria [8]. However, PSG is not practical outside the laboratory or clinic due to a number of factors. These include its relatively high cost, the specialized training, and the time burden required to conduct and interpret studies. Moreover, PSG recording procedures and equipment require controlled settings, are cumbersome, and can be disruptive to sleep itself.

Actigraphy, on the other hand, overcomes many of these barriers. Actigraphy utilizes a research-grade wrist-worn device to collect physical activity data that are later processed with algorithms to estimate sleep and wake [9–12]. These devices are portable, relatively easy to set up, have long battery life, and are less expensive and obtrusive than PSG. Actigraphy was an important advance in the sleep field because it expanded the capability of objective sleep assessment into home and field environments, and it allows continuous recording over multiple weeks. However, there are many limitations of actigraphy compared with PSG [11–17]. Actigraphy performs best when wrist movements during wake are robust, although people do not always make robust movements during awakenings and such times are likely to be misclassified as sleep. This results in the overestimation of sleep and underestimation of wake, versus PSG [11–17]. Further, because it is an indirect measure of sleep, actigraphy has less resolution and precision than PSG—limiting it to binary sleep and wake classifications only, as opposed to individual sleep stages. To achieve the highest level of precision, actigraphy algorithms need to be manually directed to the exact timing of sleep episodes during the postprocessing analysis. This places added burdens on the user to denote bed and wake times each day using sleep diaries, and on the researcher or clinician to complete the postprocessing steps.

Although actigraphy is considered the mobile sleep assessment standard, the recent and rapid development of advanced multisensor consumer devices has raised questions about the possible validity and utility of consumer devices as acceptable alternatives to actigraphy in the measurement of sleep [17–23]. The majority of consumer devices that offer sleep-tracking are in the form of "wearables" that are worn on the wrist, however, some are designed for other body areas too (e.g. finger, head, torso). Alternatively, several companies have developed nonwearable (or "nearable") devices that are placed close to the user (e.g. under a mattress, on a bedside table) to track sleep using remote detection of physiological and behavioral signals. Compared with actigraphy devices, consumer devices are lower cost and come equipped with additional features, enabled by wireless or Bluetooth connections that allow for near-real-time data processing. Although most consumer devices contain accelerometers similar to the ones in actigraphy devices, many contain other sensors (e.g. heart rate) that are used as additional inputs into proprietary sleep algorithms. With multiple physiological and behavioral data inputs and continued improvements in sensor technology, sleep-tracking devices may be able to equal or out-perform standard actigraphy. Indeed, several recent studies [24–29] of the latest generation consumer devices versus PSG have found improvements in sleep-tracking performance—suggesting that consumer devices are demonstrating promise toward stronger validation, and with further testing and development may prove useful in the maintenance or improvement of sleep health over time.

In 2019, around 30% of consumers surveyed in the United States owned a wearable device [30] with an additional 350 million wearable device units that were projected to ship in 2020 [31]. Apart from fitness-tracking (the most popular selling point and use for these devices), around 25% of US adults surveyed have used a wearable device or phone application ("app") to track their sleep at least once [32]. With more devices now offering sleep-tracking to users, the wearable sleep-tracking device market is projected to reach $7 billion by 2026 [33]. These trends indicate that a large section of the population is now tracking their sleep, and potentially using device data to inform their sleep habits and other health, wellness, and behavioral choices. However, whether these devices perform well enough to generate accurate and reliable data, and whether the choices people make concerning their health and behavior can be well-informed from the use of these devices, are critical questions in this generation of the "quantified self."

Given the large scope of current device use, rapid technological advancements, promising findings from recent performance studies, and strong interest from researchers and clinicians for their possible use as alternatives to actigraphy, it is important to continue evaluating the performance of consumer sleep-tracking devices. Accordingly, in the present study we tested many of the latest wearable and nonwearable devices against PSG, and alongside actigraphy for direct comparison. Validation testing is a multistep process, and the initial step is to test performance under well-controlled conditions. Thus, following the guidelines for implementation and reporting of sleep device performance studies recently put forth by the Sleep Research Society [23] and others [34], we aimed to provide an initial evaluation of the performance and reliability of multiple consumer sleep-tracking devices in a controlled laboratory setting with a sample of healthy young adult participants.

Methods

Participants

In total, 34 healthy young adults (12 men, 22 women) aged 28.1 ± 3.9 years (mean ± *SD*) participated in the study. Screening consisted of a self-report medical history questionnaire to assess the following exclusion criteria: age <18 or >35 years, body mass index (BMI) <18.5 or ≥30.0 kg/m^2 (24.2 ± 2.1 kg/m^2; mean ± *SD*), average nightly sleep duration <6 or >9 h, any diagnosed sleep, mental health, or other significant medical disorder, use of any prescription or over-the-counter sleep medications in the previous one month, use of any nicotine product in the previous one month, use of any illegal drug in the previous 6 months, pregnant women, any physical or living conditions affecting the ability to complete the home or laboratory protocols, and

any shift work (night shifts or rotating shifts) or travel >1 time zone within one month prior to study. Height and weight were measured by research staff at the screening appointment to calculate BMI.

The study protocol was approved by the Naval Health Research Center Institutional Review Board and was conducted in accordance with the Declaration of Helsinki. All participants provided signed informed consent prior to the study and were compensated for their participation with gift cards.

Prestudy conditions

Beginning 6 days prior to the first in-lab study visit and continuing throughout the study, participants stopped their consumption of caffeine and alcohol. For the four nights prior to the first lab visit, participants maintained consistent 8-h sleep schedules at home and refrained from taking naps. Specifically, participants were asked to self-select a target sleep schedule with time in bed (TIB) for exactly 8 h that they could maintain across all four prestudy nights at home, and they were allowed to deviate from their target sleep schedule by only up to 30 min (earlier or later) on each night while still maintaining exactly 8-h TIB.

During the prestudy period, participants wore the Actiwatch 2 (Philips Respironics, Inc.; Murrysville, PA, USA) research-grade wrist actigraphy device (hereafter referred to as "Actiwatch") and completed written sleep diaries. Participants were instructed to continuously wear the Actiwatch (except when showering or during other activities where it could get damaged) on their nondominant wrist and to complete the sleep diary entries every morning, denoting the exact bed and wake times for all four prestudy home sleep nights. Upon admission to the first lab visit, actigraphy data and sleep diary entries were checked by research staff to verify compliance to the prestudy sleep schedule.

Study eligibility criteria and compliance to the study conditions were further verified upon admission to the lab on all three study visits—participants were required to sign a form attesting that they still met each of the study eligibility criteria and adhered to all the prestudy conditions, and they were tested with an alcohol breathalyzer to verify sobriety. Upon admission on the first lab visit only, female participants were asked to provide a urine sample for research staff to test and verify nonpregnancy status.

In addition to the Actiwatch, during the four prestudy days and nights all participants continuously wore the subset of consumer wrist devices that they were subsequently tested with on their lab visits. Thus, by the first lab visit, participants were already familiarized to wearing the devices—which was intended to mitigate some of the possible effects on sleep during the lab visits that could arise from wearing multiple wrist devices at once (e.g. novelty, discomfort).

Study protocol

The study was conducted at the Naval Health Research Center in San Diego, CA, USA, and consisted of three consecutive overnight lab visits. On Lab Visit 1, participants reported to the lab ~2.5 h prior to their calculated habitual bedtime and underwent a PSG electrode application conducted by trained research staff members. During the start of each lab visit, research staff ensured that devices were worn correctly according to each device company's guidelines. Participants were allowed to use their personal electronic devices (e.g. cell phone) until 30 min prior to their bedtime, at which time their personal devices were removed from the study bedroom for the remainder of the visit.

To account for individual differences in habitual sleep schedules and estimated circadian timing, the sleep schedule of each participant's lab visits was calculated and set from the midpoint average of their bed and wake times from their four prestudy nights. TIB was exactly 8 h and their bed and wake times were set at the same clock times on all three lab visits. Participants slept in individual sound-attenuated research bedrooms, without windows or clocks. Room lighting was set to ~150 lx, as measured from a seated position in the room. At scheduled bedtime, the room lights were turned off and participants slept in darkness. Research staff monitored the study from a control station outside the bedroom equipped with an infrared video display and an audio intercom system. At wake time, research staff turned on the lights, entered the room, removed the electrodes, and synced the sleep device data to a tablet computer (iPad; Apple Inc.; Cupertino, CA, USA). Afterward, participants left the lab and were required to continue wearing all devices during the daytime and report back later that evening for the next study visit. Participants were not given any additional restrictions on their activities in between lab visits other than the prestudy conditions previously described.

Experimental sleep disruption protocol

An experimental sleep disruption protocol occurred on one of the final two overnight lab visits. This provided an opportunity to examine the effects of disrupted (i.e. fragmented) sleep patterns on device algorithm performance. Understanding the sleep-tracking performance of devices on nights with disrupted and/or poor sleep (like a sleep-maintenance insomnia profile) could yield important additional findings that would not be easily achievable given the healthy sleep cohort included in the study, and especially if only tested under ideal undisrupted sleeping conditions—as is typical in laboratory sleep research studies. The experimental sleep disruption protocol used auditory tones played through a speaker in the bedroom to awaken participants for a brief scheduled period at every hour within the sleep episode. The disrupted sleep protocol was randomized and counterbalanced between Lab Visits 2 and 3, and during the other visit, the participants were not disrupted by external stimuli during the sleep episode (see Supplemental Materials for additional details on the experimental sleep disruption protocol methods).

Consumer sleep-tracking devices tested

In total, seven consumer devices were tested in the study. Four of the devices were wrist-worn wearables: Fatigue Science Readiband (Fatigue Science; Vancouver, BC, Canada), Fitbit Alta HR (Fitbit, Inc.; San Francisco, CA, USA), Garmin Fenix 5S (Garmin, Ltd; Olathe, KS, USA), and Garmin Vivosmart 3. The three other devices tested were nonwearables (or "nearables") that plug into a wall outlet for power. One of the nonwearables was an under-mattress device that uses a piezoelectric sensor to detect heart rate, breathing, and physical movement, that was

placed 6 in. from the side of the bed and under the participant's chest: EarlySense Live (EarlySense, Ltd; Woburn, MA, USA). The two other nonwearables were bedside devices using ultra-low-power radiofrequency waves for signal detection, placed about one arm's length away on a bedside table at the same height as the top of the mattress and with the device pointed toward the participant's chest: ResMed S+ (ResMed, Inc.; San Diego, CA, USA) and SleepScore Max (SleepScore Labs; Carlsbad, CA, USA). Epoch duration for sleep data output by one of the wearables (Fitbit Alta HR) and all the nonwearable devices (EarlySense Live, ResMed S+, and SleepScore Max) was 30 s. However, three of the wearable devices (Fatigue Science Readiband, Garmin Fenix 5S, and Garmin Vivosmart 3) only output sleep data in 60-s epochs. Prior to sleep testing on each lab visit, all devices and the PSG computers were time synced to the clock time displayed on the tablet computer.

Due to practical constraints, not all devices could be tested at once or on all the participants. Therefore, each participant wore the Actiwatch and used a subset of consumer devices that included multiple wearable devices and one or more nonwearable devices (see Table S1 in the Supplemental Materials which depicts the devices used by each participant). The following are the total number of participants who used each device: Fatigue Science Readiband: 15, Fitbit Alta HR: 20, Garmin Fenix 5S: 11, Garmin Vivosmart 3: 15, EarlySense Live: 19, ResMed S+: 19, SleepScore Max: 15. Additional details on consumer sleep devices, testing methods, and analysis are included in the Supplemental Materials.

