ORIGINAL RESEARCH

Daytime Sleep-Tracking Performance of Four Commercial Wearable Devices During Unrestricted Home Sleep

Evan D Chinoy [iD][1,2], Joseph A Cuellar[1,2], Jason T Jameson[1,2], Rachel R Markwald[1]

[1]Sleep, Tactical Efficiency, and Endurance Laboratory, Warfighter Performance Department, Naval Health Research Center, San Diego, CA, USA; [2]Leidos, Inc, San Diego, CA, USA

Correspondence: Rachel R Markwald, Sleep, Tactical Efficiency, and Endurance Laboratory, Warfighter Performance Department, Naval Health Research Center, 140 Sylvester Road, San Diego, CA, 92106, USA, Tel +1 619 767 4494, Email rachel.r.markwald.civ@health.mil

Purpose: Previous studies have found that many commercial wearable devices can accurately track sleep-wake patterns in laboratory or home settings. However, nearly all previous studies tested devices under conditions with fixed time in bed (TIB) and during nighttime sleep episodes only. Despite its relevance to shift workers and others with irregular sleep schedules, it is largely unknown how devices track daytime sleep. Therefore, we tested the sleep-tracking performance of four commercial wearable devices during unrestricted home daytime sleep.

Participants and Methods: Participants were 16 healthy young adults (6 men, 10 women; 26.6 ± 4.6 years, mean ± SD) with habitual daytime sleep schedules. Participants slept at home for 1 week under unrestricted conditions (ie, self-selecting TIB) using a set of four commercial wearable devices and completed reference sleep logs. Wearables included the Fatigue Science ReadiBand, Fitbit Inspire HR, Oura Ring, and Polar Vantage V Titan. Daytime sleep episode TIB biases and frequencies of missed and false-positive daytime sleep episodes were examined.

Results: TIB bias was low in general for all devices on most daytime sleep episodes, but some exhibited large biases (eg, >1 h). Total missed daytime sleep episodes were as follows: Fatigue Science: 3.6%; Fitbit: 4.8%; Oura: 6.0%; Polar: 37.3%. Missed episodes occurred most often when TIB was short (eg, naps <4 h).

Conclusion: When daytime sleep episodes were recorded, the devices generally exhibited similar performance for tracking TIB (ie, most episodes had low bias). However, the devices failed to detect some daytime episodes, which occurred most often when TIB was short, but varied across devices (especially Polar, which missed over one-third of episodes). Findings suggest that accurate daytime sleep tracking is largely achievable with commercial wearable devices. However, performance differences for missed recordings suggest that some devices vary in reliability (especially for naps), but improvements could likely be made with changes to algorithm sensitivities.

Keywords: validation, consumer sleep technology, naps, habitual sleep, shift work, sleep diary

Introduction

Sleep-tracking technology for consumer, research, or clinical use has grown rapidly over the past decade, with new devices and algorithms being released onto the market faster than researchers can evaluate their performance.[1,2] While this upward trend in use and access to sleep-tracking technology has brought much deserved and overdue attention to the importance of sleep among health-care professionals and the public, the strengths and limitations of particular devices and algorithms are still being determined (and may change over time as technologies are regularly updated).[3,4] Developing standards and best practices will rely on robust and rapid testing under complementary protocols to evaluate different areas of device performance.[2,5–8] This information is essential for obtaining accurate and reliable sleep device data for individual consumers and for ensuring that rigorous standards are met for research or clinical uses.

151

An important area of wearable device performance is their ability to track daytime sleep episodes. However, relatively little is known about their performance under daytime conditions, as these devices are nearly always tested for their ability to track nighttime sleep episodes only. Also, devices are most often tested under controlled conditions such as in a sleep laboratory under a fixed time in bed (TIB). This type of testing protocol typically utilizes gold-standard polysomnography (PSG) as a reference, because PSG provides high-quality data on initial device sleep-tracking performance and thus has been recommended as the first step for evaluating devices as part of the "validation" process.[9] Most recent studies of newer devices have indeed found that their sleep-tracking performance is generally good, with many devices performing better than the first commercial wearable devices released around a decade ago,[10–12] and at levels that meet or exceed research-grade actigraphy[13–17] (the standard method for mobile sleep-wake tracking).[18] However, daytime sleep tracking has still been absent from device performance evaluation studies,[1,6,19] despite its clear relevance for individuals and communities who often sleep at irregular or daytime hours and in shorter or multiple daily bouts (eg, shift workers, military personnel, first responders, athletes, children, and patients with certain sleep or circadian disorders). For many of those people, daytime sleep is an essential part of their overall sleep and plays a critical role in maintaining their optimal health and performance that comes from achieving the recommended total daily sleep duration.[20] This lack of research into daytime sleep evaluation with new devices is underscored by recent guidance from a group of sleep technology experts from the Sleep Research Society, who concluded that the "current evidence does not support the use of consumer wearable devices for daytime sleep assessment" and therefore recommended that devices "should not be used to assess daytime sleep or naps" for research or clinical purposes.[6]

One of the key measures of a device's real-world sleep-tracking performance is potential bias in the TIB domain.[21] Accurate TIB readings (ie, with small bias) are critical, because if a device cannot accurately identify the actual window of time when an individual attempts to sleep, then the other sleep-tracking outcomes it generates (eg, total sleep time (TST), sleep efficiency, sleep onset latency, wake after sleep onset) will likely be biased as well. Unfortunately, many real-world habits during wake (such as sedentariness) produce patterns of physical and physiological activity (eg, reduced movement and heart rate) that may be difficult to differentiate from signals generated when someone tries to fall asleep, which could confuse device sleep-tracking algorithms. Despite its importance, TIB tracking has also not been an outcome variable in most device performance evaluation studies because testing protocols largely control for environment, participant behavior, and TIB. In our previous device performance evaluation study of nighttime sleep under unrestricted real-world conditions,[16] we found that four commercial wearable devices generally exhibited good performance with low biases for TIB and most other sleep-wake outcomes on the majority of nights. However, on some nights TIB biases were large (eg, >1 h), which is a concern for the reliability and accuracy of data from wearables that automatically and passively track sleep patterns in real-world conditions. Such biases, even when occurring infrequently, could lead to reduced trust in a device for monitoring sleep and greater risk for discontinuing use of the device altogether. Additionally, it is likely that some proprietary commercial device algorithms are designed primarily for detecting nighttime sleep, so it is unknown how well these commercial devices perform at recording daytime sleep episodes at all.

Therefore, in the current study, we evaluated the daytime sleep-tracking performance of four commercial wearable devices that we tested previously under unrestricted nighttime conditions at home.[16] In particular, we aimed to evaluate the performance of four wearable devices to track daytime sleep episodes for (1) bias in sleep timing outcomes (TIB, start time of sleep episode, end time of sleep episode) and (2) the frequencies of missed daytime sleep episodes and false-positive episodes mislabeled as sleep by the devices.

Participants and Methods
Participants

Healthy young adults (n=16, 10 women, 6 men; 26.6 ± 4.6 years, mean ± SD) participated. Screening consisted of a self-report medical history questionnaire that assessed the following exclusion criteria: age <18 or >40 years, body mass index <18.5 or ≥30.0 kg/m^2, any diagnosed sleep, mental health, or other medical disorder, use of any illegal drugs or any sleep medications (over-the-counter or prescription) in the previous month, current pregnancy, and any physical or living conditions affecting the ability to sleep uninterrupted. Participants also had to report a habitual schedule of regular

daytime sleep, which, for enrollment, was defined as sleeping between 06:00 and 22:00 with TIB ≥1 h at least twice weekly.

