Sleep Health: Journal of the National Sleep Foundation 10 (2024) 356–368

Contents lists available at ScienceDirect

Sleep Health: Journal of the National Sleep Foundation

journal homepage: www.sleephealthjournal.org

Performance of wearable sleep trackers during nocturnal sleep and periods of simulated real-world smartphone use

Adrian R. Willoughby, PhD, Hosein Aghayan Golkashani, MD, PhD, Shohreh Ghorbani, MSc, Kian F. Wong, BA, Nicholas I.Y.N. Chee, BSc, Ju Lynn Ong, PhD, Michael W.L. Chee, MBBS *

Sleep and Cognition Laboratory, Centre for Sleep and Cognition, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

ARTICLE INFO

Article history:
Received 18 January 2024
Received in revised form 16 February 2024
Accepted 27 February 2024

Keywords:
Consumer sleep tracker
Specificity
Motionless wake
Sleep hygiene
Bedtime routine
Actigraphy

ABSTRACT

Goal and aims: To test sleep/wake transition detection of consumer sleep trackers and research-grade actigraphy during nocturnal sleep and simulated peri-sleep behavior involving minimal movement.
Focus technology: Oura Ring Gen 3, Fitbit Sense, AXTRO Fit 3, Xiaomi Mi Band 7, and ActiGraph GT9X.
Reference technology: Polysomnography.
Sample: Sixty-three participants (36 female) aged 20-68.
Design: Participants engaged in common peri-sleep behavior (reading news articles, watching videos, and exchanging texts) on a smartphone before and after the sleep period. They were woken up during the night to complete a short questionnaire to simulate responding to an incoming message.
Core analytics: Detection and timing accuracy for the sleep onset times and wake times.
Additional analytics and exploratory analyses: Discrepancy analysis both including and excluding the peri-sleep activity periods. Epoch-by-epoch analysis of rate and extent of wake misclassification during peri-sleep activity periods.
Core outcomes: Oura and Fitbit were more accurate at detecting sleep/wake transitions than the actigraph and the lower-priced consumer sleep tracker devices. Detection accuracy was less reliable in participants with lower sleep efficiency.
Important additional outcomes: With inclusion of peri-sleep periods, specificity and Kappa improved significantly for Oura and Fitbit, but not ActiGraph. All devices misclassified motionless wake as sleep to some extent, but this was less prevalent for Oura and Fitbit.
Core conclusions: Performance of Oura and Fitbit is robust on nights with suboptimal bedtime routines or minor sleep disturbances. Reduced performance on nights with low sleep efficiency bolsters concerns that these devices are less accurate for fragmented or disturbed sleep.
© 2024 The Author(s). Published by Elsevier Inc. on behalf of National Sleep Foundation. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Introduction

Growing interest in personal sleep measurement to improve health and well-being has driven significant growth in the development and sales of consumer sleep trackers (CSTs). With such a proliferation of devices, it is vital to ensure that their performance is evaluated adequately and that they are used appropriately for scientific research.[1-3] As part of this effort, recent studies have found that the performance of higher-quality CSTs equals or exceeds that of "research-grade" actigraphy.[4-9]

However, a major concern for both traditional actigraphs and newer CSTs is their ability to correctly classify motionless wake.[1,10-12] Typically, these trackers have sensitivity for sleep detection of over 90%, meaning they correctly detect sleep epochs most of the time. However, their specificity, or correct detection of wakefulness when participants are in bed, is often under 60%.[5,13]

Most performance evaluation studies have been conducted on healthy participants in a sleep laboratory. However, good performance in laboratory studies does not necessarily translate into good performance outside the lab or when patients with insomnia use CSTs without supervision, as has been highlighted in the media.[14]

Peri-sleep behavior inside and outside the lab differs significantly. Research has shown that many people use electronic

* Corresponding author: Michael W. L. Chee, MBBS, Sleep and Cognition Laboratory, Centre for Sleep and Cognition, Yong Loo Lin School of Medicine, National University of Singapore, Tahir Foundation Bldg., MD1, 12 Science Drive 2, Singapore 117549, Singapore.
E-mail address: michael.chee@nus.edu.sg (M.W.L. Chee).

https://doi.org/10.1016/j.sleh.2024.02.007
2352-7218/© 2024 The Author(s). Published by Elsevier Inc. on behalf of National Sleep Foundation. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

devices while in bed before sleeping.[15-17] Device use in bed is associated with periods of motionless wake that wearable devices may be poor at classifying.

The correct classification of motionless wake is important not only in healthy people using electronic or other devices, but also in those who have difficulty initiating or maintaining sleep. Previous research has shown that performance of both "research-grade" actigraphy and CSTs declines as sleep quality decreases. This has been demonstrated both in a comparison between insomnia patients and healthy sleepers[18] and in the same participants under different conditions of sleep fragmentation.[19]

To address these issues, we designed a protocol to simulate common peri-sleep behaviors in the sleep lab, allowing both control and measurement of participant behavior, and accurate evaluation of device performance. Participants engaged in everyday activities (reading, watching videos, and texting) on a smartphone while in bed, before going to sleep and after waking up. They were also woken during the night to answer a brief questionnaire, to simulate being woken by, and replying to, an incoming message. We tested two well-known and well-developed CSTs (Oura and Fitbit), two lower-priced CSTs with unknown development histories (Axtro and Xiaomi) and one "research-grade" actigraph (ActiGraph GT9X).

This protocol allowed us to investigate the degree to which wake detection affected estimates of sleep measures and performance evaluation metrics. These findings should inform both researchers and CST users concerned about device performance on nights where there is a modest departure from an ideal bedtime routine.

Methods

Data for this study was collected on the second night of a two-night sleep protocol that evaluated performance of different categories of sleep wearable devices[9] (please see for additional methodological details and data from the first night). The protocol was approved by the Institutional Review Board of the National University of Singapore.

Sample

Participants were recruited from the National University of Singapore and the wider university community. Sixty-six participants were recruited from three different age groups to ensure comparable representation from a wider age range than commonly published on (Young: 18-30 years of age; Middle: 31-50; Older: 51-70). Data from three participants were excluded. One participant withdrew from the study after the first night. A second had such disturbed sleep (total sleep time [TST] = 2:42:30, sleep efficiency [SE] = 54.17%) that neither Oura nor Fitbit recorded any sleep. There was a data collection failure for the third participant. This resulted in a final sample of 63 participants.

Inclusion criteria were: (1) habitual sleep duration of at least 5 hours per night between 8 PM and 10 AM, (2) body mass index (BMI) less than 35 kg/m², (3) no self-reported pre-existing sleep, neurological, or psychiatric disorders, (4) no excessive daytime sleepiness (Epworth Sleepiness Scale score less than 11[20]), (5) low risk for Obstructive Sleep Apnea (indicated by the Berlin questionnaire[21]), (6) no caffeine or alcohol use within 6 hours prior to bedtime, (7) no active illness, and (8) not pregnant. Mean age, BMI, and sex distribution for the three age groups are shown in Table 1.