PSG recording and analysis

Overnight laboratory PSG recordings were acquired using a digital PSG recorder (Siesta; Compumedics USA, Inc.; Charlotte, NC, USA), with PSG data sampled and stored at 256 Hz. The PSG montage included electroencephalography recordings at six brain sites referenced to contra-lateral mastoid processes for F3-M2, F4-M1, C3-M2, C4-M1, O1-M2, and O2-M1. Left and right electrooculograms, left and right mentalis electromyograms, and two-lead electrocardiogram were also recorded in the PSG montage. Impedances were ≤10 kΩ at the start of the recordings. PSG electrode sites were measured and applied according to standard criteria (International 10–20 System of Electrode Placement).

PSG sleep stages (N1, N2, N3, and rapid eye movement [REM]) and wake were manually scored in 30-s epochs by two experienced American Academy of Sleep Medicine (AASM) certified Registered Polysomnographic Technologists (RPSGTs) who had >90% scoring agreement with AASM gold-standard PSG records. PSG data were scored using the standard criteria [8], and all three PSG nights for a given participant were scored by the same RPSGT. As stated above, the devices output data in either 30- or 60-s epochs. The standard 30-s PSG-scored epochs were used for all the sleep summary and epoch-by-epoch (EBE) analyses of the 30-s devices. For the EBE analyses, the PSG data were additionally scored in 60-s epochs to align with the devices that output data in this way (Fatigue Science Readiband, Garmin Fenix 5S, and Garmin Vivosmart 3). The same RPSGT scored these records using the standard scoring criteria [8] but applied over every 60-s epoch starting from bedtime. Several of the consumer devices also output data for individual sleep stages but, in addition to REM, only characterize the non-REM (NREM) sleep stages as two stages: "light" and "deep." Therefore, for the comparison of device sleep stage data with PSG (and in following the analysis procedures of previous device performance studies [24, 26–29, 35–37]), the PSG epochs scored as N1 and N2 were combined to form a light PSG sleep stage, and PSG epochs scored as N3 were the deep PSG sleep stage.

In addition to the total time and percentage of the 8-h sleep episodes spent awake or in each sleep stage, the following sleep summary measures were calculated: total sleep time (TST; total time spent in all sleep stages), sleep efficiency (SE; percentage of TST divided by TIB), sleep latency (time taken to fall asleep from bedtime), and wake after sleep onset (WASO; time awake in each sleep episode occurring after sleep onset). Sleep latency was calculated in two ways: (1) sleep onset latency (SOL), the time from bedtime to the first epoch scored as any sleep stage, and (2) latency to persistent sleep (LPS), the time from bedtime to the first epoch of 10 consecutive minutes scored as any sleep stage. WASO was calculated from both SOL and LPS. REM latency was also calculated, as the time from bedtime to the first epoch scored as REM.

Device data export procedures and analysis

Consumer device sleep data were exported and/or viewed from the device apps on the tablet computer or from an online website portal set up by the device company that also provides device data to the user. Additionally, upon our request, several of the device companies also provided access to the data via research portals, which were outside the standard device apps and online user account web portals. This was the case for the EBE sleep data analyzed from the EarlySense Live, ResMed S+, and SleepScore Max devices. Fitbit, Inc. does not directly provide the EBE sleep data for their devices via any of the standard data viewing or export channels. Instead, the Fitbit Alta HR EBE data were exported via Fitabase (Small Steps Labs, LLC; San Diego, CA, USA), a licensed third-party data management platform. Actiwatch actigraphy sleep and wake data were analyzed in 30-s epochs using the medium sensitivity threshold with the software package Actiware, version 6.0.9 (Philips Respironics, Inc.; Murrysville, PA, USA).

Missing data procedures

On each lab visit, the research staff carefully synced and charged all devices, actigraphy watches, and computers/tablets in order to capture sleep data for PSG, actigraphy, and all devices over the scheduled 8-h sleep episode. However, despite these best practices, issues with missing data, partial data loss, and/or time syncing of devices did occasionally occur (see the Supplemental Materials for additional detailed information on the treatment of missing data for analysis).

Statistical analysis

The sample sizes for the sleep/wake summary and EBE analyses were dictated by the availability of valid pairs of data between PSG and each device. The final sleep/wake summary analysis sample sizes are shown in the respective tables, and the total EBE sample sizes are shown in the EBE contingency tables (Tables S2 and S3) in the Supplemental Materials.

The following EBE agreement statistics were calculated for the analysis of all sleep versus wake epochs for each device, in comparison to the same epoch as scored by PSG: sensitivity (true positive rate; the proportion of PSG sleep epochs correctly detected as sleep by the device), specificity (true negative rate; the proportion of PSG wake epochs correctly detected as wake by the device), positive predictive value (PPV; proportion of device-scored sleep epochs that were true PSG sleep), negative predictive value (NPV; proportion of device-scored wake epochs that were true PSG wake), accuracy (proportion of all PSG epochs correctly detected by the device), and the prevalence and bias-adjusted kappa (PABAK; Cohen's kappa weighted to account for the amount of inequality between the number of sleep and wake epochs).

Most of the devices (Fitbit Alta HR, Garmin Fenix 5S, Garmin Vivosmart 3, EarlySense Live, ResMed S+, SleepScore Max) also output individual sleep stage classifications (light, deep, REM) for each epoch (rather than just a binary sleep versus wake classification, as was the case for the Actiwatch and Fatigue Science Readiband). Therefore, for the devices that output sleep stages, EBE agreement statistics were calculated for each stage versus the combination of all other classifications (e.g. EBE agreement for light sleep was calculated as the light sleep epochs versus the combination of all wake, deep, and REM epochs, etc.). Because there are multiple classification possibilities with sleep stage analysis, the proportions of misclassification were also determined for each stage (e.g. how often PSG light sleep epochs were misclassified by a device as deep sleep, etc.).

Bland–Altman plots [38] for sleep summary measures were constructed for each device to show all the individual night differences versus PSG as well as the overall levels for bias (the average difference between the device and PSG) and the upper and lower limits of agreement (two standard deviations) from the bias. Best-fit curves shown in Bland–Altman plots are locally estimated scatterplot smoothing (LOESS) curves, with 95% confidence bands (using standard errors). LOESS methods use low degree polynomials to model local structure in subsets of the data, displaying possible local patterns in the data without forcing a particular global form at the outset. Device summary measures were statistically compared with PSG using Student's paired *t*-tests, *Hedges'* g effect sizes, and R^2 proportional biases. Proportional bias was calculated using linear regression methods for Bland–Altman plots [39] (and has been used previously for device performance testing [29]), which indicates the reliability of the bias versus PSG (*y*-axis value) and whether it changes in proportion to the mean of the device and PSG (*x*-axis value) in the Bland–Altman plots. *p*-values at the $p < 0.05$ level were considered statistically significant.

Statistics were performed using the statistical analysis software package R, version 3.5.3 (R Foundation, Vienna, Austria).

Results

Sleep/wake summary results

Sleep/wake summary measures versus PSG for all combined nights are shown in Tables 1–4 and Supplemental Tables S4 and S5, and corresponding Bland–Altman plots in Figures 1–4 and Supplemental Figures S1 and S2. The Actiwatch significantly differed from PSG on all the sleep/wake summary measures—it overestimated the sleep-related measures (TST and SE) and underestimated the wake-related measures (SOL, LPS, and WASO from both SOL and LPS). Several consumer devices (Garmin Fenix 5S, Garmin Vivosmart 3, and EarlySense Live) also significantly overestimated TST (Table 1) compared with PSG. The SE results (Table 2) were very similar to the TST results (largely because TIB was fixed at 8 h) and thus all the devices that significantly overestimated TST also did so for SE. The Bland–Altman plots for TST and SE (Figures 1 and 2) demonstrate that the biases for each device versus PSG were generally the lowest magnitude and least variable when participants had higher TST and SE, and was generally more variable and biased on nights with lower TST and SE. This was especially the case for the Actiwatch, Fitbit Alta HR, Garmin Fenix 5S, Garmin Vivosmart 3, and SleepScore Max which each had significant levels of proportional bias for TST and SE.

Several of the consumer devices significantly differed on the sleep latency measures SOL and LPS (Table 3 and Supplemental Table S4, respectively), as compared with PSG. These differences, although significant, were low magnitude (each <5 min), less than the sleep latency biases of Actiwatch, and no consumer device differed from PSG on both sleep latency measures. SOL was underestimated by the Fitbit Alta HR and overestimated by the ResMed S+ and SleepScore Max, while the Fatigue Science Readiband underestimated LPS. Similar to TST and SE, the Bland–Altman plots for SOL

Table 1. Sleep summary: total sleep time (TST)

Device	*n*	PSG Mean ± *SD*	Device Mean ± *SD*	Bias	Lower limit	Upper limit	*t* (*p*)	Effect size	R^2 (*p*)
Actiwatch	98	418.2 ± 40.7	442.1 ± 19.7	23.9	–40.5	88.3	7.3 ***(<0.001)***	0.74	0.51 ***(<0.001)***
Fatigue Science Readiband	41	416.0 ± 47.9	429.3 ± 53.0	13.3	–93.2	119.8	1.6 (0.118)	0.26	0.01 (0.482)
Fitbit Alta HR	49	425.1 ± 33.1	427.7 ± 19.7	2.6	–42.0	47.1	0.8 (0.421)	0.09	0.41 ***(<0.001)***
Garmin Fenix 5S	29	413.1 ± 53.0	456.8 ± 21.0	43.7	–47.0	134.4	5.2 ***(<0.001)***	1.07	0.61 ***(<0.001)***
Garmin Vivosmart 3	43	414.7 ± 48.4	461.5 ± 16.4	46.8	–39.5	133.1	7.1 ***(<0.001)***	1.28	0.69 ***(<0.001)***
Earlysense Live	51	421.6 ± 34.8	435.2 ± 30.1	13.6	–45.1	72.3	3.3 ***(0.002)***	0.42	0.03 (0.214)
ResMed S+	51	422.3 ± 33.8	422.0 ± 40.0	–0.3	–70.7	70.2	–0.1 (0.953)	–0.01	0.04 (0.159)
SleepScore Max	42	413.6 ± 48.9	421.1 ± 37.1	7.5	–60.7	75.7	1.4 (0.162)	0.17	0.14 ***(0.016)***

Summary results for minutes of TST for the devices versus polysomnography (PSG). All nights with available TST data for both the device and PSG are included, with the total number of nights (*n*) indicated in each row. Mean and standard deviation (*SD*) are shown for PSG and each device. Bias represents the mean difference between PSG and the device, with positive and negative bias values indicating the device showed an overestimation or underestimation compared with PSG, respectively. Lower and upper limits of agreement represent two *SD*s from the bias. Statistical significance between each device and PSG was assessed with paired *t*-tests and corresponding *p*-values. Effect sizes (*Hedges' g*) and proportional biases (R^2) with corresponding p-values are also shown. *p*-values *at the* $p < 0.05$ level were considered statistically significant and are shown in bold and italic.