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

Study Protocol

Participants were given a set of four commercial sleep-tracking devices to use, while they slept at home for one week. They also completed detailed sleep logs to report their sleep schedule after each sleep episode. The timing and duration of their sleep episodes were self-selected and unrestricted, except for the requirement to attempt a minimum of two daytime (occurring between 06:00 and 22:00) sleep episodes during the study week. Thus, the number, timing, and sequence of daytime and nighttime sleep episodes logged by each participant varied, and only the qualifying daytime sleep episodes were included in the final analyses (see Supplemental Figures S1 and S2 for depictions of individual participant sleep schedules).

The devices were worn simultaneously during all sleep episodes, along with a research-grade actigraphy watch (Actiwatch 2; Philips Respironics, Inc.; Murrysville, PA, USA) that was not part of the current analyses. Within one hour of awakening after each sleep episode across the study week (whether occurring in daytime, nighttime, or as a nap), participants were instructed to report their bed and wake times, and the times they physically entered and exited bed, using a digital sleep log (based on the consensus sleep diary[22]) using an iPad tablet computer (Apple Inc.; Cupertino, CA, USA) that we provided. The sleep log was implemented in Smartabase (Fusion Sport; Broomfield, CO, USA), a data capture and management system application ("app"). Participants were also instructed to wear the devices as much as possible when not sleeping, but they could remove them during wake times if needed (eg, charging the devices, whenever devices could interfere with work duties, while showering or doing activities that could damage or submerge the devices). In such cases, participants were instructed to log these times using the Smartabase app.

Alcohol intake was not allowed during the study week. No other restrictions were placed on participants during the study week. They could consume caffeine and engage in other activities such as exercise during the study, but they were instructed to log these behaviors daily on the Smartabase app before each sleep episode. To improve the accuracy of sleep log entries and to ensure that the devices were charged and working correctly, researchers completed daily compliance checks online by viewing data on individual device apps and Smartabase and, if needed, promptly provided feedback to participants to enter their log data and sync devices.

Sleep-Tracking Devices

The same four commercial sleep-tracking wearable devices were tested in this study as in our previous study,[16] in which we evaluated the sleep-tracking performance of a set of devices during unrestricted nighttime sleep with a similar weeklong protocol design. However, unlike in our previous study, participants did not wear an electroencephalography headband device. The four devices tested were the Fatigue Science ReadiBand™ (version 5; Fatigue Science; Vancouver, BC, Canada), Fitbit™ Inspire HR (Fitbit, Inc.; San Francisco, CA, USA), Oura™ Ring (Gen 2; ŌURA Health Oy; Oulu, Finland), and Polar® Vantage V Titan (Polar Electro Oy; Kempele, Finland).

At the start of the study, researchers reviewed the procedures with participants and instructed them on how to wear the devices comfortably and correctly. The four wrist devices (which included Actiwatch 2; data not shown) were worn in pairs on each wrist in the same manner as in our previous study.[16] The Actiwatch 2 and Polar Vantage V Titan were worn as a pair on one wrist, and the Fatigue Science ReadiBand and Fitbit Inspire HR were worn as a pair on the other wrist (with the Polar and Fitbit devices always worn closer to the wrist, because those two devices contained photo-plethysmography [PPG] heart rate sensors that should be placed closer to the wrist). Wrist placement for each pair of devices was counterbalanced between participants, with half assigned to wear each pair on either their dominant or non-dominant wrist at the beginning of the study. After the fourth study day, participants were instructed to switch the device pairs to the other wrist, allowing an approximately equal number of assessment nights for each wrist device on either the dominant or non-dominant wrist. However, participants wore the Oura Ring on their non-dominant ring finger during the

entire study week. Participants were instructed to keep devices charged as needed and to sync the data using their respective apps on the tablet within an hour of waking up from each sleep episode. See Supplemental Materials for details on device software and firmware versions.

Data Export and Preprocessing Procedures

Device data were exported from the respective device apps and from the following online device account website portals from each device company, which allowed additional data access and management of participant accounts: Readi for the Fatigue Science ReadiBand, Oura Teams for the Oura Ring, and Polar AccessLink API for the Polar Vantage V Titan. Fitbit Inspire HR data were exported via Fitabase (Small Steps Labs, LLC; San Diego, CA, USA), a licensed third-party data management platform that allows access to Fitbit account data. Apps and online device account portals sometimes did not register data from every sleep episode that was present in the sleep log. Therefore, sleep episode data that were not present originally on the online device account portal but were present on the device app, or vice versa, were all exported or transcribed from either data source for inclusion in the final analyses. At the time of data collection, certain data export features were not available in Oura Teams and have since been added, including the export of all "sleep periods" (this includes episodes categorized as "rest" in addition to those categorized as sleep or naps). Although these additional data were available, some identified "rest" periods still did not meet the sleep duration criteria for a nap as defined by Oura (TST between 15 min and 3 h). Thus, "rest" periods with less than 15 min of scored TST did not qualify as naps and were not included in the analyses. Additionally, when a sleep episode occurs that is identified by the Oura algorithm as a nap or likely nap, the Oura app prompts the user to "confirm" the nap so that it can be added into the sleep data output (which also uses the nap to update the daily readiness and sleep scores). However, the nap confirmation prompt is available only within the same day. Since data collection in this study was passive and participants were not instructed specifically to confirm naps using the app prompts, it is possible that some naps were "unconfirmed" by the participant. As a result, these naps may instead have been categorized as a "rest" period and would thus have appeared in the Oura app activity data log, not in the app's sleep data. However, unconfirmed naps would appear in the Oura Teams sleep periods export file and were used for analysis. For each device, TIB was the total time from the start to the end of the recording for each sleep episode (ie, the TST and the total wake occurring within each recorded episode).

All 16 participants completed the weeklong study protocol. However, because sleep schedules were self-selected and unrestricted, the total number of sleep episodes that each participant had over the study week varied and the number of episodes that were classified into "daytime" versus "nighttime" categories varied as well. Across participants, a total of 83 reported sleep episodes (ranging from 2 to 9 per participant) met the criteria for daytime sleep. For this analysis, daytime episodes were defined as those with TIB starting between 06:00 and 20:00 and ending before midnight of the same calendar day. Additional information on individual sleep schedules and qualifying daytime sleep episodes for analysis are depicted in Supplementary Figures S1 and S2.

Devices each collected sleep data passively and automatically, which is the default by which most users track their sleep with wearable devices. Our goal was to evaluate the devices in conditions that closely mirrored the daily use and real-world settings of typical users. To this end, we used only the sleep-tracking data that were originally provided by the device apps and data export portals regarding whether a device recorded a sleep episode and the timing/duration of that episode (ie, without additional editing to the device app inputs/outputs that some devices allow for users or researchers, to potentially correct TIB errors or log missed sleep episodes). For additional context, Supplemental Figure S3 depicts data from three example participants showing all their device TIBs alongside their sleep log TIBs across the study week.