Table 1
Demographic details of the participants sorted by age group

	Overall	Age group			*p*-value
		Young	Middle	Older	
Age (y)	39.56 (15.51)	23.19 (2.23)[a]	37.65 (4.90)[b]	59.95 (6.17)[c]	< .001
Sex	27 M/36 F	6 M/15 F	12 M/11 F	9 M/10 F	.26
BMI (kg/m^2)	23.84 (3.33)	22.49 (2.44)[a]	24.86 (4.12)[b]	24.09 (2.70)[ab]	.054

Abbreviation: BMI, body mass index.
For age and BMI, means (and standard deviation) are reported. Statistical significance column shows ANOVA main effect (for age and BMI) and chi-squared (for sex) significance. Values with different letters are significantly different from each other (pairwise comparison: $p < .05$).

Design, study setting, and procedures

The study was conducted in a sleep research laboratory with controlled bedroom light and temperature and sound attenuation. To simulate common bedtime behaviors that could cause a wearable device to report sleep when a person was lying still but awake, participants completed a number of peri-sleep activities on a smartphone (Samsung A30s running Android 11) before, during, and after their night's sleep.

Directly before going to sleep, participants engaged in three Pre-Sleep activities: reading news articles, watching nature videos, and texting with a chatbot (https://www.kuki.ai). Each activity was designed to take approximately 10 minutes to complete. The order of activities was counterbalanced across participants. Full details of the peri-sleep activities can be found in Section 1.1 of the Supplementary Material.

To simulate responding to an incoming message during the night, participants were woken by an alarm and completed a short questionnaire (adapted from an existing one[22,23]) on the smartphone. They were woken up 5 minutes after the first appearance of N3 sleep, within 45 minutes of initial sleep onset. If N3 was not detected, the participant was woken at 45 minutes in N2 sleep. If they were in REM sleep after 45 minutes, they were woken during the next instance of N2 sleep. This Interruption split sleep into two main periods, Sleep Period 1 (SP1) and Sleep Period 2 (SP2).

If not already awake in the morning, participants were woken by an alarm at their requested wake-up time. They then completed two Post-Sleep activities designed to take around 15 minutes, which consisted of reading four news articles and completing a standardized questionnaire to evaluate their night's sleep and current mood.

An app that recorded phone usage (App Usage[24]) was used to record the timing and duration of phone use during the activities (app use start and end times were rounded to the nearest 30 seconds to align with polysomnography [PSG] epochs). Bedroom lights remained off for the entire time the participants were in bed (including the Pre- and Post-Sleep activity periods). Participants could sit or lie in any comfortable position in bed while they completed the peri-sleep activities. The timeline of events is shown in Fig. 1; sleep onset and wake times were determined from PSG, bedtimes were derived from phone use logs.

Focus technology

Five wearable devices were evaluated: the Oura Ring Gen3, running Oura Sleep Staging Algorithm 2.0 (OSSA 2.0; Oura Health Oy, Oulu, Finland), the Fitbit Sense (Fitbit Inc, San Francisco, CA), the ActiGraph GT9X (ActiGraph Inc, Pensacola, FL), the Xiaomi Mi Band 7 (Xiaomi Inc, Beijing, China) and the AXTRO Fit 3 Fitness Tracker (AXTRO Pte. Ltd, Singapore). For convenience, we refer to these devices by their manufacturer's name. The wearable devices and associated apps were updated to the latest version prior to the start of the study (see Supplementary Table S1 for firmware and software details).

Four of the five devices were worn concurrently (Oura, Fitbit, ActiGraph, and either Xiaomi or Axtro; this was done both for participant comfort and to keep the devices as close as possible to the

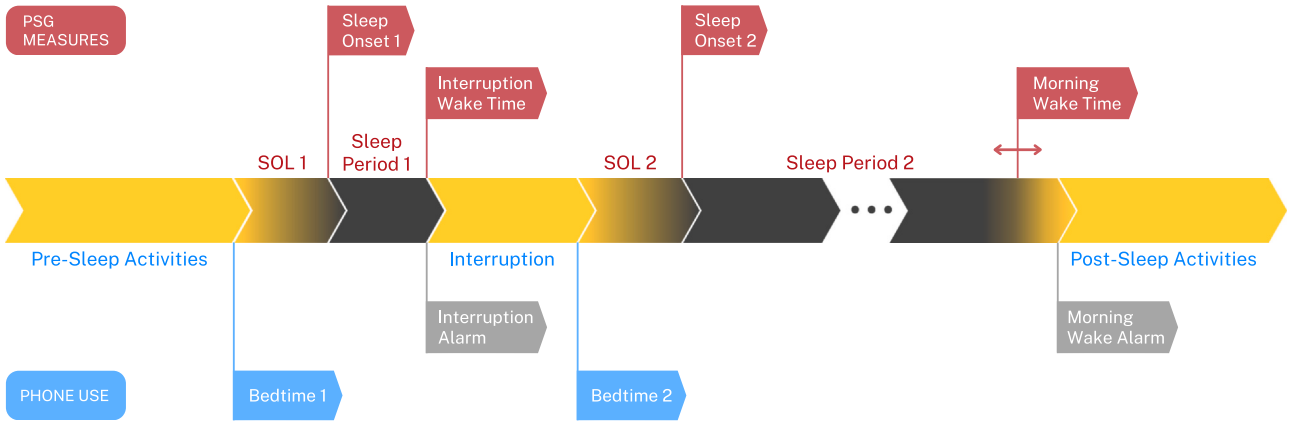

Fig. 1. Protocol timeline. PSG-derived timepoints are shown in red, phone use determined timepoints are shown in blue. Periods of wake are in yellow and periods of sleep are in black; the color gradient represents sleep/wake transitions. The horizontal arrow on the Morning Wake Time flag indicates variable wake time in relation to the morning alarm. SOL, Sleep Onset Latency

correct wrist position). The Oura ring was worn on either the index or ring finger of the nondominant hand; the other devices were worn on the nondominant wrist. Sleep-staged epochs for each device were resampled to match the 30-second epochs of the PSG recording (see Section 1.3 of the Supplementary Material).

To demarcate the period for sleep scoring, ActiGraph requires manual entry of markers. To compare the effects of marker position, we analyzed data scored with the markers placed in two different positions. One data set was created by placing the markers when participants got into bed and out of bed (i.e., including the Pre-Sleep and Post-Sleep periods; termed ActiGraph-IB, for "In Bed"). The second data set was created by placing the markers at bedtime and at the participants' requested wake time (i.e., excluding the Pre-Sleep and Post-Sleep periods, equivalent to the traditional "lights off" and "lights on" times; termed ActiGraph-BT, for "Bedtime").

For three participants, Fitbit failed to record data correctly and for one participant ActiGraph data was lost due to a technical error. Final participant numbers for each device were Oura: 63, Fitbit: 60, ActiGraph: 62, Xiaomi: 29, and Axtro: 31.