Table 2. Sleep summary: sleep efficiency (SE)

Device	*n*	PSG Mean ± SD	Device Mean ± SD	Bias	Lower limit	Upper limit	*t* (*p*)	Effect size	R^2 (*p*)
Actiwatch	98	87.1 ± 8.5	92.1 ± 4.1	5.0	−8.5	18.4	7.3 (**<0.001**)	0.74	0.51 (**<0.001**)
Fatigue Science Readiband	41	86.7 ± 10.0	89.4 ± 11.0	2.8	−19.4	24.9	1.6 (0.117)	0.26	0.01 (0.487)
Fitbit Alta HR	49	88.6 ± 6.9	89.4 ± 4.0	0.9	−8.4	10.2	1.3 (0.191)	0.16	0.45 (**<0.001**)
Garmin Fenix 5S	29	86.1 ± 11.0	96.6 ± 2.9	10.6	−9.0	30.1	5.8 (**<0.001**)	1.29	0.82 (**<0.001**)
Garmin Vivosmart 3	43	86.4 ± 10.1	96.5 ± 3.1	10.1	−7.3	27.6	7.6 (**<0.001**)	1.34	0.76 (**<0.001**)
Earlysense Live	51	87.8 ± 7.3	90.8 ± 6.1	2.9	−9.2	15.1	3.4 (**0.001**)	0.43	0.04 (0.148)
ResMed S+	51	88.0 ± 7.0	88.0 ± 8.3	0.0	−14.7	14.7	0.0 (0.996)	0.00	0.04 (0.158)
SleepScore Max	42	86.2 ± 10.2	87.8 ± 7.8	1.6	−12.6	15.8	1.5 (0.150)	0.18	0.13 (**0.017**)

Summary results for the percentage of SE for the devices versus polysomnography (PSG). See Table 1 caption for additional table details.

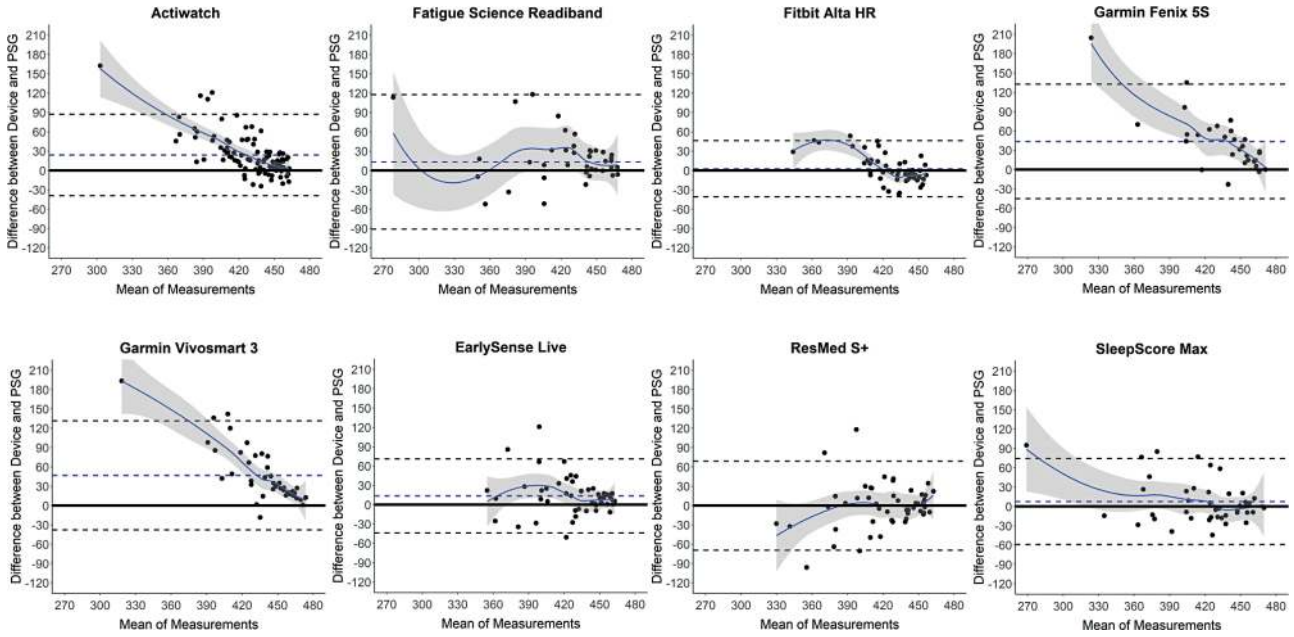

Figure 1. Bland–Altman plots: total sleep time (TST). Bland–Altman plots depicting the mean bias (blue dashed line) and upper and lower limits of agreement (two standard deviations from bias; black dashed lines) for minutes of TST for the devices compared with polysomnography (PSG). Black circles are individual nights. Solid blue curves represent the best-fit of data, with surrounding gray shaded regions representing 95% confidence bands. The solid black line at zero represents no difference, with positive and negative *y*-axis values indicating an overestimation or underestimation, respectively, compared with PSG.

(Figure 3) and LPS (Supplemental Figure S1) demonstrate that nights with lower sleep latency (i.e. faster time to fall asleep) were closer to the zero line of agreement with PSG, and the nights with higher sleep latencies were more variable and biased. This pattern is further corroborated by the significant proportional biases found for all devices, either on one or both of the sleep latency measures.

WASO was calculated from both sleep latency measures (from SOL in Table 4, from LPS in Supplemental Table S5) and was significantly underestimated versus PSG in several devices (Garmin Fenix 5S, Garmin Vivosmart 3, EarlySense Live, and SleepScore Max). All devices that differed in WASO from PSG did so whether calculated from SOL or LPS, indicating negligible differences in WASO between the two sleep latency thresholds. Corresponding Bland–Altman plots (Figure 4 and Supplemental Figure S2) again demonstrate that for nights with lower WASO (and thus higher TST and SE) the agreement between devices and PSG was better, and on nights when there was higher WASO the devices were more variable. Further, all devices except the Fatigue Science Readiband and ResMed S+ had significant levels of proportional bias for WASO, from either one or both sleep latency measures.

Sleep stage summary results

Sleep stage summary comparisons to PSG are shown for the six devices that output sleep stage data for light, deep, and REM, respectively, in Tables 5–7 and corresponding Bland–Altman plots in Figures 5–7. All six devices significantly differed from PSG in their estimation of light sleep, all with overestimations (except the EarlySense Live which underestimated light sleep). Three devices (EarlySense Live, ResMed S+, SleepScore Max) significantly differed on deep sleep, all with overestimations. Additionally, three devices (Fitbit Alta HR, ResMed S+, SleepScore Max) differed on REM sleep, all with underestimations. Bland–Altman plots (Figures 5–7) generally depict large spreads of data across both axes, and greater individual night variability for the sleep stages compared with the sleep/wake measures. The EarlySense Live had significant proportional bias across all stages, however, no other devices had a significant proportional bias for either

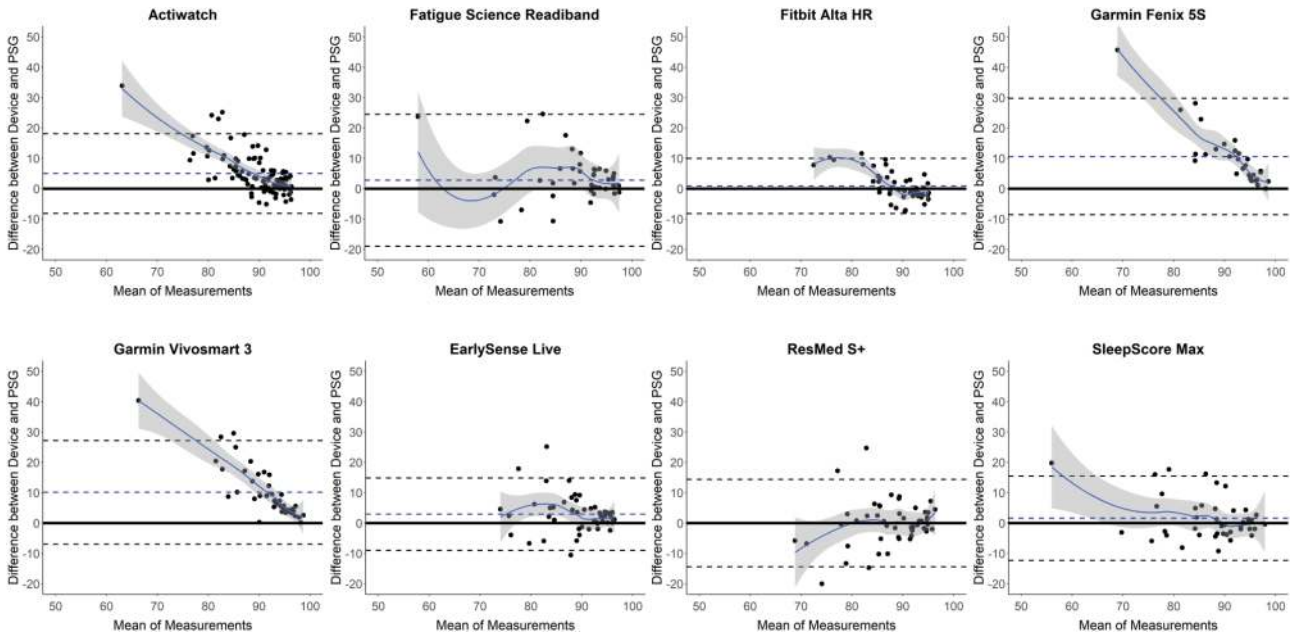

Figure 2. Bland–Altman plots: sleep efficiency (SE). Bland–Altman plots depicting the percentage of SE for the devices compared with polysomnography (PSG). See Figure 1 caption for additional details on the interpretation of Bland–Altman plots.

Table 3. Sleep summary: sleep onset latency (SOL)

Device	*n*	PSG Mean ± *SD*	Device Mean ± *SD*	Bias	Lower limit	Upper limit	*t* (*p*)	Effect size	R^2 (*p*)
Actiwatch	102	9.7 ± 8.5	2.1 ± 1.2	–7.6	–24.1	8.8	–9.4 (**<0.001**)	–1.25	0.93 (**<0.001**)
Fatigue Science Readiband	42	9.8 ± 7.4	9.0 ± 10.6	–0.7	–18.2	16.7	–0.5 (0.593)	–0.08	0.17 (**0.007**)
Fitbit Alta HR	57	8.9 ± 7.7	5.8 ± 4.7	–3.1	–19.0	12.8	–2.9 (**0.005**)	–0.48	0.22 (**<0.001**)
Garmin Fenix 5S	30	9.5 ± 8.1	10.3 ± 13.9	0.8	–26.4	28.0	0.3 (0.750)	0.07	0.26 (**0.004**)
Garmin Vivosmart 3	44	9.6 ± 7.3	8.5 ± 7.5	–1.1	–12.1	9.9	–1.3 (0.192)	–0.15	0.00 (0.789)
Earlysense Live	55	9.7 ± 9.6	10.5 ± 6.9	0.8	–15.2	16.8	0.8 (0.451)	0.10	0.14 (**0.004**)
ResMed S+	54	10.1 ± 9.6	14.1 ± 15.9	4.0	–25.1	33.1	2.0 (**0.049**)	0.30	0.26 (**<0.001**)
SleepScore Max	44	9.6 ± 7.2	14.0 ± 14.3	4.4	–18.9	27.6	2.5 (**0.017**)	0.38	0.45 (**<0.001**)

Summary results for minutes of SOL for the devices versus polysomnography (PSG). See Table 1 caption for additional table details.