Statistical Analysis

The two aims of this study were to evaluate daytime sleep episodes tracked from wearable devices for (1) bias in sleep timing outcomes (TIB, start time of sleep episode, end time of sleep episode) and (2) the frequencies of missing sleep episode recordings and false-positive episodes mislabeled as sleep. For Aim 1, we adopted the analysis standards (with minor modifications) recommended for evaluating summary bias for sleep devices as outlined by Menghini et al,[23] which we had also applied in our previous device performance evaluation studies.[15,16] Since the focus of Aim 1 was on evaluating sleep timing outcomes (and not the epoch-by-epoch classifications for sleep-wake or sleep stages that we

reported for these devices previously[16]), statistical output for this study included the means and bias summary tables and Bland-Altman plots[24] to depict performance against the sleep episode times reported in the reference sleep log. Additionally, repeated measures correlations[25] between the differences in the times reported for physically entering and exiting bed and the biases for device sleep start and end times were also calculated to evaluate whether observed device sleep timing biases were potentially related to the reported amount of time spent physically in bed outside of the TIB window for sleep (ie, as a measure of potential sedentary waking activity in bed such as reading, smartphone use, or watching television). For Aim 2, bar charts were constructed to depict the frequency of devices detecting versus missing daytime sleep episode recordings in aggregate and across hourly TIB bins. Performance for Aim 2 was also evaluated against daytime sleep episodes reported in the reference sleep log.

For two participants, each had one daytime sleep episode recorded by Polar Vantage V Titan that was an extreme outlier on sleep timing outcomes (ie, sleep recordings continued for 291 and 758 min, respectively, after the end of those sleep episodes, despite reporting removal of the devices soon after those sleep episodes ended). Therefore, the Polar device data for those two episodes were included only in the frequency of missing sleep recordings analysis for Aim 2, but were excluded from the sleep timing bias analyses for Aim 1. No other data recorded from the devices were excluded from any analyses.

All analyses were conducted using the statistical computing language R, version 4.1.2 (R Foundation, Vienna, Austria).

Results

Sleep episode timing summary results are shown in Table 1 and corresponding Bland-Altman plots in Figures 1–3. Expanded sleep-wake summary results are presented in Supplementary Table S1.

Daytime Sleep Episode TIB Bias

For TIB (Table 1 and Figure 1), the devices exhibited low bias for mean biases and most individual episodes. However, each device also had a few episodes that were more variable, such as those with biases >1 h under or over the TIB reported in the sleep log. Corresponding to the variability observed in biases, which included both under- and over-estimations for the devices compared with the sleep log, mean biases were greater when expressed in absolute terms but

Table 1 Bland-Altman Summary Agreement

Outcome	Device	n	Device Mean (SD)	Reference Mean (SD)	Mean Bias (SD)	Mean Absolute Bias (SD)
TIB	Fatigue Science ReadiBand	80	328.1 (123.7)	344.0 (125.8)	−15.8 (28.8)	23.5 (22.9)
	Fitbit Inspire HR	79	351.4 (113.8)	353.4 (115.6)	−1.9 (45.0)	26.3 (36.4)
	Oura Ring	78	362.0 (108.8)	350.6 (115.6)	11.4 (32.6)	22.5 (26.1)
	Polar Vantage V Titan	50	394.0 (81.2)	405.9 (73.7)	−11.8 (27.0)	21.3 (20.2)
Start Time	Fatigue Science ReadiBand	80	641.8 (184.9)	634.0 (186.7)	7.8 (20.0)	15.1 (15.2)
	Fitbit Inspire HR	79	625.2 (173.7)	620.0 (173.9)	5.1 (27.6)	15.3 (23.5)
	Oura Ring	78	626.9 (171.9)	625.8 (177.1)	1.2 (18.9)	11.1 (15.3)
	Polar Vantage V Titan	50	601.8 (153.0)	590.7 (147.3)	11.0 (17.1)	13.9 (14.8)
End Time	Fatigue Science ReadiBand	80	970.0 (156.4)	978.0 (151.4)	−8.0 (20.4)	14.0 (16.8)
	Fitbit Inspire HR	79	976.6 (152.9)	973.4 (151.7)	3.2 (35.9)	17.5 (31.4)
	Oura Ring	78	989.0 (152.9)	976.4 (153.4)	12.6 (26.5)	17.0 (23.9)
	Polar Vantage V Titan	50	995.8 (143.9)	996.6 (142.5)	−0.8 (22.0)	13.3 (17.5)

Notes: Summary agreement between devices and reference for the Bland-Altman plots for TIB, sleep episode start time, and sleep episode end time. The sleep log served as reference for all outcome measures. All daytime sleep episodes with available data for both the device and reference were included in analyses, with the total number of nights (n) indicated in each row. Mean and SD are shown for each device and reference, with all values provided in minutes. Values for sleep start and end times are minutes since midnight. Mean bias and mean absolute bias columns compare the device and reference in linear bias terms (not regression bias terms, which are provided within the expanded bias statistics in Supplementary Table S1). Positive and negative values in the mean bias column indicate overestimations or underestimations, respectively, for the device versus the reference sleep log.
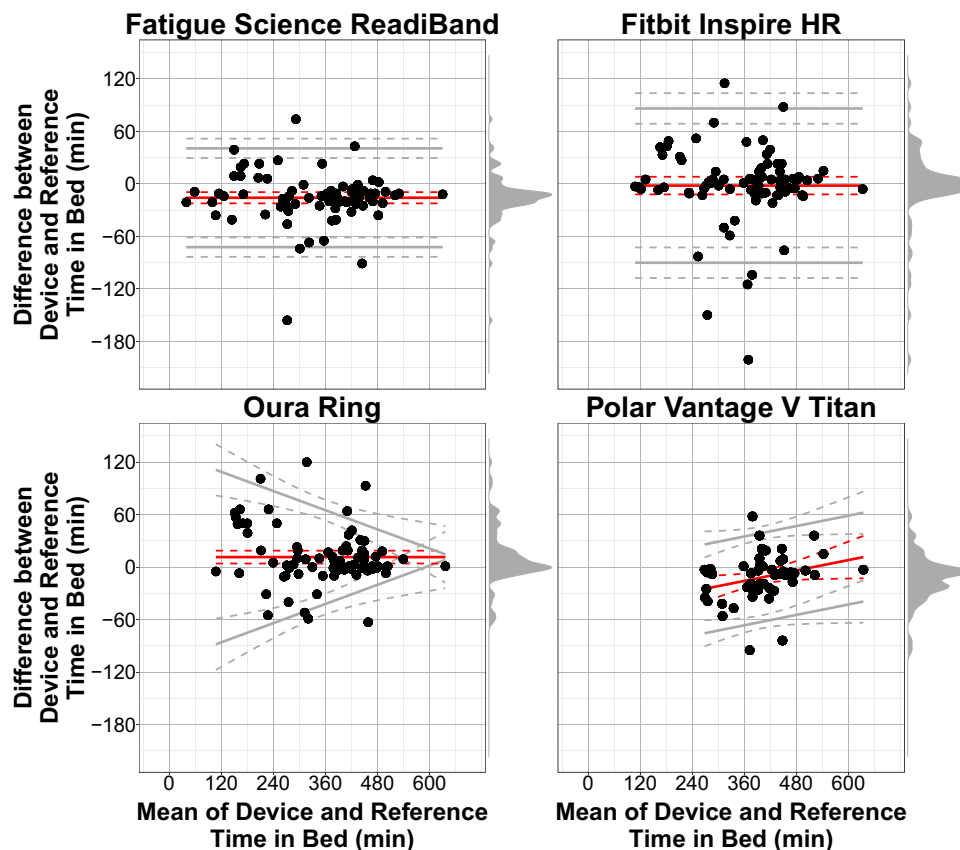**Abbreviations**: SD, standard deviation; TIB, time in bed.