Reference technology

PSG was collected using the SOMNOtouch RESP system (SOMNOmedics GmbH, Randersacker, Germany). Electroencephalogram data was recorded from electrodes C3 and C4 referenced to the contralateral mastoids (M2 and M1). Electro-oculogram, electromyogram, and electrocardiogram were recorded using standard practices. The PSG was scored by consensus using three systems (Neurobit: Neurobit Inc, New York; Somnolyzer 24×7: The Siesta Group Schlafanalyse GmbH, Vienna, Austria; U-Sleep: https://sleep.ai.ku.dk[25]), according to a hybrid American Academy of Sleep Medicine[26] (AASM) and Rechtschaffen and Kales[27] system[9] (see Section 1.4 of the Supplementary Material).

Core analytics and main outcome variables

Analyses were conducted using Matlab version R2022b (The MathWorks Inc, Natick, MA[28]), R version 4.1.1 (R Core Team, Vienna, Austria[29]), and IBM SPSS Statistics version 27.0 (IBM Corp, Armonk, NY[30]).

Detection and timing accuracy for the initial bedtime, and the Sleep Onset Time (SOT) and Wake Time (WT) associated with Sleep Periods 1 and 2 were examined. SOT was defined as the first epoch of any sleep stage and WT as the first epoch of wake. Detection accuracy for each device was calculated as the proportion of sleep/wake transitions reported during the correct time window (defined in the relevant Results section). Timing accuracy is reported as the mean (and standard deviation) and median (and median absolute deviation) difference from the PSG-derived SOT or WT. The Mean Absolute Error (MAE; calculated as the mean of the absolute differences between the device and PSG-derived times) is also reported as an indication of error magnitude. Participants who fell asleep during the peri-sleep activities were excluded from the relevant analyses (seven participants fell asleep during the Pre-Sleep period, none during the Interruption period and two during the Post-Sleep period).

Additional analytics and exploratory analyses

Traditional sleep measures and performance evaluation metrics were calculated in an epoch-by-epoch manner following standard procedures.[3] This analysis was restricted to the 50 participants who had complete PSG, Oura, Fitbit, and ActiGraph data from both nights and who stayed awake during the peri-sleep periods. To assess how the devices were affected by the simulated peri-sleep behavior, we compared sleep measures and performance metrics across both nights of the protocol. Additionally, to examine how inclusion of peri-sleep activity affected these sleep measures, we compared the

Table 2
Start time and duration of peri-sleep activity periods for participants in each age group

			Age group			
Activity period	Measure	Overall	Young	Middle	Older	*p*-value
Pre-sleep	Start time (HH:MM)	23:16 (68.3 min)	23:55 (55.7 min)^{a}	23:19 (58.3 min)^{a}	22:30 (66.3 min)^{b}	< .001
	Duration (min)	36.04 (7.32)	33.88 (6.30)^{a}	33.72 (6.04)^{a}	41.24 (7.43)^{b}	< .001
Interruption	Start time (HH:MM)	00:36 (71.9 min)	01:04 (61.6 min)^{a}	00:32 (58.4 min)^{ab}	00:09 (88.0 min)^{b}	.05
	Duration (min)	3.30 (1.66)	2.71 (1.81)^{a}	2.91 (0.98)^{a}	4.42 (1.66)^{b}	< .001
Post-sleep	Start time (HH:MM)	07:06 (70.1 min)	07:34 (62.3 min)^{a}	07:01 (70.2 min)^{ab}	06:39 (69.7 min)^{b}	.04
	Duration (min)	15.90 (6.80)	12.81 (4.03)^{a}	15.30 (7.09)^{a}	20.03 (7.08)^{b}	< .01

Mean (and standard deviation) start time and duration of peri-sleep activity periods. Statistical significance column shows ANOVA main effect significance. Values with different letters are significantly different from each other (pairwise comparison: $p < .05$).

Table 3
PSG-derived sleep measures for all participants in each age group

PSG sleep measure	Overall	Age group			*p*-value
		Young	Middle	Older	
TIB (min)	433.21 (43.13)	425.60 (48.16)	428.33 (41.81)	447.53 (37.14)	.22
TST (min)	379.49 (42.66)	384.33 (45.74)[ab]	391.54 (40.79)[a]	359.55 (35.74)[b]	.04
WASO (min)	43.45 (32.38)	29.69 (20.96)[a]	31.28 (17.95)[a]	73.39 (37.06)[b]	< .001
SE (%)	87.88 (7.98)	90.48 (6.29)[a]	91.47 (4.53)[a]	80.68 (8.45)[b]	< .001
SP1 SOL (min)	10.46 (12.90)	11.64 (15.13)[ab]	5.50 (5.20)[a]	15.15 (15.16)[b]	.04
SP2 SOL (min)	10.14 (16.65)	9.76 (18.74)	8.43 (14.92)	12.63 (16.79)	.72
Light sleep (min)	222.89 (42.99)	200.64 (46.41)[a]	232.43 (31.35)[b]	235.92 (40.10)[b]	.01
Deep sleep (min)	65.95 (32.31)	92.48 (30.60)[a]	60.21 (23.91)[b]	44.08 (23.76)[b]	< .001
REM sleep (min)	90.74 (24.90)	91.86 (22.78)[ab]	98.91 (19.44)[a]	79.61 (29.67)[b]	.04

Abbreviations: Deep sleep, N3 sleep; Light sleep, N1 and N2 sleep; REM sleep, rapid eye movement sleep; SE, sleep efficiency; SOL, sleep onset latency; SP1/2, sleep period 1/2; TIB, time in bed; TST, total sleep time; WASO, wake after sleep onset.
Mean (and standard deviation) PSG-derived sleep measures are shown. Statistical significance column shows ANOVA main effect significance. Values with different letters are significantly different from each other (pairwise comparison: $p < .05$).

full data recording period (including the peri-sleep periods) with data excluding the peri-sleep periods.

Results

Peri-sleep activity period timing and duration

Table 2 shows the timing and duration of the peri-sleep activity periods for the three age groups. In general, the activity periods started earlier, and were longer, for the Older group than the Young and Middle age groups.

PSG-derived sleep measures

PSG-derived sleep measures for each age group are shown in Table 3. Overall sleep patterns are consistent with research characterizing changes in sleep with age.[31] Participants in the Older group had shorter and less efficient sleep than the other groups; participants in both the Older and Middle groups showed less Deep sleep and more Light sleep than those in the Young group.

Bedtime detection

Bedtime is typically defined as the time at which a person goes to bed with the intention to sleep. In laboratory studies, this usually corresponds to "lights off" time. In this study, participants got into bed, and lights were turned off, *prior to* the Pre-Sleep period. Therefore, bedtime was defined as the first epoch following completion of the Pre-Sleep activities, indicating the point at which participants first tried to fall asleep.