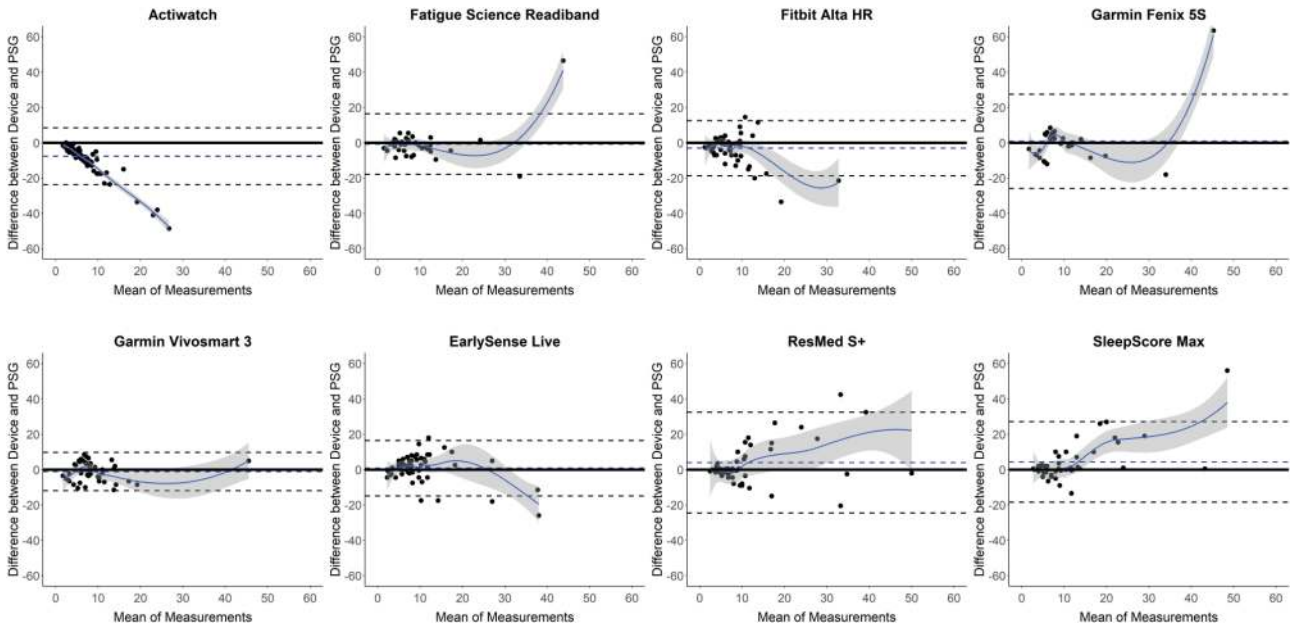

Figure 3. Bland–Altman plots: sleep onset latency (SOL). Bland–Altman plots depicting the minutes of SOL for the devices compared with polysomnography (PSG). See Figure 1 caption for additional details on the interpretation of Bland–Altman plots.

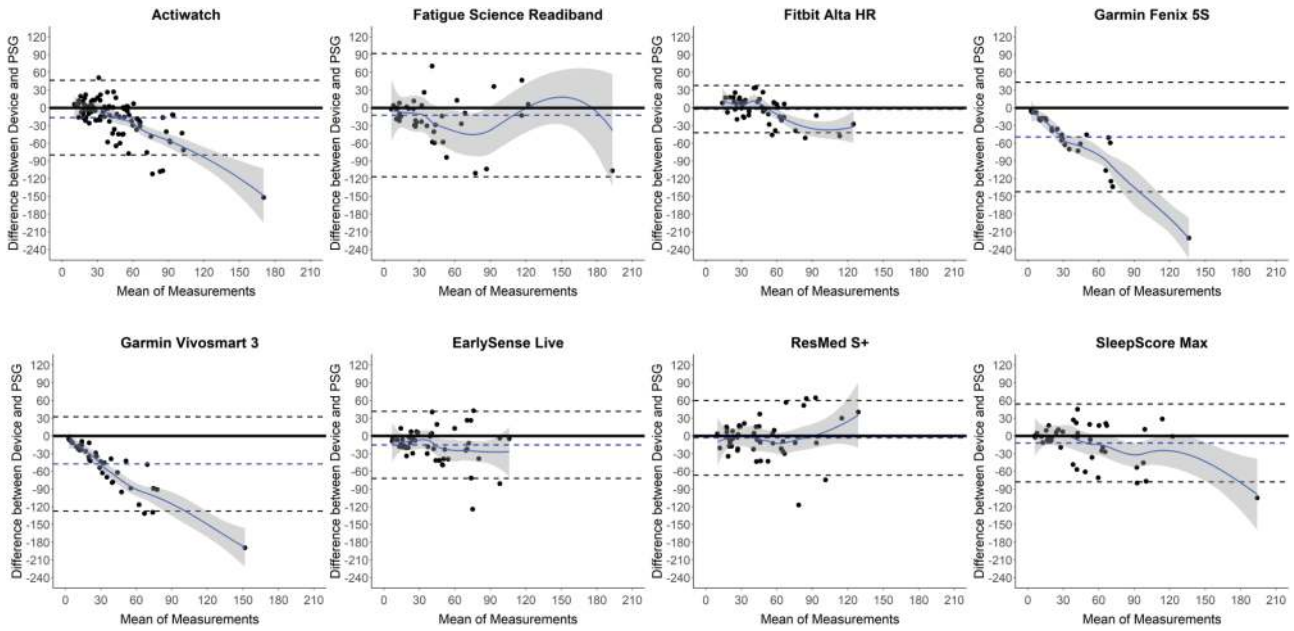

Figure 4. Bland–Altman plots: wake after sleep onset (WASO). Bland–Altman plots depicting the minutes of WASO from sleep onset latency (SOL) for the devices compared with polysomnography (PSG). See Figure 1 caption for additional details on the interpretation of Bland–Altman plots.

Table 4. Sleep summary: wake after sleep onset (WASO)

Device	*n*	PSG Mean ± *SD*	Device Mean ± *SD*	Bias	Lower limit	Upper limit	*t* (*p*)	Effect size	R^2 (*p*)
Actiwatch	98	52.5 ± 39.5	35.9 ± 19.4	−16.6	−81.0	47.9	−5.1 **(<0.001)**	−0.53	0.47 **(<0.001)**
Fatigue Science Readiband	41	54.1 ± 47.5	41.6 ± 52.3	−12.5	−119.0	94.0	−1.5 (0.140)	−0.25	0.01 (0.506)
Fitbit Alta HR	49	46.6 ± 30.8	44.5 ± 19.4	−2.1	−42.9	38.7	−0.7 (0.472)	−0.08	0.35 **(<0.001)**
Garmin Fenix 5S	29	57.2 ± 52.9	7.7 ± 11.7	−49.5	−144.2	45.1	−5.6 **(<0.001)**	−1.27	0.87 **(<0.001)**
Garmin Vivosmart 3	43	55.6 ± 48.1	8.0 ± 12.8	−47.6	−129.0	33.8	−7.7 **(<0.001)**	−1.34	0.85 **(<0.001)**
Earlysense Live	51	49.2 ± 32.2	33.9 ± 26.7	−15.3	−73.2	42.6	−3.8 **(<0.001)**	−0.52	0.05 (0.124)
ResMed S+	51	48.3 ± 31.1	44.9 ± 34.0	−3.4	−67.8	61.1	−0.8 (0.457)	−0.10	0.01 (0.460)
SleepScore Max	42	56.7 ± 48.6	44.6 ± 34.1	−12.1	−79.5	55.3	−2.3 **(0.025)**	−0.29	0.22 **(0.002)**

Summary results for total minutes of WASO, from sleep onset latency (SOL), for the devices versus polysomnography (PSG). See Table 1 caption for additional table details.

Table 5. Sleep summary: light sleep

Device	*n*	PSG Mean ± *SD*	Device Mean ± *SD*	Bias	Lower limit	Upper limit	*t* (*p*)	Effect size	R^2 (*p*)
Fitbit Alta HR	49	236.6 ± 28.5	256.7 ± 30.1	20.0	−54.1	94.2	3.8 **(<0.001)**	0.68	0.00 (0.714)
Garmin Fenix 5S	29	238.3 ± 36.2	267.3 ± 35.8	29.0	−74.4	132.4	3.0 **(0.005)**	0.80	0.00 (0.957)
Garmin Vivosmart 3	43	238.3 ± 31.9	273.0 ± 36.1	34.7	−60.5	129.8	4.8 **(<0.001)**	1.01	0.02 (0.431)
Earlysense Live	51	237.3 ± 31.0	215.0 ± 51.5	−22.3	−133.6	89.0	−2.9 **(0.006)**	−0.52	0.22 **(<0.001)**
ResMed S+	51	235.9 ± 31.1	253.0 ± 34.3	17.1	−58.5	92.6	3.2 **(0.002)**	0.52	0.01 (0.468)
SleepScore Max	42	236.5 ± 32.2	259.1 ± 38.7	22.7	−51.7	97.0	4.0 **(<0.001)**	0.63	0.04 (0.196)

Summary results for total minutes of light sleep for the devices versus polysomnography (PSG). For PSG, light sleep was calculated as the combination of N1 and N2 sleep stages. Results are shown for all devices that output sleep stage classifications. See Table 1 caption for additional table details.

light or deep sleep. For REM sleep, proportional bias was also significant for the two Garmin devices.

REM latency (Table S6 and Supplemental Figure S3)—the time to enter REM sleep, and in this analysis represents the ability of a device to track the first NREM–REM cycle of the night—significantly differed from PSG for (nearly) all six devices that track sleep stages. There was a nonsignificant trend for EarlySense Live to underestimate REM latency, while the other five devices all significantly overestimated REM latency. Further, the Fitbit Alta HR, Garmin Vivosmart 3, and ResMed S+ had significant levels of proportional bias for REM latency.

EBE classification of sleep versus wake

EBE agreement of sleep versus wake compared with PSG for all nights is shown in Table 8. Sensitivity for devices versus PSG was very high (all ≥0.93), indicating a high ability for devices

Table 6. Sleep summary: deep sleep

Device	*n*	PSG Mean ± *SD*	Device Mean ± *SD*	Bias	Lower limit	Upper limit	*t* (*p*)	Effect size	R^2 (*p*)
Fitbit Alta HR	49	81.1 ± 28.1	75.0 ± 21.1	–6.0	–73.8	61.7	–1.2 (0.219)	–0.24	0.08 (0.052)
Garmin Fenix 5S	29	63.3 ± 22.9	69.4 ± 29.3	6.1	–53.0	65.1	1.1 (0.278)	0.23	0.07 (0.170)
Garmin Vivosmart 3	43	66.7 ± 23.1	70.4 ± 28.9	3.7	–57.5	64.9	0.8 (0.431)	0.14	0.06 (0.130)
Earlysense Live	51	79.6 ± 30.3	115.5 ± 42.9	35.9	–66.8	138.6	5.0 **(<0.001)**	0.96	0.11 **(0.017)**
ResMed S+	51	81.5 ± 27.8	96.0 ± 30.5	14.5	–54.8	83.8	3.0 **(0.004)**	0.49	0.01 (0.504)
SleepScore Max	42	66.8 ± 23.7	87.4 ± 28.9	20.7	–36.3	77.7	4.7 **(<0.001)**	0.77	0.05 (0.166)

Summary results for total minutes of deep sleep for the devices versus polysomnography (PSG). For PSG, deep sleep was calculated as the N3 sleep stage. Results are shown for all devices that output sleep stage classifications. See Table 1 caption for additional table details.