Figure 1 Bland-Altman plots: TIB.
Notes: Plots depict the mean bias (solid red line) and upper and lower limits of agreement (LOA; solid gray lines) for deviation in TIB for the devices compared with the reference sleep log. Black circles are individual nights. Dashed lines represent the 95% confidence intervals around the bias and LOA lines. Gray shaded regions on the right y-axis are density plots showing the distribution of individual night biases. Zero on the y-axis represents no difference, with positive and negative y-axis values indicating an overestimation or underestimation, respectively, compared with the reference. Diagonal mean bias lines indicate significant proportional bias. Non-parallel LOA lines indicate significant heteroscedasticity.
Abbreviations: LOA, limits of agreement; TIB, time in bed.

were all <30 min. The only device displaying a significant proportional mean bias for TIB was the Polar Vantage V Titan, which was in a positive direction.

Daytime Sleep Episode Start Time Bias

The start time of sleep episodes for each device (Table 1 and Figure 2) also had low biases for the mean and most individual episodes. Like TIB, a few episodes for each device also exhibited large underestimations or overestimations compared with the sleep log. Significant proportional mean biases occurred for the Oura Ring (negative) and the Polar Vantage V Titan (positive).

Daytime Sleep Episode End Time Bias

For sleep episode end times (Table 1 and Figure 3), device biases remained low for most episodes and the mean biases. Like TIB and episode start time biases, episode end times also had a few episodes with large biases. Only the Fatigue Science ReadiBand exhibited a significant proportional mean bias, which was in a positive direction.

Correlations Between Daytime Sleep Episode Start and End Time Biases and Physical in-Bed Times

Table 2 displays correlations that tested whether the biases of start and end times of daytime sleep episodes were associated with the differences in the times that participants reported physically entering or exiting bed (which sometimes
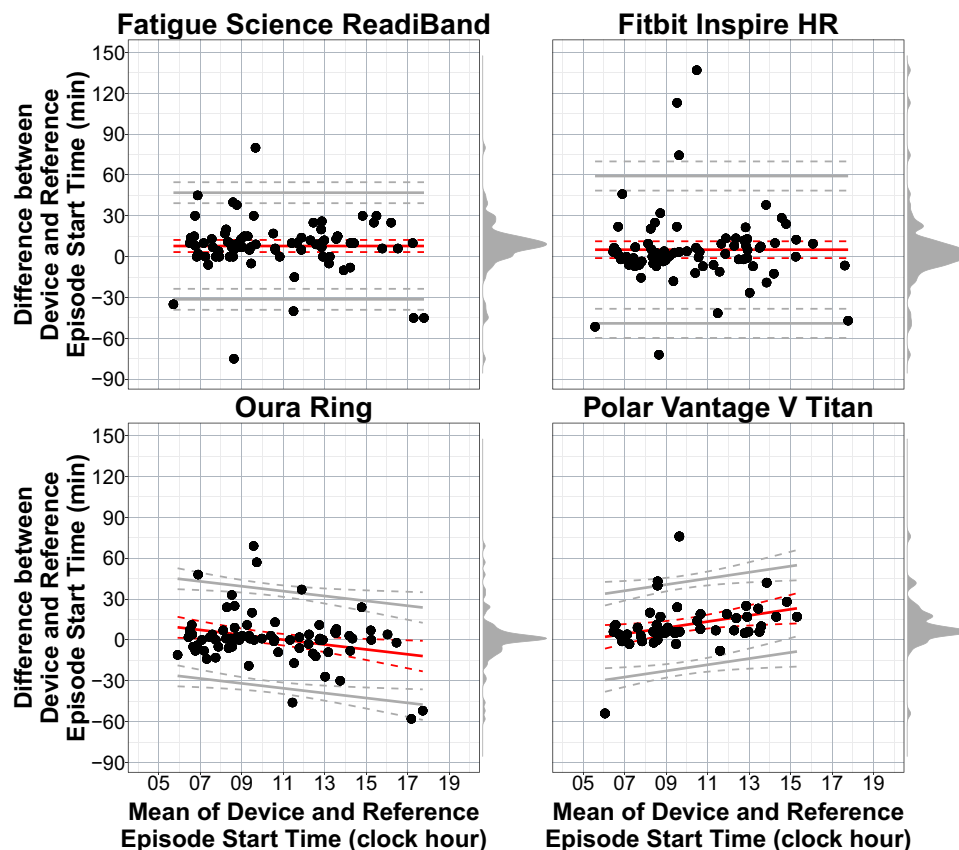
Figure 2 Bland-Altman plots: sleep episode start time.
Notes: Plots depict the mean bias (solid red line) and upper and lower limits of agreement (solid gray lines) for deviation in sleep episode start time of TIB for the devices compared with the reference sleep log. See Figure 1 notes for additional figure details.
Abbreviation: TIB, time in bed.

differ from the start and end times of the sleep episode TIB, such as when someone spends time in sedentary wake activities in bed prior to or after their attempted sleep episode). Across all devices, none of the correlations for the start time or end time of daytime sleep episodes reached statistical significance, suggesting that device start and end time biases for TIB were not associated with bed entry or exit times.

Frequency of Missing Daytime Sleep Episode Recordings

Histograms depicting the frequency of devices detecting or missing the reported daytime sleep episodes are shown in Figure 4. The Fatigue Science ReadiBand, Fitbit Inspire HR, and Oura Ring missed only a few episodes overall (3.6%, 4.8%, and 6.0%, respectively). For Fitbit and Oura, these missing episodes were mostly naps with shorter TIB. Notably, the Polar Vantage V Titan failed to detect the most episodes overall (37.3%), including all episodes with reported TIB <4 h.

Additionally, devices sometimes produced false-positive daytime sleep recordings (ie, detecting additional daytime episodes when they were not reported in the sleep logs; see Supplemental Figure S4). This type of error occurred less frequently than the error frequency for devices missing reported episodes, as described previously. Fatigue Science ReadiBand was the only device with a cluster of false-positive episodes, occurring most frequently for low TIBs (eg, extra naps recorded). The overall percentage of false-positive episodes mislabeled as sleep versus the total number of daytime episodes detected by each device were as follows: Fatigue Science ReadiBand (10.0%), Fitbit Inspire HR (4.9%), Oura Ring (3.7%), Polar Vantage V Titan (1.9%).
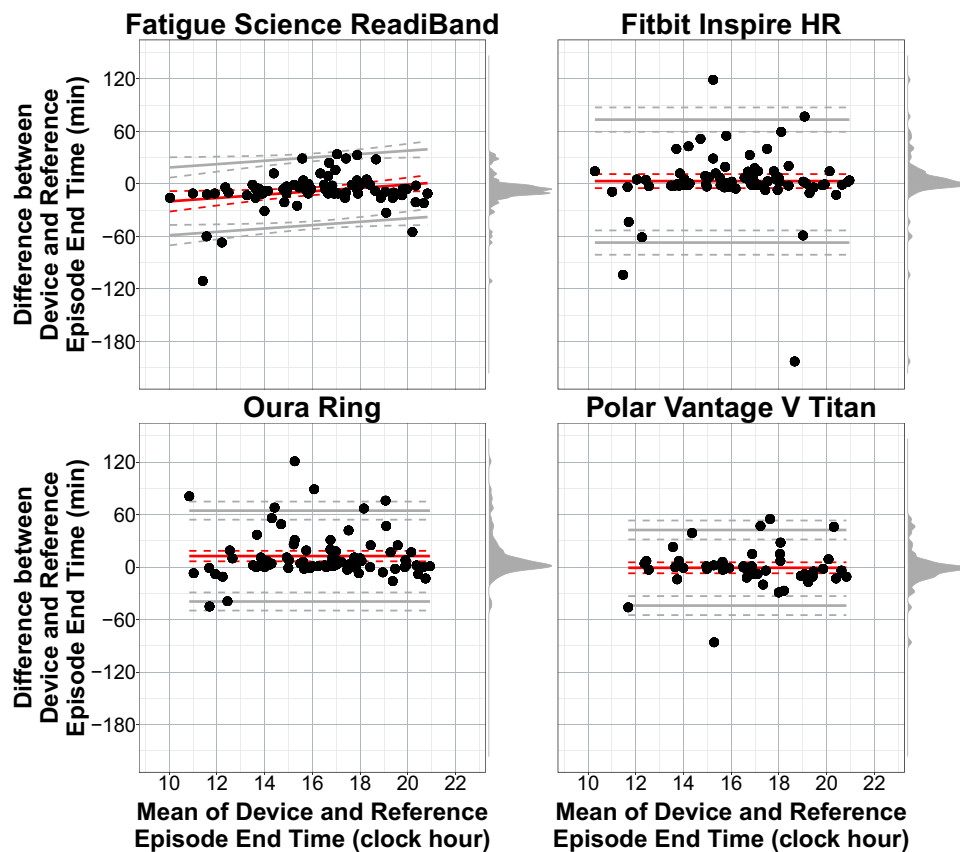
Figure 3 Bland-Altman plots: sleep episode end time.
Notes: Plots depict the mean bias (solid red line) and upper and lower limits of agreement (solid gray lines) for deviation in sleep episode end time of TIB for the devices compared with the reference sleep log. See Figure 1 notes for additional figure details.
Abbreviation: TIB, time in bed.