Of the devices evaluated, only Oura and Fitbit provided automated measures of bedtime. Both devices reported bedtime as later than it actually was; Oura showed a mean difference of 8.57 minutes (SD = 24.65) and Fitbit 9.75 minutes (SD = 23.97). This is consistent with findings indicating that wearables typically underestimate SOL.[32] Fig. 2 shows the distribution of differences in bedtimes reported by

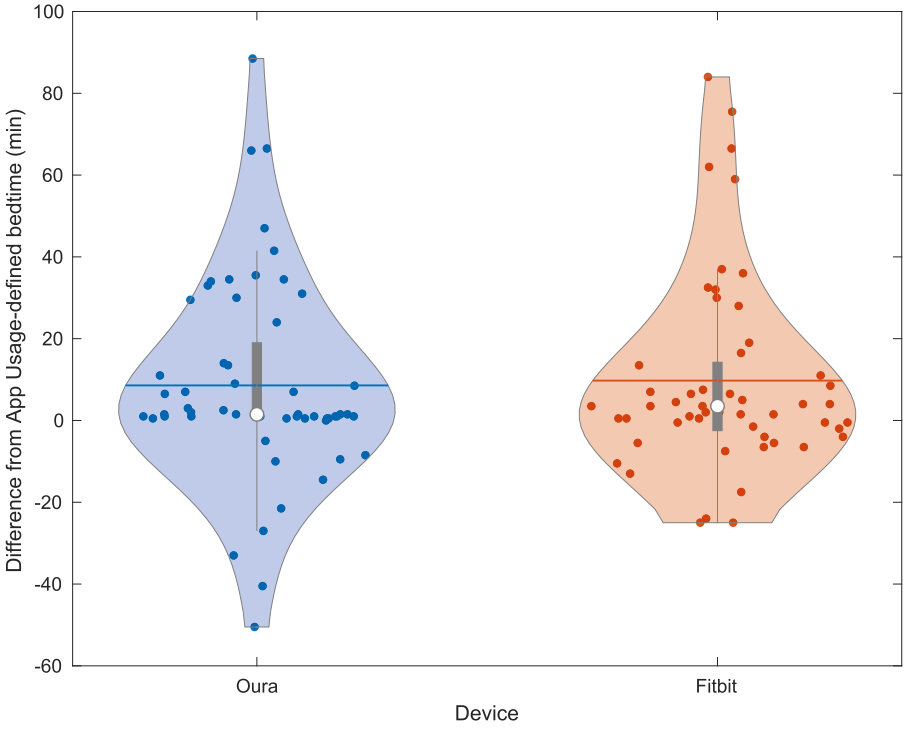

Fig. 2. Differences between bedtimes reported by Oura and Fitbit compared with bedtime defined as the first epoch following the Pre-Sleep period. Colored dots represent individual participants; the white dot shows the median; the horizontal line shows the mean; the gray bar indicates the first/third quartile; and the gray line indicates the upper/lower adjacent value (same for all violin plots)

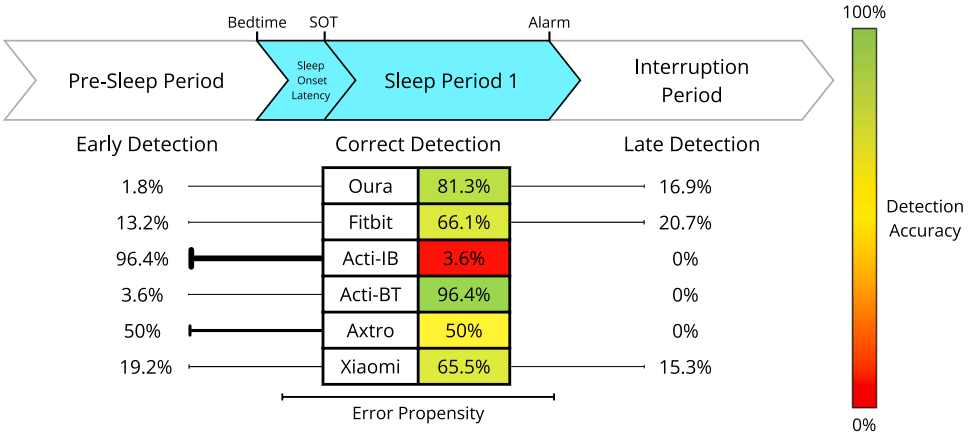

Fig. 3. Detection accuracy for Sleep Period 1 Sleep Onset Time (SP1 SOT). The blue region in the timeline indicates the window that was considered correct detection. The table below shows color-coded accuracy for each device in detecting SOT in the correct window. Lines indicate the extent and direction of errors

Oura and Fitbit compared to the phone usage defined bedtime for all participants.[33] While the mean and median differences were relatively small, the MAE was notably larger (Oura: 16.43 (SD = 20.19) minutes; Fitbit: 15.76 (SD = 20.45) minutes). There were also several instances when devices misjudged bedtime by over an hour.

Sleep Period 1

An epoch-by-epoch analysis (see Section 2.1 of the Supplementary Material) showed that ActiGraph misclassified wake as sleep more often than Oura and Fitbit. However, epoch-by-epoch analyses characterize classification performance without regard to error distribution and may not reflect how accurately devices detect sleep/wake transitions during the night. In the following sections, we describe detection and timing accuracy for these timepoints in temporal order.

Sleep Period 1 SOT could be detected in one of three time windows: incorrectly during the Pre-Sleep period (before bedtime), during the correct window after bedtime and before the Interruption (when the participant was either trying to fall asleep or was asleep) or after the Interruption. Fig. 3 shows the detection rate for each device in each time window. Table 4 (first section) presents this information together with the distribution of instances, and the average difference from the PSG-derived SOT across participants.

ActiGraph-BT detected SP1 SOT accurately most of the time, with no sleep onsets detected late, and only two before bedtime. However, ActiGraph-IB accuracy was dramatically lower, with sleep onset detected early almost all the time. Oura performed fairly accurately, reporting sleep onset early only once, but detecting it late nine times. Fitbit and Xiaomi performed similarly with a roughly even split in errors detecting the sleep onset early and late. Axtro, while never missing the sleep onset, reported it before bedtime for half of the participants.

Fig. 4 shows the distribution of SP1 SOT estimation accuracy; that is, the difference between device-reported and PSG-derived SOT. Oura, Fitbit, ActiGraph-BT, and Xiaomi cluster around zero, indicating reasonably good estimation of SOT for most participants. Large differences reflect early and late detection of SOT (see Supplementary Figs. S1 and S4). ActiGraph-IB SOT estimation is almost 40 minutes early, consistent with its tendency to misclassify wake as sleep during the Pre-Sleep period. Axtro also did not estimate SOT accurately, reporting SOT approximately 15 minutes early.

Sleep Efficiency and Sleep Period 1 SOT detection accuracy

SE is often used as an indicator of sleep continuity. Fig. 5 shows, for each device, the relationship between SE and accuracy of SP1 SOT estimation. Members of the three age groups are shown in different colors. For most devices, at higher SE there is a cluster of points relatively close to zero. The spread of this cluster is low, suggesting that when SE is high, SOT is estimated accurately. However, at lower SE (below ~85%) points are more spread out, indicating that SOT estimation is less accurate. The slope of the regression lines suggest that ActiGraph reported SOT increasingly earlier for participants with lower SE.

Mid-sleep interruption

All participants were woken up from N2 or N3 sleep to complete the activities during the Interruption period. The Interruption Wake Time (IWT) could only be estimated if the device identified sleep beforehand; therefore, this analysis includes only those participants where the device recorded sleep prior to the Interruption.