Table 7. Sleep summary: rapid eye movement (REM) sleep

Device	*n*	PSG Mean ± *SD*	Device Mean ± *SD*	Bias	Lower limit	Upper limit	*t* (*p*)	Effect size	R^2 (*p*)
Fitbit Alta HR	49	107.4 ± 20.6	96.0 ± 22.9	–11.4	–60.2	37.3	–3.3 **(0.002)**	–0.52	0.01 (0.426)
Garmin Fenix 5S	29	111.4 ± 23.8	120.0 ± 37.8	8.6	–74.9	92.1	1.1 (0.277)	0.27	0.19 **(0.018)**
Garmin Vivosmart 3	43	109.7 ± 22.6	118.1 ± 40.4	8.4	–75.9	92.6	1.3 (0.200)	0.25	0.28 **(<0.001)**
Earlysense Live	51	104.7 ± 22.2	104.7 ± 43.1	0.0	–82.9	82.9	0.0 (0.997)	0.00	0.36 **(<0.001)**
ResMed S+	51	104.9 ± 20.7	73.0 ± 25.1	–31.9	–85.5	21.8	–8.5 **(<0.001)**	–1.37	0.04 (0.159)
SleepScore Max	42	110.4 ± 22.4	74.6 ± 23.0	–35.8	–95.3	23.7	–7.8 **(<0.001)**	–1.57	0.00 (0.856)

Summary results for total minutes of REM sleep for the devices versus polysomnography (PSG). Results are shown for all devices that output sleep stage classifications. See Table 1 caption for additional table details.

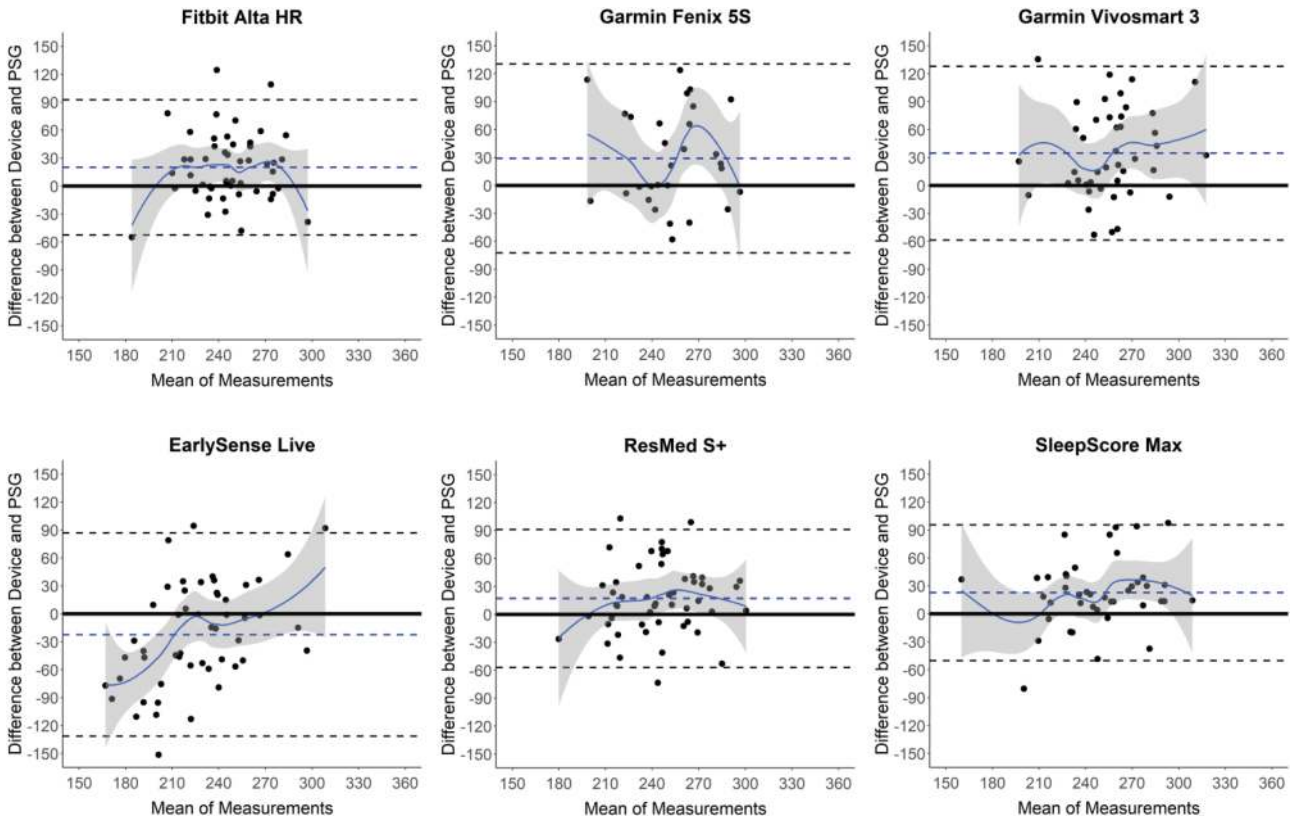

Figure 5. Bland–Altman plots: light sleep. Bland–Altman plots depicting the minutes of light sleep for the devices compared with polysomnography (PSG). For PSG, light sleep was calculated as the combination of N1 and N2 sleep stages. Only devices that output data on sleep stages are depicted. See Figure 1 caption for additional details on the interpretation of Bland–Altman plots.

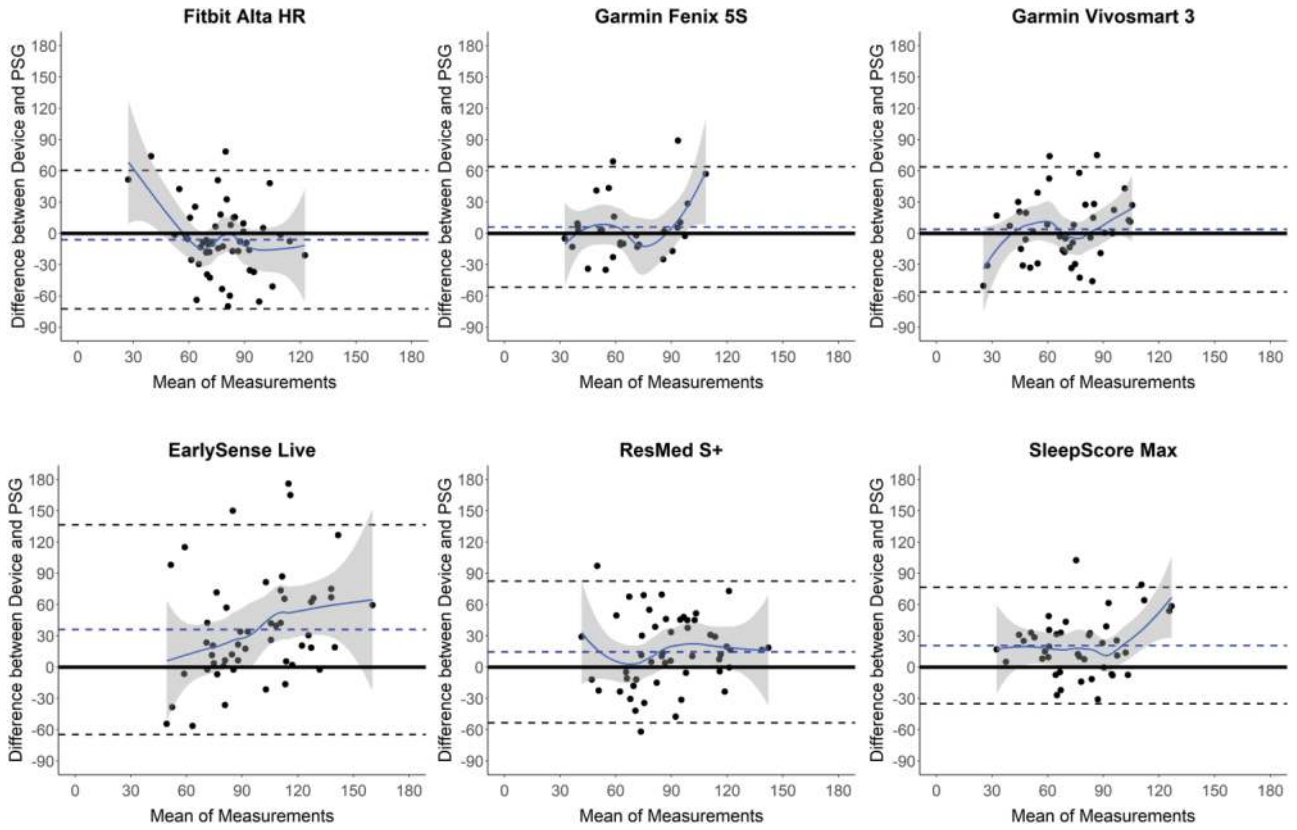

Figure 6. Bland–Altman plots: deep sleep. Bland–Altman plots depicting the minutes of deep sleep for the devices compared with polysomnography (PSG). For PSG, deep sleep was calculated as the N3 sleep stage. Only devices that output data on sleep stages are depicted. See Figure 1 caption for additional details on the interpretation of Bland–Altman plots.

to correctly detect PSG sleep epochs. However, specificity for the devices was variable, ranging from 0.18 to 0.54—indicating a lower ability for the devices to correctly detect PSG wake epochs than sleep epochs. Specificity for five of the seven consumer devices was substantially higher than the specificity of the Actiwatch (0.39), while both Garmin device models had the lowest specificities (0.18 and 0.19), indicating much worse performance as compared with the Actiwatch and all of the other consumer devices in detecting PSG wake epochs.

Other measures of EBE agreement with PSG (also shown in Table 8), largely reflect the sensitivity and specificity results. The values for PPV were all high and in a narrow range, indicating a high ability of device-scored sleep epochs to be reflected as PSG sleep. However, the NPV values were somewhat lower and more variable, indicating an overall worse ability for the device-scored wake epochs than the device-scored sleep epochs to be correctly reflected in the PSG scoring. Across devices, accuracy was high, and PABAK values were medium-to-high.

EBE classifications of individual sleep stages

Sleep stage EBE agreement versus PSG are shown for light, deep, and REM sleep stages in Tables 9, 10, and 11, respectively. In general, compared with the sensitivity and specificity of the sleep/wake EBE classifications (described previously), across sleep stages the sensitivity was relatively lower with a wider range of values and specificity was relatively higher with a more narrow range of values—indicating an overall poorer and inconsistent ability of devices to correctly detect PSG sleep stage epochs.

Among all sleep stages and devices, levels of sensitivity were in the medium range, while specificity levels were medium for light sleep but high for deep and REM. Notably, the Fitbit Alta HR had the highest values across most of the light and REM EBE agreement measures, while no specific device stood out for deep sleep. When devices misclassified PSG sleep stage epochs (Table 12), there were particularly high error rates for all devices to misclassify PSG wake, deep, and REM epochs as light sleep. Likewise, when PSG differed from the device-scored epochs (Supplemental Table S7), device epochs scored as wake, deep, and REM were often classified instead as light sleep. Misclassification errors between the other possible stage classifications were all comparatively low.