Discussion

During real-world conditions, we found that four commercial wearable devices exhibited low bias in general when measuring daytime sleep episode timing, but some devices had mixed or poor performance for detecting all the reported daytime sleep episodes (especially when TIB was short). Devices sometimes had large biases (eg, >1 h) for estimating the sleep timing of daytime episodes, but there were few instances overall. The direction of the bias was variable (ie, both

Table 2 Correlations Between Daytime Sleep Episode Start and End Time Biases and Physical in-Bed Times

Bias Outcome	Device	Correlation Statistic	p-value
Start Time	Fatigue Science ReadiBand	0.040	0.754
	Fitbit Inspire HR	0.104	0.412
	Oura Ring	0.078	0.545
	Polar Vantage V Titan	−0.038	0.819
End Time	Fatigue Science ReadiBand	−0.167	0.184
	Fitbit Inspire HR	−0.223	0.076
	Oura Ring	−0.131	0.304
	Polar Vantage V Titan	−0.217	0.184

Notes: Correlations testing associations between sleep episode TIB start and end time biases and the differences in times when participants physically entered and exited bed, respectively. Participants reported all times in the sleep logs. Statistics are repeated measures correlations.
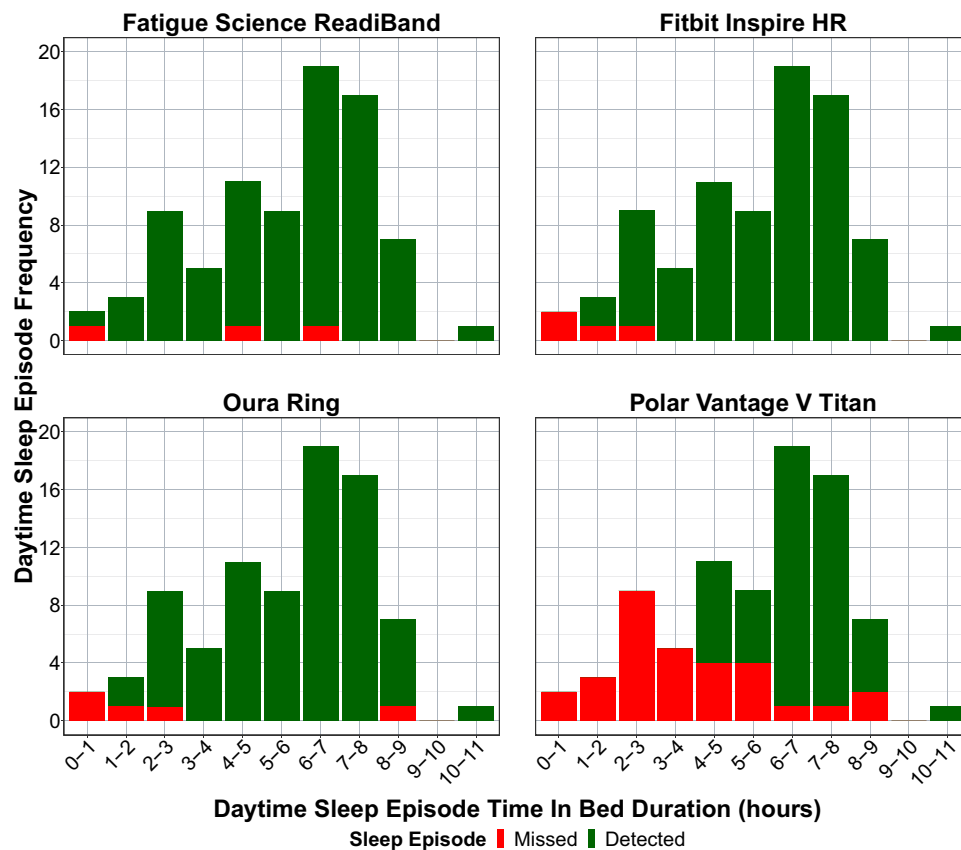**Abbreviation**: TIB, time in bed.

Figure 4 Frequency of missing daytime sleep episode recordings.
Notes: Plots depict stacked histogram bars for the number of detected and missing daytime sleep episode recordings for the devices compared with those reported in the reference sleep log. Green bars represent correctly detected episodes and red bars represent missing episodes, for each 1-h TIB bin. TIB refers to the minutes of TIB reported for each daytime sleep episode in the sleep log, which may not necessarily equal the TIB of the corresponding daytime sleep episode recording by the device.
Abbreviation: TIB, time in bed.

underestimations and overestimations occurred), which may have contributed to the low mean biases and few proportional biases observed for the devices across sleep timing outcomes. Daytime sleep timing biases were not correlated with the time that participants reported spending physically in bed before and after their sleep episode TIB, indicating that TIB biases may sometimes be due to potential behavioral and physiological factors other than engaging in sedentary wake activities in bed that occur before or after their attempted sleep episode. Differences observed between devices in their ability to correctly detect daytime sleep episodes, which also differed by TIB, suggest that different device sleep-tracking algorithms likely have varying sensitivities or thresholds for triggering the recording of daytime sleep and naps in general. Polar Vantage V Titan was especially poor at detecting shorter daytime episodes, missing over a third of total episodes including all episodes with TIB <4 h; however, this finding aligns with the parameters of its sleep-tracking algorithm to not detect naps of that duration. This performance gap between devices has implications for their reliability and accuracy when used with shift workers or others who often sleep or nap during daytime hours. Overall, wearable devices exhibit promising performance for tracking the timing of daytime sleep episodes, but the mixed performance for daytime sleep episode detection (especially for shorter naps) raises some concerns for the reliability and accuracy of certain devices for daytime sleep tracking.