The IWT was considered correctly detected if it was identified in the time window from the epoch prior to the PSG-defined awakening caused by the alarm (to allow for discrepancies caused by splitting 1-minute epochs for some devices) to Sleep Period 2 SOT (i.e., while the participant was awake following the Interruption). There are two potential types of errors; one where the participant was incorrectly classified as awake prior to the Interruption alarm and one where wakefulness during the Interruption period was missed. In both cases, the correct transition from sleep to wake caused by the alarm was not identified correctly. Fig. 6 illustrates the detection windows and IWT detection rates, with additional results shown in the second section of Table 4.

Oura, Fitbit, and ActiGraph were very sensitive to awakening during the Interruption; it was missed only once. Axtro and Xiaomi, however, missed most or all of the Interruptions. Fig. 7 shows the IWT bias when it was not missed (including instances when the device reported the wake Interruption before its detection by PSG). Oura, Fitbit, and ActiGraph frequently detected the wake Interruption within one epoch (30 seconds) of the PSG wake epoch (see Supplementary Fig. S1 for the Oura outlier). When Axtro successfully detected the wake Interruption, it was reported on average over 5 minutes after PSG. Xiaomi is not shown in this figure because it missed all of the IWTs (see Supplementary Fig. S2).

Table 4
Detection accuracy, timing accuracy, and error magnitude for Sleep Period 1 Sleep Onset Time, Interruption Wake Time, Sleep Period 2 Sleep Onset Time and Morning Wake Time

		Device					
		Oura	Fitbit	ActiGraph-IB	ActiGraph-BT	Axtro	Xiaomi
Sleep Period 1 Sleep Onset Time	Detection accuracy	0.82 (1/56 early) (9/56 late)	0.66 (7/53 early) (11/53 late)	0.04 (53/55 early) (0/55 late)	**0.96 (2/55 early) (0/55 late)**	0.5 (14/28 early) (0/28 late)	0.65 (5/26 early) (4/26 late)
	Bias: Median (MAD, min)	5.75 (7.04)	2 (10.38)	−36 (10.38)	**−1 (5.19)**	−15.5 (30.39)	3.25 (12.23)
	Bias: Mean (SD, min)	13.04 (23.65)	6.33 (22.00)	−38.87 (18.31)	**−5.31 (14.10)**	−14.91 (28.38)	5.54 (35.30)
	MAE (SD, min)	16.93 (20.99)	15.16 (18.32)	39.07 (17.87)	**8.11 (12.67)**	25.41 (19.15)	22.46 (27.44)
Interruption Wake Time	Detection accuracy	**0.94 (3/54 early) (0/54 missed)**	0.83 (8/47 early) (0/47 missed)	0.63 (22/62 early) (1/62 missed)		0.23 (0/31 early) (24/31 missed)	0 (0/23 early) (23/23 missed)
	Bias: Median (MAD, min)	0.5 (0)	−0.5 (0.74)	**0 (0.74)**		5 (2.97)	-
	Bias: Mean (SD, min)	−0.68 (7.80)	−0.60 (1.18)	**−0.52 (1.71)**		5.43 (3.87)	-
	MAE (SD, min)	1.68 (7.64)	0.66 (1.15)	**0.96 (1.50)**		5.43 (3.87)	-
Sleep Period 2 Sleep Onset Time	Detection accuracy	**1 (63/63 correct)**	**1 (60/60 correct)**	**1 (62/62 correct)**		**1 (9/9 correct)**	**1 (7/7 correct)**
	Bias: Median (MAD, min)	2 (2.97)	**0 (4.45)**	−0.75 (5.56)		3.50 (7.41)	3 (4.45)
	Bias: Mean (SD, min)	1.44 (10.37)	−3.1 (17.57)	−5.96 (17.76)		5.11 (17.26)	**0.71 (18.18)**
	MAE (SD, min)	**5.64 (8.79)**	7.33 (16.25)	9.10 (16.35)		10.44 (14.30)	11.43 (13.37)
Morning Wake Time	Detection accuracy	0.98 (1/61 early) (0/61 missed)	0.97 (2/58 early) (0/58 missed)	**1 (60/60 correct)**	**1 (60/60 correct)**	0.48 (0/31 early) (16/31 missed)	0.71 (0/28 early) (8/28 missed)
	Bias: Median (MAD, min)	**0.5 (0.74)**	**−0.5 (0.74)**	−1.5 (0.74)	−1.5 (0.74)	4.5 (6.67)	7.75 (2.97)
	Bias: Mean (SD, min)	−2.47 (14.17)	−4.03 (15.25)	**−1.5 (1.74)**	**−1.5 (1.85)**	3.3 (10.83)	7 (4.70)
	MAE (SD, min)	3.12 (14.03)	4.28 (15.18)	**1.68 (1.56)**	**1.68 (1.68)**	8.17 (7.58)	7.7 (3.35)

Abbreviations: MAD, median absolute deviation; MAE, mean absolute error.
Detection accuracy: Proportion (and number) of participants where the SOT/WT was detected in the correct window. Bias: Average difference between device and PSG-derived SOT/WT. The best-performing device for each timepoint and each measure is shown in bold (there are ties for some measures).

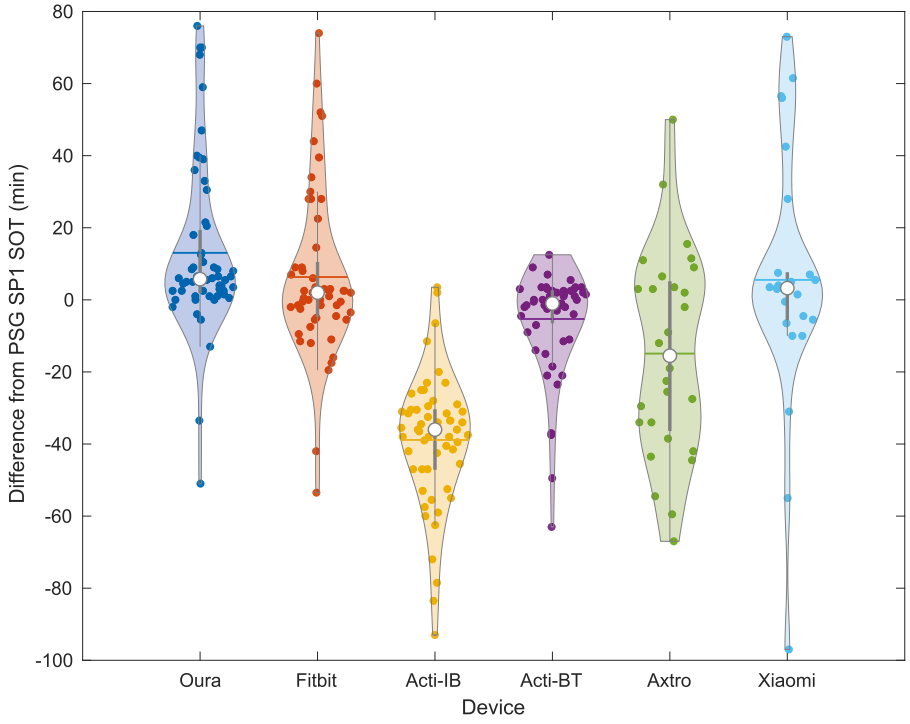

Fig. 4. Sleep Period 1 Sleep Onset Time difference from the first PSG Sleep Onset Time for each device

Sleep Period 2

All participants were able to fall asleep again following the Interruption. For each device, Sleep Period 2 SOT was defined as the first wake/sleep transition following the Interruption alarm up to 1 hour after the PSG-defined SP2 SOT (Fig. 8). SP2 SOT could only be calculated for participants where the device either recorded a wake period related to the Interruption or where Sleep Period 1 was missed (in which case, the first sleep onset recorded was for SP2). SP2 SOT could not be detected if the device identified SP1 but no Interruption wake period (see Supplementary Fig. S2). All devices detected SP2 SOT during the correct window when they identified wake during the Interruption period (see third section of Table 4).