Discussion

Overall, the consumer sleep-tracking devices we tested had high sensitivity but relatively lower specificity, indicating a tendency for the devices to accurately detect sleep but to less accurately detect wake compared with the gold-standard sleep measurement technique PSG. Mixed results were found for the ability of devices to accurately detect sleep stages at the level of individual epochs or sleep summary measures. Notably, on several important performance measures, many of the consumer devices performed as well or even better than actigraphy. Like actigraphy, most devices also showed proportional bias on sleep/

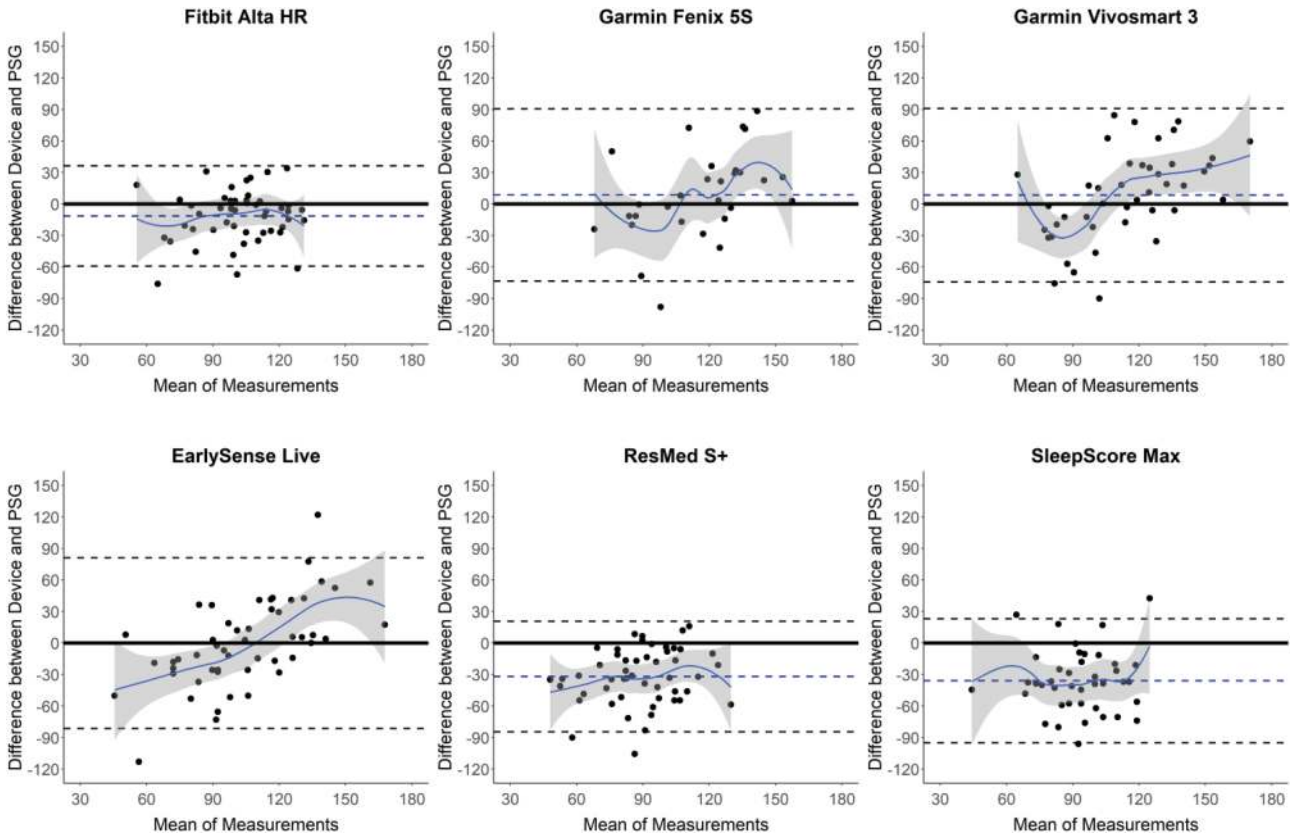

Figure 7. Bland–Altman plots: rapid eye movement (REM) sleep. Bland–Altman plots depicting the minutes of REM sleep for the devices compared with polysomnography (PSG). Only devices that output data on sleep stages are depicted. See Figure 1 caption for additional details on the interpretation of Bland–Altman plots.

Table 8. Epoch-by-epoch (EBE) agreement: sleep versus wake

Device	Sensitivity	Specificity	PPV	NPV	Accuracy	PABAK
Actiwatch	0.97	0.39	0.91	0.63	0.89	0.78
Fatigue Science Readiband	0.94	0.45	0.92	0.55	0.88	0.75
Fitbit Alta HR	0.95	0.54	0.94	0.58	0.90	0.80
Garmin Fenix 5S	0.99	0.18	0.88	0.74	0.88	0.74
Garmin Vivosmart 3	0.99	0.19	0.89	0.74	0.88	0.76
EarlySense Live	0.96	0.47	0.93	0.62	0.90	0.79
ResMed S+	0.93	0.51	0.93	0.51	0.88	0.75
SleepScore Max	0.94	0.50	0.92	0.56	0.88	0.75

Proportions for EBE agreement metrics are shown for sleep epochs (versus wake epochs) on all nights for the devices, compared with the corresponding epochs from polysomnography (PSG). Higher values (closer to 1.0) indicate better performance on that metric. PPV = positive predictive value, NPV = negative predictive value, PABAK = prevalence and bias-adjusted kappa.

Table 9. Epoch-by-epoch (EBE) agreement: light sleep epochs

Device	Sensitivity	Specificity	PPV	NPV	Accuracy	PABAK
Fitbit Alta HR	0.76	0.67	0.70	0.74	0.72	0.42
Garmin Fenix 5S	0.68	0.54	0.58	0.64	0.60	0.19
Garmin Vivosmart 3	0.70	0.55	0.60	0.66	0.63	0.24
EarlySense Live	0.57	0.69	0.64	0.62	0.63	0.25
ResMed S+	0.67	0.61	0.63	0.65	0.64	0.27
SleepScore Max	0.68	0.60	0.62	0.66	0.64	0.26

Proportions for EBE agreement metrics are shown for light sleep epochs (versus the combination of all other classifications—wake, deep, and REM) on all nights, compared with the corresponding polysomnography (PSG) epochs. Results are shown for all devices that output sleep stage classifications. See Table 8 caption for additional table details.

Table 10. Epoch-by-epoch (EBE) agreement: deep sleep epochs

Device	Sensitivity	Specificity	PPV	NPV	Accuracy	PABAK
Fitbit Alta HR	0.53	0.92	0.58	0.91	0.86	0.71
Garmin Fenix 5S	0.56	0.92	0.55	0.92	0.87	0.73
Garmin Vivosmart 3	0.56	0.92	0.54	0.93	0.87	0.73
EarlySense Live	0.68	0.84	0.46	0.93	0.81	0.62
ResMed S+	0.59	0.88	0.50	0.91	0.83	0.66
SleepScore Max	0.59	0.88	0.44	0.93	0.84	0.67

Proportions for EBE agreement metrics are shown for deep sleep epochs (versus the combination of all other classifications—wake, light, and REM) on all nights, compared with the corresponding polysomnography (PSG) epochs. Results are shown for all devices that output sleep stage classifications. See Table 8 caption for additional table details.

Table 11. Epoch-by-epoch (EBE) agreement: rapid eye movement (REM) sleep epochs

Device	Sensitivity	Specificity	PPV	NPV	Accuracy	PABAK
Fitbit Alta HR	0.69	0.94	0.77	0.91	0.89	0.77
Garmin Fenix 5S	0.54	0.84	0.51	0.86	0.77	0.53
Garmin Vivosmart 3	0.50	0.82	0.46	0.84	0.75	0.48
EarlySense Live	0.64	0.89	0.62	0.90	0.84	0.67
ResMed S+	0.50	0.95	0.71	0.87	0.85	0.69
SleepScore Max	0.49	0.95	0.74	0.86	0.84	0.68

Proportions for EBE agreement metrics are shown for REM sleep epochs (versus the combination of all other classifications—wake, light, and deep) on all nights, compared with the corresponding polysomnography (PSG) epochs. Results are shown for all devices that output sleep stage classifications. See Table 8 caption for additional table details.

Table 12. Epoch-by-epoch (EBE) agreement: device sleep stage misclassification errors

	Wake epochs			Light sleep epochs			Deep sleep epochs			REM sleep epochs		
Device	Light	Deep	REM	Wake	Deep	REM	Wake	Light	REM	Wake	Light	Deep
Fitbit Alta HR	0.35	0.01	0.09	0.06	0.12	0.06	0.03	0.43	0.01	0.05	0.24	0.02
Garmin Fenix 5S	0.54	0.06	0.22	0.02	0.10	0.20	0.00	0.40	0.04	0.00	0.46	0.04
Garmin Vivosmart 3	0.53	0.08	0.20	0.02	0.09	0.18	0.00	0.41	0.03	0.00	0.42	0.04
EarlySense Live	0.35	0.03	0.15	0.06	0.25	0.12	0.02	0.27	0.03	0.02	0.32	0.02
ResMed S+	0.38	0.01	0.09	0.08	0.19	0.06	0.05	0.35	0.01	0.06	0.42	0.02
SleepScore Max	0.40	0.01	0.09	0.07	0.20	0.05	0.05	0.35	0.01	0.05	0.44	0.02

Proportions for EBE misclassification errors of sleep stage epochs versus polysomnography (PSG). PSG-scored classifications are the larger column categories, with the three possible device-scored misclassifications under each category. Results are shown for all devices that output sleep stage classifications.

wake summary measures, tending to perform worse against PSG on nights with poorer/disrupted sleep (lower SE) or longer sleep latencies. Taken together, these findings indicate that several new consumer sleep-tracking devices demonstrate promise as potential valid alternatives to the current mobile sleep monitoring standard of actigraphy, at least for the measurement of sleep versus wake.

Many of the current findings are consistent with recent studies that also tested the performance of consumer sleep-tracking devices versus PSG. In general, we found that many of the sleep/wake summary biases that are common for actigraphy [11–17] and other consumer devices [22, 23, 40–44]—such as overestimating sleep (TST, SE) and underestimating wake (WASO)—were also observed for most consumer devices in the current study. Actigraphy significantly differed from PSG in all the major summary metrics, while many of the consumer devices either did not significantly differ from PSG or their biases were more modest than for actigraphy—with a few exceptions (primarily the Garmin devices, which typically had more extreme biases than actigraphy). The two primary outcomes for EBE agreement, sensitivity and specificity, exhibited the commonly found pattern for consumer devices to have high sensitivity and low or medium levels of specificity [22, 23]. This pattern is similar to the known bias for actigraphy to also have high sensitivity and lower specificity, and thus to better detect sleep epochs than wake epochs [11–17]. Importantly, though, most of the consumer devices (five of the seven) performed either as well as or better than actigraphy on specificity, the primary indicator of a device's wake-detection capability. The overall greatest specificity was for the Fitbit Alta HR, and it is notable that recent studies with the same device also found equal or greater specificity than actigraphy in comparison with PSG [27, 29]. Because actigraphy is the standard technique for mobile measurement of sleep/wake and has been validated against PSG, it is reasonable to suggest that one of the best initial benchmarks for judging the validity of consumer devices should be evaluating their performance relative to actigraphy versus PSG. Thus, based on sleep/wake summary outcomes and EBE sensitivity and specificity as primary comparisons, our findings demonstrate positive evidence for the capability of many new consumer devices to track PSG sleep and wake as accurately or better than actigraphy.