The finding that device-determined TIB is tracked with low bias for most episodes largely indicates promising daytime-sleep tracking performance for the accuracy of devices to measure sleep timing when an episode is recorded. Accurate detection of TIB is a critical performance outcome because measurement of the other major sleep summary outcomes of interest (eg, TST, sleep efficiency, wake after sleep onset, sleep onset latency) rely on first identifying the correct TIB window. Therefore, the few episodes we found with large TIB biases will also bias other sleep summary

outcomes, which has implications for the reliability of sleep-tracking data in general and is a concern for their real-world use (especially for individuals who often sleep during the day).[21] For all devices, biases occurred both as overestimations and underestimations. Therefore, the overall mean biases were often low because there were similar magnitudes of positive and negative individual bias values (averaging out close to zero). Thus, it is also important to consider measures of distribution for biases that contextualize the (sometimes) large variability in observed biases (eg, mean absolute bias, SD, and LOA). The importance of accurate TIB tracking was also considered in our previous study examining performance of these same devices for nighttime sleep, in which we also found that most episodes exhibited low TIB bias with some notable exceptions (eg, bias >1 h).[16] To enhance the use of devices for personal, research, or clinical purposes, sleep-tracking devices and algorithms should prioritize improving the accuracy of TIB tracking.

In the current study, we also examined potential biases at the start or end time of daytime sleep episodes. As with TIB, mean biases were low for all devices in their start or end time detection, only a few episodes exhibited large biases, and there was no evidence of any relationship with the extra time that participants spent in bed outside of the TIB window. Thus, the cause of bias may be in other potential factors (eg, demographics and behavior) that should be examined in future studies to better understand why a device may be biased to start or end at the wrong time, which would improve sleep tracking in general.

The main area of daytime sleep-tracking performance where we found differences between devices was in the frequency of detecting or missing daytime episode recordings. This was especially the case for the Polar Vantage V Titan, which missed the greatest number of episodes overall (37.3%) and failed to detect any daytime episodes with TIB <4 h, a finding that has obvious implications for using Polar devices in daytime sleepers. However, this finding is not surprising given the information Polar provides regarding its sleep-tracking algorithm,[26] which states that only the longest sleep episode per day (defined as 18:00 to 18:00) is tracked and that episodes <4 h are not detected at all. These specific algorithm design choices currently make Polar devices largely unreliable sleep-tracking tools for those who nap or split their sleep episodes across multiple bouts per day, such as operational or clinical populations described previously. Oura Ring also missed a few of the shorter daytime episodes, which may also be concerning for its use in such populations. According to Oura, its sleep algorithm can detect naps for episodes with TST between 15 min and 3 h.[27] It should also be noted that some newer features for the Oura Ring that are available to researchers on the data management platform Oura Teams may provide additional sleep and nap data beyond the data available to users on the Oura app (eg, extra "rest" periods), which is where some of the nap data were found for our analyses. Oura requires users to "confirm" naps on the same day they occur, with a prompt appearing on the Oura app (see Methods for more details). We could not verify which naps were "confirmed" by participants or not, therefore each participant's Oura nap data were constructed from both the app and Oura Teams exports. For the shortest nap episodes, if the TST was under 15 min it would not reach Oura's defined threshold for a nap and may have led to the nap not appearing in the sleep data outputs – thus, possibly accounting for some of the missing short daytime naps that we found for Oura. The Fitbit Inspire HR and Fatigue Science ReadiBand had the most reliable recording performances, each missing only a few daytime episodes. Fitbit states that naps >1 h are tracked,[28] which aligns well with our finding that the two naps logged that were <1 h were missed, whereas almost all longer episodes were correctly detected. Although ReadiBand had a high rate for correctly detecting daytime sleep episodes regardless of TIB, ReadiBand notably had the most false-positive episodes recorded, especially short naps, detecting several that were not reported by participants. ReadiBand has been shown to track daytime sleep episodes in recent real-world observational studies in which it was used to measure the sleep patterns of shift-working nurses and police officers,[29,30] although without reported reference metrics to examine potential performance biases (eg, TIB biases or missed and false-positive sleep episodes) for comparison to the current study. In the current study, differences between devices in their ability to detect episodes are potentially determined by designed settings in their proprietary sleep-tracking algorithms (eg, thresholds for activity, heart rate, TIB, or TST) and could be changed or updated in the future to better detect daytime sleep and naps. Thus, additional studies on daytime sleep tracking are needed to examine performance parameters and limitations or improvements, while new devices and algorithms continue to be released on the consumer market.

Previous studies have not specifically focused on the performance evaluation of naps or daytime sleep episodes with commercial wearable devices,[1,6,19] therefore the findings of this study are novel. However, there is evidence from a few

previous studies that wearable devices would not perform well for tracking naps or daytime sleep. A recent study found that the Garmin[TM] Vivosmart HR device (Garmin, Olathe, KS, USA) was not able to detect daytime sleep episodes in police trainees when they switched from working day shifts (sleeping at night) to working night shifts (primarily sleeping during the daytime).[31] Additionally, a study with a previous-generation Fitbit device (Charge HR model) found that in a group of young athletes most of the daytime sleep episodes were missed, especially shorter naps with TIB of 1 or 2 h.[32] Finally, a study using a since-discontinued Jawbone device found that it could not reliably record sleep during daytime nap episodes in a multiple sleep latency test protocol.[33] In contrast, a recent study of the WHOOP[TM] strap device (WHOOP, Boston, MA, USA) under a controlled laboratory setting compared with PSG tested daytime sleep episodes, with a 7-h fixed TIB occurring in the afternoon-evening with an analysis of its auto-detection sleep-tracking feature.[34] While the authors found the device generally had good sleep-tracking performance that was comparable to when TIB was entered manually for the same episodes, they also reported that no sleep episodes were missed with its auto-detection feature. The study, however, did not include naps or shorter daytime episodes, nor was TIB a reported outcome, and therefore those results cannot be compared directly with the bias outcomes of the current study.

With the growing popularity of wearable devices for tracking sleep, physiology, and behavior, the relevance for understanding how devices perform across the 24-h day cannot be overstated, especially for groups at risk of safety, performance, and health decrements from irregular or insufficient sleep patterns (eg, shift workers, military personnel, patients with sleep or circadian disorders). Unsurprisingly, several review papers and editorials have highlighted the need for more studies examining naps or daytime sleep tracking with wearable devices.[1,6,19] There is also a growing demand from sleep researchers who have expressed a preference for using wearable devices that feature nap detection for TIB as low as 20 min.[35] In the current study, only a few naps with TIB <1 h were observed for analysis, however we found that the missed episodes were mostly from episodes with low TIB including many naps. Therefore, from the current study, it is largely unknown whether commercial wearable devices can reliably track naps with TIB as low as the researcher-preferred minimum of 20 min, although it is theoretically possible that algorithms can be made sensitive enough for devices to better detect very short naps. It is likely that studies or initiatives to use wearables to track sleep patterns or promote sleep health in groups such as shift workers will suffer from low compliance or adoption if devices do not first demonstrate good performance for detecting naps or sleep episodes across the 24-h day.[36] Research-grade actigraphy traditionally has been used for real-world tracking of sleep across the day and night in clinical and operational populations, but newer multi-sensor commercial wearable devices with their proprietary algorithms (or customized research algorithms for sleep or circadian rhythm tracking using data from wearables[31,37–39]) theoretically could surpass standard actigraphy algorithms for daytime sleep tracking. This enhanced performance is necessary since actigraphy, which typically contains an accelerometry sensor input only, is often overly sensitive to classifying naps during times of sedentary activity.[40] Also, to improve accuracy of daytime sleep assessments with actigraphy, it is still recommended for participants to complete a reference sleep log every day to inform researchers when sleep was actually attempted to reduce the occurrence of potential false-positive naps.[18,40] Modern commercial wearable devices have the potential to track important real-world health statistics and clinical outcomes beyond sleep.[41,42] Based on the promise of what wearables could deliver for everyday use and clinical endpoints,[41,42] there is interest in the use and improvement of wearables from several different sectors (eg, research, industry, healthcare, and government) and an expansion of the already growing wearables market. These advancements will benefit the sleep field too,[2] if wearables can be improved to reliably track the sleep of those with irregular sleep patterns such as shift workers and operational communities who make up a large and necessary part of the workforce and have known risks to their health, performance, and safety.[43–45]

This study has several strengths. Multiple devices were tested concurrently, which resulted in higher throughput of results and easier comparison of performance between devices. This study was a direct follow-up to our previous study[16] that tested the same set of four commercial wearable devices during nighttime sleep in a similar unrestricted home protocol, which now extends and better defines the performance of these devices. This was one of the first studies to focus on evaluating the daytime sleep-tracking performance of commercial devices in general. Device performance was assessed under real-world, unrestricted conditions, which yielded a wide range of episode TIBs. Finally, as in our previous performance evaluation studies,[15,16] this study was also conducted independently without involvement or potential conflicts from the device companies.