In general, when devices detected SP2 SOT, it was fairly accurate, as illustrated by the clustering around zero in Fig. 9. There were some instances in which participants took a long time to return to sleep and devices reported SOT considerably earlier than PSG (see Supplementary Fig. S2).

Morning wake up

A morning alarm woke participants at their requested wake-up time, if they were not already awake. The Morning Wake Time (MWT) was considered missed if no transitions from sleep to wake were detected from 45 minutes before the final PSG awakening to the end of the Post-Sleep period (Fig. 10). Table 4 (bottom section) shows the detection accuracy and timing for the MWT (either in response to the alarm, or the final awakening before the alarm). Oura, Fitbit, and ActiGraph successfully detected a MWT for each participant; however, Axtro and Xiaomi frequently missed it.

Fig. 11 shows the differences between PSG and device estimates of MWT when it was not missed. Oura, Fitbit, and ActiGraph are all fairly accurate apart from a few instances where the MWT was estimated much earlier than the PSG-derived MWT. These were cases where the participant woke up early (which was detected by the device) and then fell back to sleep before waking up for the final time (which was not detected by the device; see Supplementary Fig. S3). Both Axtro and Xiaomi frequently failed to detect the morning awakening; furthermore, when it was detected, it was reported later than the PSG-derived MWT.

Sleep measures and performance evaluation metrics across nights

Table 5 shows PSG sleep measures and device discrepancy (upper section), and device performance metrics (lower section) for both nights of the protocol. The second night was analyzed both including the peri-sleep periods (labeled Night 2 IB, for “In Bed”) and excluding them (labeled Night 2 BT, for “Bedtime”).

PSG data showed that, when excluding the peri-sleep periods, TST and SE were similar on Night 1 and Night 2. However, both measures were underestimated by Oura and Fitbit on Night 2 compared with Night 1 (upper bold values in Table 5).

Including the peri-sleep periods in the analysis had a significant impact on performance evaluation metrics. For both Oura and Fitbit, specificity increased, while for ActiGraph specificity decreased. Kappa scores for Oura and Fitbit also increased, while that for ActiGraph remained unchanged (lower bold values in Table 5).

Discussion

Main results and implications

Using a novel protocol to simulate common peri-sleep behavior, we found that modest deviation from a good bedtime routine does not significantly compromise the performance of higher-quality CSTs. Consistent with data from the first night of the study,[9] the high-quality CSTs performed considerably better than the lower-cost ones.

Oura and Fitbit displayed good detection accuracy for Sleep Period 1 SOT. While Oura’s detection accuracy was higher than Fitbit’s, it was biased towards detecting sleep onset later than it actually was. This is consistent with results from the first night of the protocol where Oura significantly overestimated sleep onset latency. While Axtro and Xiaomi performed moderately well for the initial

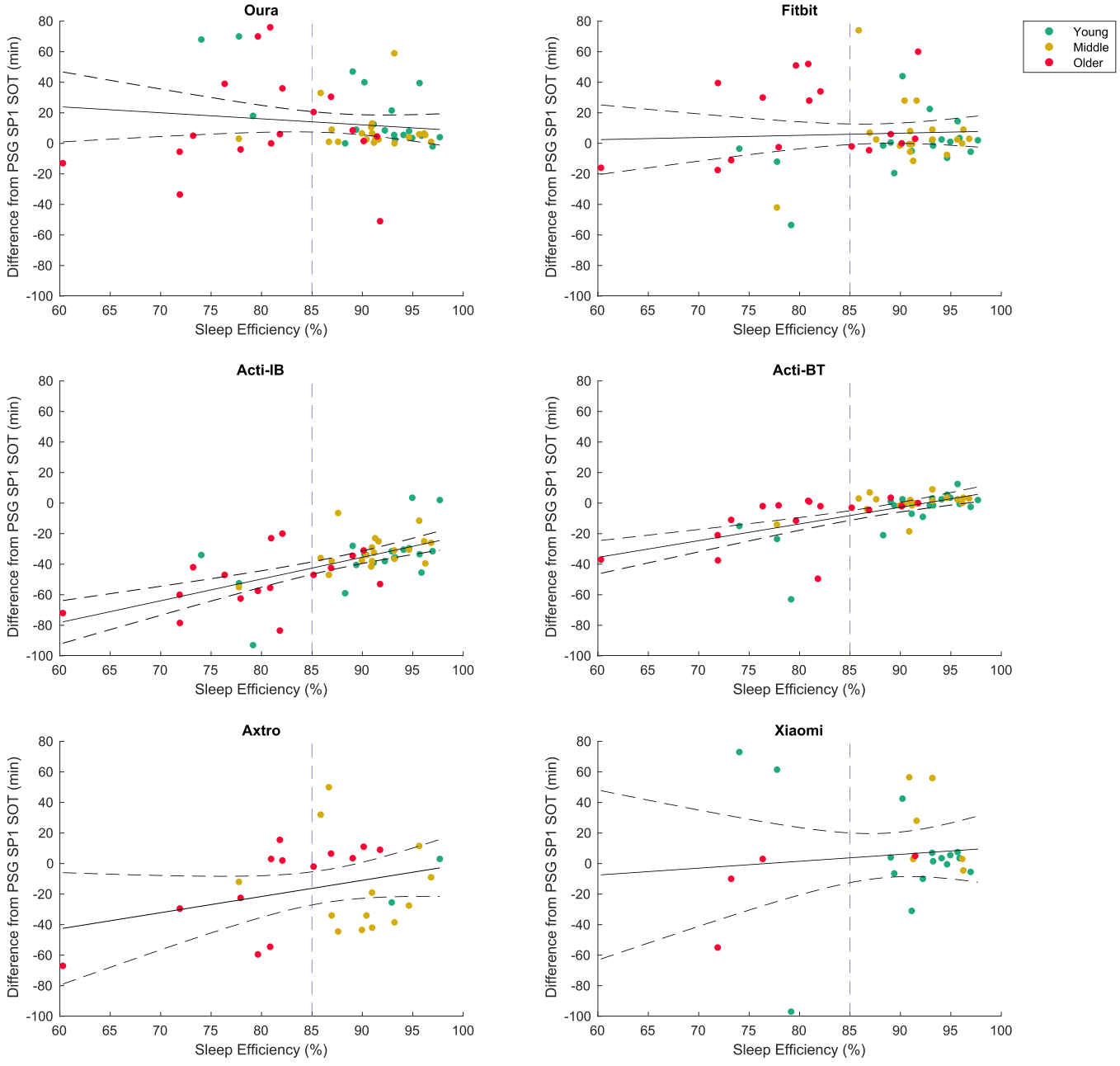

Fig. 5. Scatter plots showing the relationship between sleep efficiency (SE) and Sleep Period 1 Sleep Onset Time accuracy. Participants from the Young group are shown in Green, the Middle group in Gold, and the Older group in Red. The regression line is plotted in black with confidence intervals (dashed black lines). The vertical dashed blue line marks 85% SE

sleep onset, their inability to detect awakenings may severely limit their usefulness.