While actigraphy is limited to a binary classification of only sleep versus wake, many of the consumer devices also output metrics for individual sleep stages. Similar to many other recent studies that evaluated device sleep stage-tracking performance [26–29, 35–37, 45], we found that for both sleep summary metrics as well as EBE agreement that the consumer devices demonstrated mixed and often poor results for detection of sleep stages versus PSG. All devices significantly differed from PSG in total light sleep, mostly with overestimations. Half the devices also differed from PSG in the total amount of either deep or REM sleep, doing so in a consistent direction (overestimations for deep and underestimations for REM). However, all the sleep stage summary results should especially be considered in the context of the Bland–Altman plots, which had large spreads of individual night data across both axes. Thus, the devices that *did not* have significant sleep stage biases versus PSG may only have averaged out to be similar to PSG because of high variability, with a mix of underestimations *and* overestimations across individual nights. Compared with the overall sleep/wake EBE results, sensitivity was relatively lower and specificity was relatively higher for each of the sleep stages, indicating that consumer devices perform only at a medium level for correctly detecting PSG sleep stages. Most devices failed to correctly identify 30%–50% of both deep sleep and REM sleep, on average. Thus, there is an overall substantial likelihood for devices to misclassify all sleep stages, and the misclassified sleep stages are most often called light sleep.

In this study, we included only healthy individuals with no reported sleep problems and fixed TIB to 8 h—thus, our findings are best generalized to the performance of devices in healthy sleepers on nights with a clinically recommended TIB. However, there was variability between nights and we calculated proportional bias to determine whether that variability affected the night-to-night reliability of devices versus PSG. While proportional bias always occurred for actigraphy, we found a mix of results for the consumer devices. Overall, there were significant proportional biases for at least several devices on most summary measures, which were driven by the higher levels of variability and bias on nights with poorer/disrupted sleep (i.e. longer sleep latency, lower TST and SE, and higher WASO). Thus, the proportional bias patterns that actigraphy has on nights with poor/disrupted sleep are also found in consumer devices. For sleep stages, the devices overall did not show many proportional biases, which may be due to the high variability when estimating sleep stages versus PSG in general—therefore few proportional bias patterns could emerge. Importantly, these proportional biases could have implications for the overall accuracy of devices on poor/disrupted nights of sleep in healthy individuals, and generally for those with insomnia. Of note, a recent device performance study in insomnia patients [29] using actigraphy and the same Fitbit model (Alta HR) found less bias overall and fewer proportional biases versus PSG than for actigraphy. Together, these findings suggest that device performance is worse on nights with poor/disrupted sleep, but that most devices still perform as well as or better than actigraphy, even as summary measures vary across healthy or sleep-disordered populations.

Considering all the findings, which device(s) performed best? This is a somewhat difficult question to answer, as there is no established model for weighting or rank-ordering the various device performance metrics. Therefore, our approach to answering this question is primarily driven by considering what makes a consumer sleep-tracking device useful for sleep research or clinical sleep testing, i.e. can a device perform as well as or better than—and, thus, could it be considered for use as an alternative to—standard actigraphy?

As previously discussed, actigraphy has served as the mobile sleep assessment standard for several decades, with its performance and limitations having been reviewed extensively [11–17]. Regarding its sleep/wake-tracking performance, actigraphy is: (1) good at detecting sleep epochs (reflected in high EBE sensitivity levels) but it overestimates sleep measures such as TST and SE, (2) worse at detecting wake epochs (reflected in poor-to-medium EBE specificity levels) and underestimates wake measures such as SOL and WASO, and (3) more accurate on nights with good sleep than poor/disrupted sleep. Let us consider each point with the current findings: (1) We found that actigraphy and all the consumer devices had high sensitivity levels, indicating that consumer devices perform as well as actigraphy in EBE sleep detection. Also, actigraphy significantly overestimated TST and SE while only some of the consumer devices did so, suggesting equivalent or better performance for consumer devices. (2) Specificity for actigraphy was low but was substantially higher in most consumer devices (all except the Garmin devices, which were much worse), and actigraphy significantly underestimated SOL and WASO whereas only some consumer devices did so. Therefore, on wake detection—the greatest weakness of actigraphy performance—most devices performed better than actigraphy. (3) Actigraphy had the usual pattern of worse performance on nights with poor/disrupted sleep, indicated by the Bland–Altman individual night data and significant proportional biases. Most consumer devices demonstrated a similar proportional bias pattern, indicating that they also tend to perform less accurately on poorer/disrupted nights of sleep. Thus, based on these major aspects of actigraphy performance, we conclude that most of the devices we tested performed as well as or better than actigraphy.

Although we conclude that most of the devices performed well, did any specific devices perform best or worst? Across most of the performance metrics, it appears that the Fitbit Alta HR consistently performed either as the best or among the best, and thus could be considered the top performer of the devices we tested. On summary outcomes, the Fitbit Alta HR tracked the sleep/wake metrics very closely with low bias. On EBE outcomes, the Alta HR notably had the highest level of specificity. A recent review [46] found that newer-generation Fitbit models have improved sleep-tracking performance compared with earlier models, and that Fitbit performance is overall good for sleep/wake classification but not for sleep stages—consistent with our current findings. Further, regarding the Alta HR model, several recent studies [27–29, 36]—comprising different age groups and sleep health statuses—also found good sleep-tracking performance versus PSG but still found relatively poor performance for tracking sleep stages. Future studies should further investigate whether new Fitbit models remain as top-performing sleep-tracking devices. The Fatigue Science Readiband, EarlySense Live, ResMed S+, and SleepScore Max were the other top-performing devices, each demonstrating relatively high performance across most metrics that was also as good as or better than actigraphy. The worst-performing devices were the two Garmins, which ranked last on most performance metrics and with biases that were often more extreme than actigraphy. Based on their high

EBE NPV values, it is possible that the poor performance of the Garmin devices is in part due to having a relatively higher threshold for wake (and a lower threshold for sleep) than the other devices, which resulted in the Garmin devices performing poorly for tracking both sleep and wake because so much actual PSG wake was missed—but the wake epochs it did detect were more likely to indeed be PSG wake than the other devices.

One of the strengths of our study is that it was carried out independently [23]. To our knowledge, except for the Fitbit Alta HR, the other device models we tested have not undergone prior independent performance testing against PSG. However, there were notable nonindependent studies (i.e. studies that were directly funded/supported by the device company, and one or more employees were authors) conducted for two of the devices. The first study [24], for the EarlySense Live, found that it performed well against PSG and even better in many regards than our current findings (e.g. they reported similar sensitivity but much higher specificity, and no differences from PSG in TST or sleep stages). The second study [26], for the ResMed S+, also found good performance versus PSG and was better than actigraphy in most sleep/wake outcomes (e.g. a similar level of sensitivity but much higher specificity than the current study, but they found an overestimation of TST and some different sleep stage biases). Thus, except for the Fitbit Alta HR, to our knowledge, the current study represents the first *independent* performance study for these devices and, more broadly, either the first or one of the first studies overall for the devices. Additional studies are therefore warranted to further evaluate the sleep-tracking performance of these devices beyond the current results.

We should also consider the *method* of actigraphy and how that relates to the use of consumer devices. For actigraphy, it remains a standard practice in postprocessing to input TIB data into the actigraphy program to ensure accuracy [9–12]. For consumer devices, however, there are no intended postprocessing steps, as most device algorithms are designed to carry out sleep-tracking passively and automatically. It is of concern though that the ability of device algorithms to determine TIB without user input may lead to errors that affect their accuracy. This can be caused by an algorithm "detecting" the start or end of a sleep episode earlier or later than it should have, resulting in a shortened or lengthened TIB (see Supplemental Materials for more discussion on this issue). If consumer devices are to be considered for use as alternatives to actigraphy in research and/or clinical testing, it would be important to conduct some basic level of postprocessing (like with actigraphy) to improve data integrity. In this study, we conducted such postprocessing (as needed) in order to capture device data over the 8-h TIB, allowing for direct comparison of devices with actigraphy and PSG. It is encouraging that some newer devices actually do allow the user to edit or postprocess their sleep data in the device app (e.g. to correct TIB errors, or to input missed sleep episodes). Individual users and future studies could benefit from the use of such tools to improve the accuracy of sleep-tracking data from consumer devices.

There were both strengths and limitations to the study. In addition to this being an independent study, the other major strengths were the study design, scope of the study, and testing conditions. For example, testing multiple device models and different types of devices (wearable and nonwearable), utilizing gold-standard comparisons (PSG and actigraphy), including a sleep disruption condition night, having well-defined participant inclusion criteria, and maintaining prestudy and within-study conditions important for laboratory sleep testing. Further, the data analysis is broad and comprehensive, and complies with recently recommended standards [23, 34] for performance testing of sleep-tracking devices. Limitations of the study include: not being able to test all seven consumer devices at once or on all participants (due to comfort and other physical constraints, availability of devices, and to prevent possible interference between devices), occasional missing data or partial data loss, the controlled laboratory setting and PSG/device applications can affect sleep itself, and the fixed sleep episode times and durations do not reflect natural variation in TIB. We did not collect race/ethnicity demographic data from participants, however, future studies should do so to examine possible race/ethnicity differences in device performance. This can be relevant for testing wearables that have photoplethysmography (PPG)-based heart rate sensors (like the Fitbit and Garmin devices), due to concerns regarding PPG sensor accuracy in people with darker skin tones and minority race/ethnic groups who are often under-represented in device studies in general [47]. Device companies occasionally update their device models, apps, and sleep-tracking algorithms—therefore the current results correspond best to the versions used during the study period. Nonetheless, our group has previously published data [29, 48] showing updates to one specific device utilized here (Fitbit Alta HR) during a ~2-year data collection period did not impact sensitivity, specificity, and accuracy values for any sleep measure. Additionally, we used the medium sensitivity threshold setting for the Actiwatch analysis, as this is the default and most commonly used actigraphy setting for healthy young adults. It is possible that use of different actigraphy settings could have produced slightly different results, however that may come with tradeoffs (e.g. a lower threshold may result in increased EBE specificity, but that may also cause lower EBE sensitivity). Ultimately, researchers should use the recommended actigraphy settings that best match the participants and conditions of their study.

In summary, our findings demonstrate that many (but not all) new consumer sleep-tracking devices perform as well as, or better than, actigraphy on sleep/wake summary and EBE performance metrics—indicating that some consumer devices are promising in their initial validity versus the gold-standard PSG. Additional device performance testing is warranted, especially in other populations and settings, to further define and reveal the strengths, weaknesses, and limitations, as well as broaden the scope for the use of consumer devices to track sleep. To initially address ecological validity, in our prior study the same participants with insomnia were tested in the lab [29] and at home [48], and the Fitbit Alta HR had comparable performance between settings. Compared with PSG and actigraphy, there are benefits to utilizing consumer devices due to their wide availability, diverse device types, the low burden on the user, relatively low cost, multiple sensors and functions, and ability to rapidly sync with phones or computers to provide real-time data. Thus, devices have the capacity to be used daily, help establish normative sleep data for an individual, and be used toward meaningful and actionable endpoints. For example, sleep-tracking is urgently needed in areas where PSG and actigraphy cannot be adequately used, such as operational settings (like the military) where device data could be utilized as inputs in fatigue models for optimized scheduling, risk mitigation, and performance

enhancement. Clinically, devices could be used to identify when an individual's sleep pattern (or sudden change in sleep pattern) may signify a sleep disorder that should be evaluated, and for tracking compliance and progress in a sleep disorder treatment plan. The wide use, rapid technological advancement, and promising initial research findings demonstrating the improved sleep-tracking performance of many recent-generation consumer devices warrant further testing versus gold-standards in different conditions, populations, and settings in order to evaluate their wider validity and utility—toward their consideration as possible valid alternatives to actigraphy.