This study also had limitations. The study design of self-selected and unrestricted sleep schedules resulted in sleep episodes with varying TIBs to test. However, the TIB bins were not balanced, and there were fewer episodes with shorter TIBs (naps) to evaluate than the longer main sleep episodes. Though we included many real-world shift workers, the study enrolled only young healthy adults and may not be representative of shift worker or operational populations who often have health issues and sleep disorders[43–45] that could affect wearable device performance. This was an initial study of the largely unknown performance of commercial devices to track daytime sleep; therefore, future studies into this area would benefit from including larger and more diverse samples of participants and by tracking sleep over periods longer than 1 week. The reference used for sleep schedule data was a digital self-report sleep log that enabled real-time compliance checks, but errors in reporting or recall may have been possible. We did not examine potential demographic differences in device performance. It is possible that wearable devices with PPG sensors, like some of those in the current study, may be prone to inaccuracies in heart rate measurements due to race, ethnicity, or skin tone,[46] which we plan to examine in a forthcoming analysis. Finally, we chose to test a variety of wearable devices from established companies that are widely used by the public or by operational communities. However, the consumer market has a high turnover of devices, so it is possible that the specific device models or algorithms we tested in the current study may be updated or no longer available in the future.

Conclusion

Our findings indicate that when daytime sleep episodes were recorded by the four wearable devices, TIB outcomes were largely tracked with low bias. However, performance differences between devices were revealed in their ability to reliably record daytime sleep episodes under certain conditions. In particular, naps and shorter sleep episodes with TIB <4 h were missed altogether by the Polar Vantage V Titan, whereas the Fatigue Science ReadiBand, Fitbit Inspire HR, and Oura Ring missed few daytime episodes. However, ReadiBand was over-sensitive in detecting sleep and recorded several extra false-positive daytime sleep episodes that were not actual logged sleep episodes, especially naps with low TIB. These performance differences in detecting daytime sleep and naps are likely caused by differences between algorithm settings, and device performance in this area may improve if the algorithm's sensitivity and use of thresholds (eg, for minimum duration of naps) are changed in the future. While we and others have found that many wearable devices (including this same set of four devices[16]) largely perform well for tracking epoch-by-epoch and summary outcomes for sleep and wake, previous device evaluation study protocols were limited to testing sleep at night only and usually under fixed TIB. Thus, the assumption that devices may also perform well for tracking daytime sleep has been overlooked by initial studies, and this study represents one of the first to test such assumptions rigorously by examining daytime sleep-tracking performance specifically. These findings have implications for the use of devices in people who nap or sleep during the daytime and/or have irregular sleep schedules in general, such as shift workers (eg, hospital workers, first responders, and operational communities like the military) or individuals with sleep or circadian disorders. The use of unrestricted sleep schedules yielded naturalistic variations in TIB, including shorter naps where the majority of missed recordings for some devices were revealed, thereby providing greater translation of results to real-world device use. Additional studies evaluating these and other wearable devices under similar and different testing conditions and populations are needed to confirm and extend current findings and better refine the strengths, limitations, and recommendations for wearable devices across the 24-h day.

Abbreviations

LOA, limits of agreement; PPG, photoplethysmography; PSG, polysomnography; SD, standard deviation; TIB, time in bed.

Data Sharing Statement

The datasets generated and/or analyzed during the current study are not publicly available due to security protocols and privacy regulations, but they may be made available on reasonable request by the Naval Health Research Center Institutional Review Board (contact phone +16195538400).

Acknowledgments

Disclosure

The authors report no conflicts of interest in this work.

References

1. Goldstein C. Current and future roles of consumer sleep technologies in sleep medicine. *Sleep Med Clin*. 2020;15(3):391–408. doi:10.1016/j.jsmc.2020.05.001
2. Baumert M, Cowie MR, Redline S, et al. Sleep characterization with smart wearable devices: a call for standardization and consensus recommendations. *Sleep*. 2022;45(12):zsac183. doi:10.1093/sleep/zsac183
3. de Zambotti M, Cellini N, Menghini L, Sarlo M, Baker FC. Sensors capabilities, performance, and use of consumer sleep technology. *Sleep Med Clin*. 2020;15(1):1–30. doi:10.1016/j.jsmc.2019.11.003
4. Lujan MR, Perez-Pozuelo I, Grandner MA. Past, present, and future of multisensory wearable technology to monitor sleep and circadian rhythms. *Front Digit Health*. 2021;3:721919. doi:10.3389/fdgth.2021.721919
5. Khosla S, Deak MC, Gault D, et al. Consumer sleep technology: an American Academy of Sleep Medicine position statement. *J Clin Sleep Med*. 2018;14(5):877–880. doi:10.5664/jcsm.7128
6. Depner CM, Cheng PC, Devine JK, et al. Wearable technologies for developing sleep and circadian biomarkers: a summary of workshop discussions. *Sleep*. 2020;43(2):zsz254. doi:10.1093/sleep/zsz254
7. Schutte-Rodin S, Deak MC, Khosla S, et al. Evaluating consumer and clinical sleep technologies: an American Academy of Sleep Medicine update. *J Clin Sleep Med*. 2021;17(11):2275–2282. doi:10.5664/jcsm.9580
8. de Zambotti M, Menghini L, Grandner MA, et al. Rigorous performance evaluation (previously, "validation") for informed use of new technologies for sleep health measurement. *Sleep Health*. 2022;8(3):263–269. doi:10.1016/j.sleh.2022.02.006
9. Grandner MA, Rosenberger ME. Chapter 12 - actigraphic sleep tracking and wearables: historical context, scientific applications and guidelines, limitations, and considerations for commercial sleep devices. In: Grandner MA, editor. *Sleep and Health*. Academic Press; 2019:147–157. doi:10.1016/B978-0-12-815373-4.00012-5
10. Montgomery-Downs HE, Insana SP, Bond JA. Movement toward a novel activity monitoring device. *Sleep Breath*. 2012;16(3):913–917. doi:10.1007/s11325-011-0585-y
11. Meltzer LJ, Hiruma LS, Avis K, Montgomery-Downs H, Valentin J. Comparison of a commercial accelerometer with polysomnography and actigraphy in children and adolescents. *Sleep*. 2015;38(8):1323–1330. doi:10.5665/sleep.4918
12. Haghayegh S, Khoshnevis S, Smolensky MH, Diller KR, Castriotta RJ. Accuracy of Wristband Fitbit models in assessing sleep: systematic review and meta-analysis. *J Med Internet Res*. 2019;21(11):e16273. doi:10.2196/16273
13. Lee XK, Chee NIYN, Ong JL, 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. doi:10.5664/jcsm.7932
14. Kahawage P, Jumabhoy R, Hamill K, de Zambotti M, Drummond SPA. 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. doi:10.1111/jsr.12931
15. Chinoy ED, Cuellar JA, Huwa KE, et al. Performance of seven consumer sleep-tracking devices compared with polysomnography. *Sleep*. 2021;44(5):zsaa291. doi:10.1093/sleep/zsaa291
16. Chinoy ED, Cuellar JA, Jameson JT, Markwald RR. Performance of four commercial wearable sleep-tracking devices tested under unrestricted conditions at home in healthy young adults. *Nat Sci Sleep*. 2022;14:493–516. doi:10.2147/NSS.S348795
17. Grandner MA, Bromberg Z, Hadley A, et al. Performance of a multisensor smart ring to evaluate sleep: in-lab and home-based evaluation of generalized and personalized algorithms. *Sleep*. 2023;46(1):zsac152. doi:10.1093/sleep/zsac152
18. Ancoli-Israel S, Cole R, Alessi C, Chambers M, Moorcroft W, Pollak CP. The role of actigraphy in the study of sleep and circadian rhythms. *Sleep*. 2003;26(3):342–392. doi:10.1093/sleep/26.3.342
19. Lambrechtse P, Ziesenitz VC, Cohen A, van den Anker JN, Bos EJ. How reliable are commercially available trackers in detecting daytime sleep [letter to the editor]. *Br J Clin Pharmacol*. 2018;84(3):605–606. doi:10.1111/bcp.13475