ActiGraph's Sleep Period 1 SOT accuracy was significantly influenced by marker placement. Both detection and timing accuracy were markedly better for the ActiGraph-BT data, when the algorithm excluded the Pre-Sleep period. This improvement is not surprising because wake misclassification during the Pre-Sleep period, resulting in earlier estimation of sleep onset, is not considered in the ActiGraph-BT data; however, it does highlight the importance of correct marker placement for accurate sleep measurement. This pattern of results is consistent with ActiGraph-IB's overestimation of TST and SE, resulting from wake misclassified as sleep in the peri-sleep periods. It also suggests that in a controlled lab environment, where participants do not engage in typical peri-sleep activities, accuracy of sleep measures may be artificially inflated. Oura and Fitbit's performance during these periods suggests that this may be less of a concern for these devices.

Sleep Period 1 SOT timing accuracy was less reliable for those with more fragmented sleep. This is consistent with previous research, including Night 1 of this protocol[9,18,19] and reflects the difficulty these devices have in accurately measuring sleep that is significantly disrupted. Sleep fragmentation is associated with normal aging[34] as well as neurodegenerative disorders such as Alzheimer's disease[35] and Parkinson's disease[36]; therefore, data from wearable devices in these populations is likely to be less reliable than from those with well-consolidated sleep.

Detection accuracy for both the Interruption and Morning Wake Times was very high for Oura, Fitbit, and ActiGraph; however, these transitions were missed by Axtro and Xiaomi over half the time. ActiGraph's ability to detect both wake and sleep onset transitions

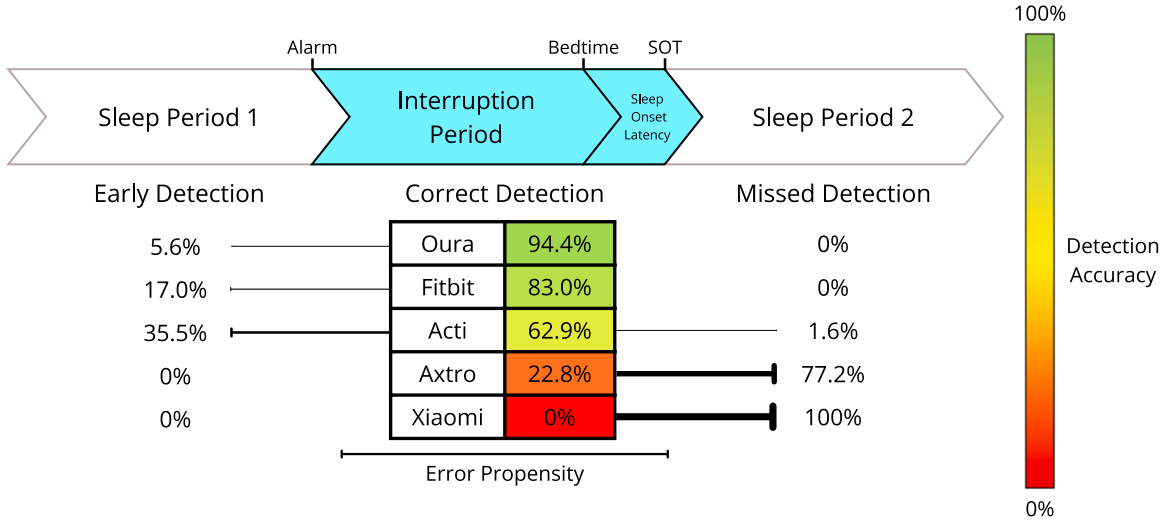

Fig. 6. Detection accuracy for the Interruption Wake Time (IWT). The blue region in the timeline indicates the window that was considered correct detection. The table below shows color-coded accuracy for each device in detecting the IWT in the correct window. Lines indicate the extent and direction of errors

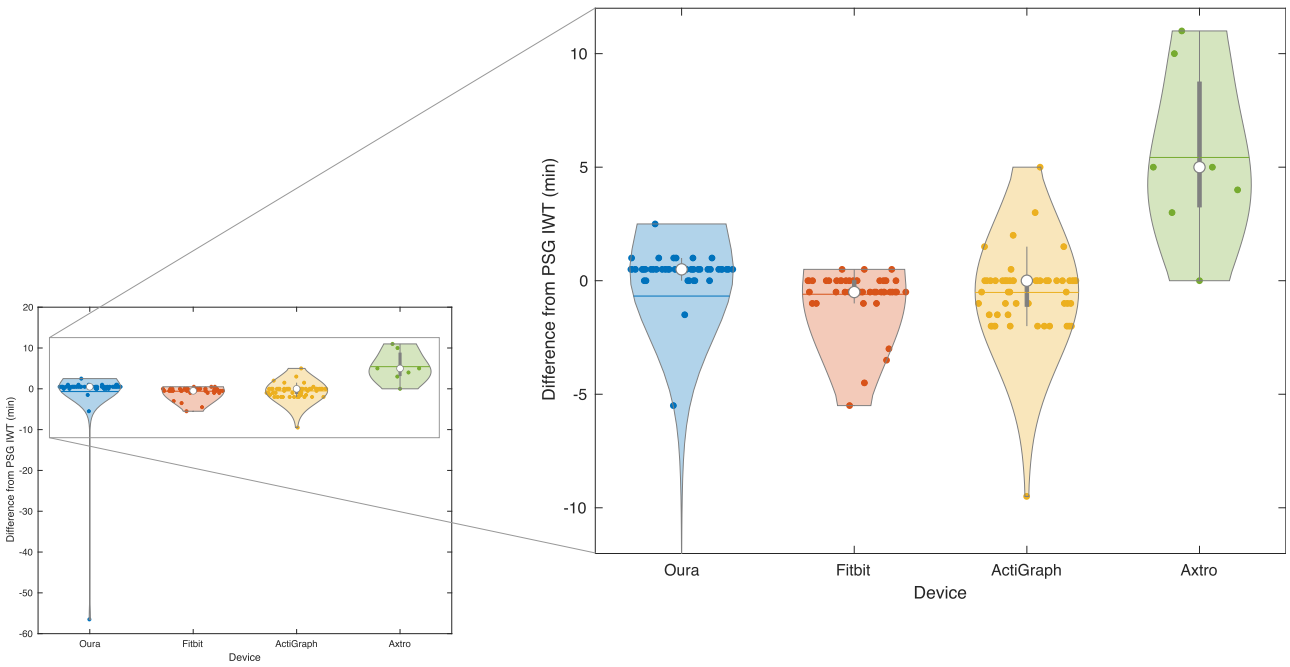

Fig. 7. Interruption Wake Time (IWT) accuracy for all devices and participants when it was not missed. The inset shows a zoomed-in version of the graph on the left. Xiaomi is not shown because it missed all Interruptions