Acknowledgments

PSG scoring services were performed by RPSGT staff from the Sleep and Behavioral Neuroscience Center at the University of Pittsburgh Medical Center. The authors wish to thank the study participants, as well as research staff members from the Warfighter Performance Department at the Naval Health Research Center who contributed to data collection. We want to acknowledge the following companies for providing access to EBE device data: EarlySense, Ltd., ResMed, Inc., Fatigue Science, SleepScore Labs.

Disclosure Statement

The authors declare no financial or nonfinancial conflicts of interest. Additionally, none of the authors have any conflicts of interest related to the companies whose devices were evaluated in the study. There was no funding or other financial support to this research from any device companies, nor did the companies have any involvement in any stage of the research including study design, selection of devices, data analysis, interpretation, writing or reviewing of the manuscript.

Funding

This research was funded by the Office of Naval Research, Code 34.

Disclaimer

I am a military service member or federal/contracted employee of the US Government. This work was prepared as part of my official duties. Title 17, U.S.C. §105 provides that copyright protection under this title is not available for any work of the US Government. Title 17, U.S.C. §101 defines a US Government work as work prepared by a military service member or employee of the US Government as part of that person's official duties. This work was supported by the Office of Naval Research, Code 34, under work unit no. N1701. The views expressed in this article reflect the results of research conducted by the authors and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, nor the US Government. The study protocol was approved by the Naval Health Research Center Institutional Review Board in compliance with all applicable federal regulations governing the protection of human subjects. Research data were derived from an approved Naval Health Research Center Institutional Review Board protocol, number NHRC.2017.0008.

References

1. Breslau N, *et al*. Sleep disturbance and psychiatric disorders: a longitudinal epidemiological study of young adults. *Biol Psychiatry*. 1996;**39**(6):411–418.
2. Knutson KL, *et al*. The metabolic consequences of sleep deprivation. *Sleep Med Rev*. 2007;**11**(3):163–178.
3. Luyster FS, *et al*.; Boards of Directors of the American Academy of Sleep Medicine and the Sleep Research Society. Sleep: a health imperative. *Sleep*. 2012;**35**(6):727–734.
4. Baglioni C, *et al*. Sleep and mental disorders: a meta-analysis of polysomnographic research. *Psychol Bull*. 2016;**142**(9):969–990.
5. Belenky G, *et al*. Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: a sleep dose–response study. *J Sleep Res*. 2003;**12**(1):1–12.
6. Van Dongen HPA, *et al*. The cumulative cost of additional wakefulness: dose-response effects on neurobehavioral functions and sleep physiology from chronic sleep restriction and total sleep deprivation. *Sleep*. 2003;**26**(2):117–126. doi:10.1093/sleep/26.2.117
7. Goel N, *et al*. Neurocognitive consequences of sleep deprivation. *Semin Neurol*. 2009;**29**(4):320–339.
8. Iber C, *et al*. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*. Westchester, IL: American Academy of Sleep Medicine; 2007.
9. Cole RJ, *et al*. Automatic sleep/wake identification from wrist activity. *Sleep*. 1992;**15**(5):461–469.
10. Sadeh A, *et al*. The role of actigraphy in the evaluation of sleep disorders. *Sleep*. 1995;**18**(4):288–302.
11. Ancoli-Israel S, *et al*. The role of actigraphy in the study of sleep and circadian rhythms. *Sleep*. 2003;**26**(3):342–392.
12. Morgenthaler T, *et al*. Practice parameters for the use of actigraphy in the assessment of sleep and sleep disorders: an update for 2007. *Sleep*. 2007;**30**(4):519–529. doi:10.1093/sleep/30.4.519
13. Paquet J, *et al*. Wake detection capacity of actigraphy during sleep. *Sleep*. 2007;**30**(10):1362–1369.
14. Sadeh A. The role and validity of actigraphy in sleep medicine: an update. *Sleep Med Rev*. 2011;**15**(4):259–267. doi:10.1016/j.smrv.2010.10.001
15. Van de Water ATM, *et al*. Objective measurements of sleep for non-laboratory settings as alternatives to polysomnography--a systematic review. *J Sleep Res*. 2011;**20**(1 Pt 2):183–200. doi:10.1111/j.1365-2869.2009.00814.x
16. Marino M, *et al*. Measuring sleep: accuracy, sensitivity, and specificity of wrist actigraphy compared to polysomnography. *Sleep*. 2013;**36**(11):1747–1755.
17. Grandner MA, *et al*. Chapter 12 – Actigraphic sleep tracking and wearables: historical context, scientific applications and guidelines, limitations, and considerations for commercial sleep devices. In: Grandner MA, ed. *Sleep and Health*. Cambridge, MA: Academic Press; 2019: 147–157. doi:10.1016/B978-0-12-815373-4.00012-5
18. Kolla BP, *et al*. Consumer sleep tracking devices: a review of mechanisms, validity and utility. *Expert Rev Med Devices*. 2016;**13**(5):497–506.
19. Khosla S, *et al*.; American Academy of Sleep Medicine Board of Directors. Consumer sleep technology: an American Academy of Sleep Medicine position statement. *J Clin Sleep Med*. 2018;**14**(5):877–880.
20. Baron KG, *et al*. Feeling validated yet? A scoping review of the use of consumer-targeted wearable and mobile

technology to measure and improve sleep. *Sleep Med Rev*. 2018;**40**:151–159.
21. Bianchi MT. Sleep devices: wearables and nearables, informational and interventional, consumer and clinical. *Metabolism*. 2018;**84**:99–108.
22. de Zambotti M, *et al*. Wearable sleep technology in clinical and research settings. *Med Sci Sports Exerc*. 2019;**51**(7):1538–1557.
23. Depner CM, *et al*. Wearable technologies for developing sleep and circadian biomarkers: a summary of workshop discussions. *Sleep*. 2020;**43**(2). doi:10.1093/sleep/zsz254
24. Tal A, *et al*. Validation of contact-free sleep monitoring device with comparison to polysomnography. *J Clin Sleep Med*. 2017;**13**(3):517–522.
25. Jumabhoy R, *et al*. Wrist-worn activity monitoring devices overestimate sleep duration and efficiency in health adults. *Sleep*. 2017;**40**(Suppl 1):A288–A289. doi:10.1093/sleepj/zsx050.778
26. Schade MM, *et al*. Sleep validity of a non-contact bedside movement and respiration-sensing device. *J Clin Sleep Med*. 2019;**15**(7):1051–1061. doi:10.5664/jcsm.7892
27. Lee XK, *et al*. Validation of a consumer sleep wearable device with actigraphy and polysomnography in adolescents across sleep opportunity manipulations. *J Clin Sleep Med*. 2019;**15**(9):1337–1346.
28. Moreno-Pino F, *et al*. Validation of Fitbit Charge 2 and Fitbit Alta HR against polysomnography for assessing sleep in adults with obstructive sleep apnea. *J Clin Sleep Med*. 2019;**15**(11):1645–1653.
29. Kahawage P, *et al*. Validity, potential clinical utility, and comparison of consumer and research-grade activity trackers in Insomnia Disorder I: in-lab validation against polysomnography. *J Sleep Res*. 2020;**29**(1):e12931.
30. Liu S. *Fitness & Activity Tracker – Statistics & Facts*. Published May 22, 2019. https://www.statista.com/topics/4393/fitness-and-activity-tracker/. Accessed January 28, 2020.
31. Moar J. *Where Now for Wearables?* Published March 21, 2018. https://www.juniperresearch.com/document-library/white-papers/where-now-for-wearables. Accessed January 28, 2020.
32. Kunst A. *Percentage of U.S. Adults that Use Apps to Track their Sleep as of 2017, by Gender*. Published December 20, 2019. https://www.statista.com/statistics/699434/us-adults-that-use-apps-to-track-sleep-by-gender/. Accessed January 28, 2020.
33. Syngene Research. *Global Smart Sleep Tracking Device Market Analysis 2019*. Published November 2019. https://www.researchandmarkets.com/reports/4857841/global-smart-sleep-tracking-device-market. Accessed January 28, 2020.
34. Menghini L, *et al*. A standardized framework for testing the performance of sleep-tracking technology: step-by-step guidelines and open-source code. *Sleep*. 2020. doi:10.1093/sleep/zsaa170
35. de Zambotti M, *et al*. A validation study of Fitbit Charge 2™ compared with polysomnography in adults. *Chronobiol Int*. 2018;**35**(4):465–476. doi:10.1080/07420528.2017.1413578
36. Cook JD, *et al*. Ability of the Fitbit Alta HR to quantify and classify sleep in patients with suspected central disorders of hypersomnolence: a comparison against polysomnography. *J Sleep Res*. 2019;**28**(4):e12789.
37. de Zambotti M, *et al*. The sleep of the ring: comparison of the ŌURA Sleep Tracker against polysomnography. *Behav Sleep Med*. 2017;17(2):124–136. doi:10.1080/15402002.2017.1300587
38. Bland JM, *et al*. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;**1**(8476):307–310.
39. Bland JM, *et al*. Measuring agreement in method comparison studies. *Stat Methods Med Res*. 1999;**8**(2):135–160.
40. Scott H, *et al*. A systematic review of the accuracy of sleep wearable devices for estimating sleep onset. *Sleep Med Rev*. 2020;**49**:101227.
41. Montgomery-Downs HE, *et al*. Movement toward a novel activity monitoring device. *Sleep Breath*. 2012;**16**(3):913–917.
42. de Zambotti M, *et al*. Evaluation of a consumer fitness-tracking device to assess sleep in adults. *Chronobiol Int*. 2015;**32**(7):1024–1028.
43. Meltzer LJ, *et al*. Comparison of a commercial accelerometer with polysomnography and actigraphy in children and adolescents. *Sleep*. 2015;**38**(8):1323–1330.
44. de Zambotti M, *et al*. Validation of sleep-tracking technology compared with polysomnography in adolescents. *Sleep*. 2015;**38**(9):1461–1468.
45. Mantua J, *et al*. Reliability of sleep measures from four personal health monitoring devices compared to research-based actigraphy and polysomnography. *Sensors (Basel)*. 2016;**16**(5):646. doi:10.3390/s16050646
46. Haghayegh S, *et al*. Accuracy of wristband Fitbit models in assessing sleep: systematic review and meta-analysis. *J Med Internet Res*. 2019;**21**(11):e16273.
47. Colvonen PJ, *et al*. Limiting racial disparities and bias for wearable devices in health science research. *Sleep*. 2020;**43**(10). doi:10.1093/sleep/zsaa159
48. Hamill K, *et al*. Validity, potential clinical utility and comparison of a consumer activity tracker and a research-grade activity tracker in insomnia disorder II: outside the laboratory. *J Sleep Res*. 2020;**29**(1):e12944.