20. Watson NF, Badr MS, Belenky G, et al. Recommended amount of sleep for a healthy adult: a joint consensus statement of the American Academy of Sleep Medicine and Sleep Research Society. *Sleep*. 2015;38(6):843–844. doi:10.5665/sleep.4716

21. Chinoy ED, Markwald RR. Use of technology for real-world sleep and circadian research. In: *Encyclopedia of Sleep and Circadian Rhythms*. 2nd ed. Elsevier; 2023:156–168. doi:10.1016/B978-0-12-822963-7.00200-0

22. Carney CE, Buysse DJ, Ancoli-Israel S, et al. The consensus sleep diary: standardizing prospective sleep self-monitoring. *Sleep*. 2012;35(2):287–302. doi:10.5665/sleep.1642

23. Menghini L, Cellini N, Goldstone A, Baker FC, de Zambotti M. A standardized framework for testing the performance of sleep-tracking technology: step-by-step guidelines and open-source code. *Sleep*. 2021;44(2):zsaa170. doi:10.1093/sleep/zsaa170

24. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;1(8476):307–310. doi:10.1016/S0140-6736(86)90837-8

25. Bakdash JZ, Marusich LR. Repeated measures correlation. *Front Psychol*. 2017;8:456. doi:10.3389/fpsyg.2017.00456

26. Polar Sleep Plus Stages™ FAQS. Available from: https://support.polar.com/en/sleep-plus-stages-faqs. Accessed February 2, 2023.

27. Oura Nap Detection. Available from: https://support.ouraring.com/hc/en-us/articles/1500009653181-Nap-Detection. Accessed February 2, 2023.

28. How do I track my sleep with my Fitbit device? Available from: https://help.fitbit.com/articles/en_US/Help_article/1314.htm. Accessed February 2, 2023.

29. James L, James SM, Wilson M, et al. Sleep health and predicted cognitive effectiveness of nurses working 12-hour shifts: an observational study. *Int J Nurs Stud*. 2020;112:103667. doi:10.1016/j.ijnurstu.2020.103667

30. James L, Caruso CC, James S. Pilot test of NIOSH training for law enforcement on shift work and long work hours. *J Occup Environ Med*. 2022;64(7):599–606. doi:10.1097/JOM.0000000000002534

31. Erickson ML, Wang W, Counts J, et al. Field-based assessments of behavioral patterns during shiftwork in police academy trainees using wearable technology. *J Biol Rhythms*. 2022;37(3):260–271. doi:10.1177/07487304221087068

32. Sargent C, Lastella M, Romyn G, Versey N, Miller DJ, Roach GD. How well does a commercially available wearable device measure sleep in young athletes? *Chronobiol Int*. 2018;35(6):754–758. doi:10.1080/07420528.2018.1466800

33. Cook JD, Prairie ML, Plante DT. Ability of the multisensory jawbone UP3 to quantify and classify sleep in patients with suspected central disorders of hypersomnolence: a comparison against polysomnography and actigraphy. *J Clin Sleep Med*. 2018;14(5):841–848. doi:10.5664/jcsm.7120

34. Miller DJ, Roach GD, Lastella M, et al. A validation study of a commercial wearable device to automatically detect and estimate sleep. *Biosensors*. 2021;11(6):185. doi:10.3390/bios11060185

35. Devine JK, Schwartz LP, Choynowski J, Hursh SR. Expert demand for consumer sleep technology features and wearable devices: a case study. *IoT*. 2022;3(2):315–331. doi:10.3390/iot3020018

36. Devine JK, Schwartz LP, Hursh SR. Technical, regulatory, economic, and trust issues preventing successful integration of sensors into the mainstream consumer wearables market. *Sensors*. 2022;22(7):2731. doi:10.3390/s22072731

37. Walch O, Huang Y, Forger D, Goldstein C. Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device. *Sleep*. 2019;42(12):zsz180. doi:10.1093/sleep/zsz180

38. Cheng P, Walch O, Huang Y, et al. Predicting circadian misalignment with wearable technology: validation of wrist-worn actigraphy and photometry in night shift workers. *Sleep*. 2021;44(2):zsaa180. doi:10.1093/sleep/zsaa180

39. Huang Y, Mayer C, Cheng P, et al. Predicting circadian phase across populations: a comparison of mathematical models and wearable devices. *Sleep*. 2021;44(10):zsab126. doi:10.1093/sleep/zsab126

40. Kanady JC, Drummond SPA, Mednick SC. Actigraphic assessment of a polysomnographic-recorded nap: a validation study. *J Sleep Res*. 2011;20(1 Pt 2):214–222. doi:10.1111/j.1365-2869.2010.00858.x

41. Dunn J, Kidzinski L, Runge R, et al. Wearable sensors enable personalized predictions of clinical laboratory measurements. *Nat Med*. 2021;27(6):1105–1112. doi:10.1038/s41591-021-01339-0

42. Mason AE, Kasl P, Hartogensis W, et al. Metrics from wearable devices as candidate predictors of antibody response following vaccination against COVID-19: data from the second TemPredict study. *Vaccines*. 2022;10(2):264. doi:10.3390/vaccines10020264

43. Cheng P, Drake C. Shift work disorder. *Neurol Clin*. 2019;37(3):563–577. doi:10.1016/j.ncl.2019.03.003

44. Good CH, Brager AJ, Capaldi VF, Mysliwiec V. Sleep in the United States military. *Neuropsychopharmacology*. 2020;45(1):176–191. doi:10.1038/s41386-019-0431-7

45. Gurubhagavatula I, Barger LK, Barnes CM, et al. Guiding principles for determining work shift duration and addressing the effects of work shift duration on performance, safety, and health: guidance from the American Academy of Sleep Medicine and the Sleep Research Society. *Sleep*. 2021;44(11):zsab161. doi:10.1093/sleep/zsab161

46. Colvonen PJ, DeYoung PN, Bosompra NOA, Owens RL. Limiting racial disparities and bias for wearable devices in health science research. *Sleep*. 2020;43(10):zsaa159. doi:10.1093/sleep/zsaa159