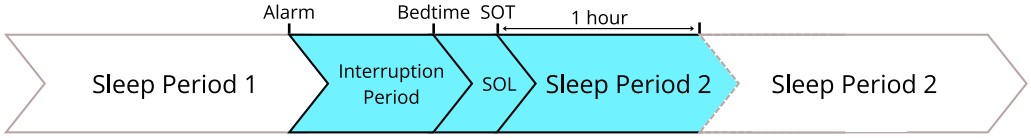

Fig. 8. Detection period for Sleep Period 2 Sleep Onset Time (SP2 SOT). The blue region indicates the window that was considered correct detection. All devices detected SP2 SOT in the correct window

when peri-sleep behavior was excluded suggests that it is sensitive to the movement associated with the sleep transitions in the protocol (e.g., putting the phone on the bedside table and finding a comfortable position to sleep in, or retrieving the phone and getting in position to do the activities). However, on either side of the movement, when the participant was still, ActiGraph reported the participant as asleep.

The marginally better sleep on Night 2 compared with Night 1 might be due to the first-night effect[37] and participants wearing the Dreem headband which was not well-tolerated by some participants on Night 1.[9,38] Both Oura and Fitbit significantly underestimated TST and SE on Night 2, but not Night 1. This might be because the wake interruption resulted in these CSTs missing the first SOT for some

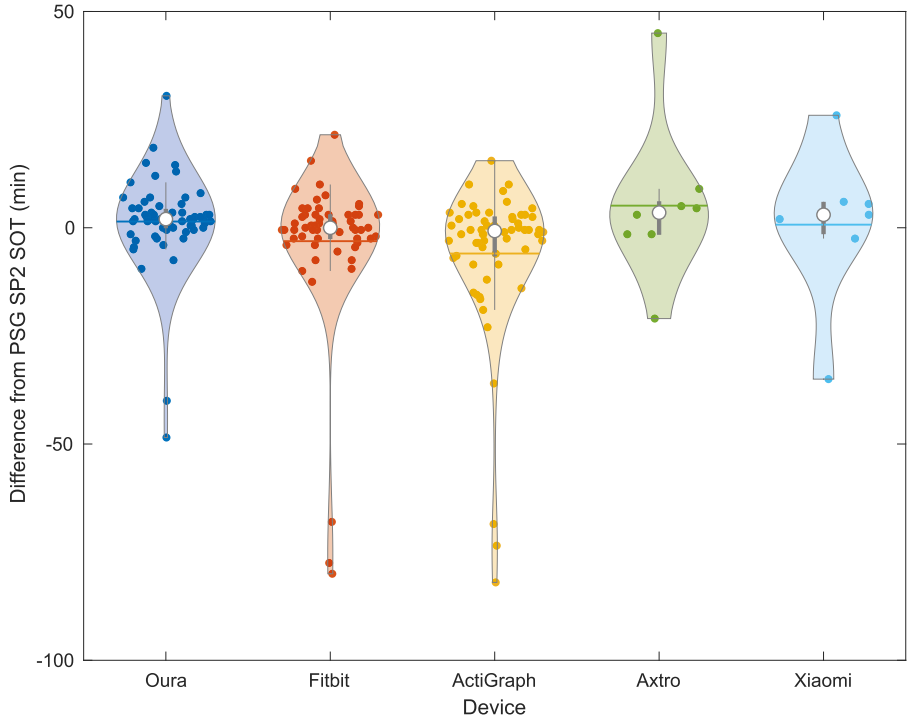

Fig. 9. Sleep Period 2 Sleep Onset Time accuracy for each device

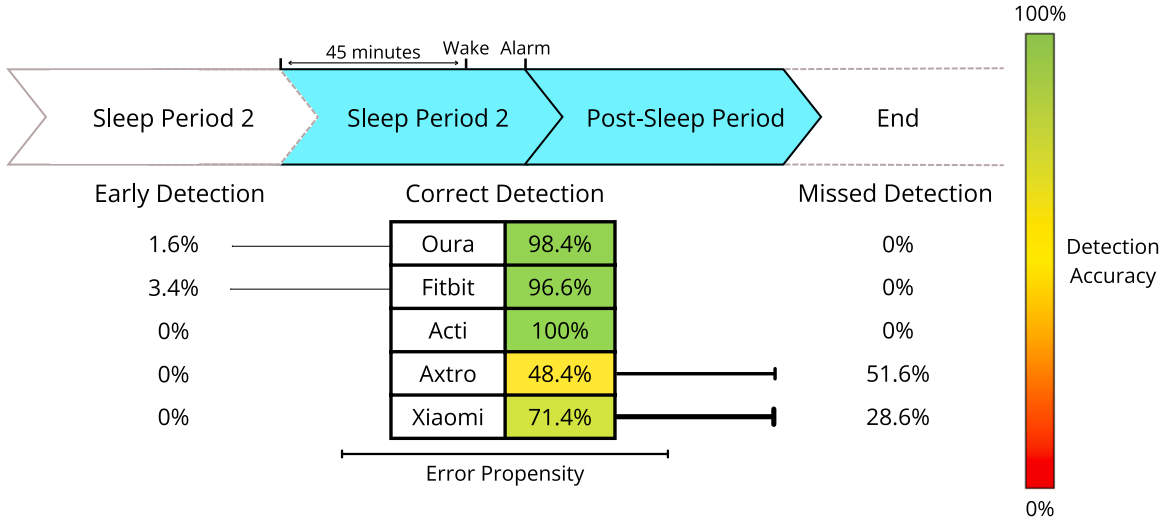

Fig. 10. Detection accuracy for the Morning Wake Time (MWT). The blue region in the timeline indicates the window that was considered correct detection. The table below shows color-coded accuracy for each device in detecting the MWT in the correct window. Lines indicate the extent and direction of errors

participants. This may, in turn, have resulted in a shorter average TST and lower SE. This explanation implies that an event which effects sleep at one point in time may have broader consequences for sleep measures at a different time.

There were significant differences in performance metrics depending on whether the peri-sleep periods were included in the analysis or not. Both specificity and Kappa increased for Oura and Fitbit because additional periods of predominantly correct wake classification were included in the analysis. For ActiGraph, specificity decreased and Kappa remained the same. This is because the classification accuracy of the periods added to the analysis were at chance levels (see Supplementary Table S2). The effect of this was to reduce specificity. However, the Kappa statistic was unaffected because it accounts for chance agreement, which was similar for the peri-sleep periods and the wake periods between bedtime and the morning alarm.

One reason that well-developed CSTs outperform "research-grade" actigraphy may be that they use data from multiple sensors, compared to actigraphs (typically including photoplethysmography and temperature sensors[39]). Several recent studies have shown that including data from additional sources in classification algorithms improves performance[40,41]; more broadly, research has also demonstrated the value of combining data sources to overcome limitations of individual measures.[42]

Another possible reason for CSTs' superior performance is that they use algorithms derived from larger training sets that include data from a

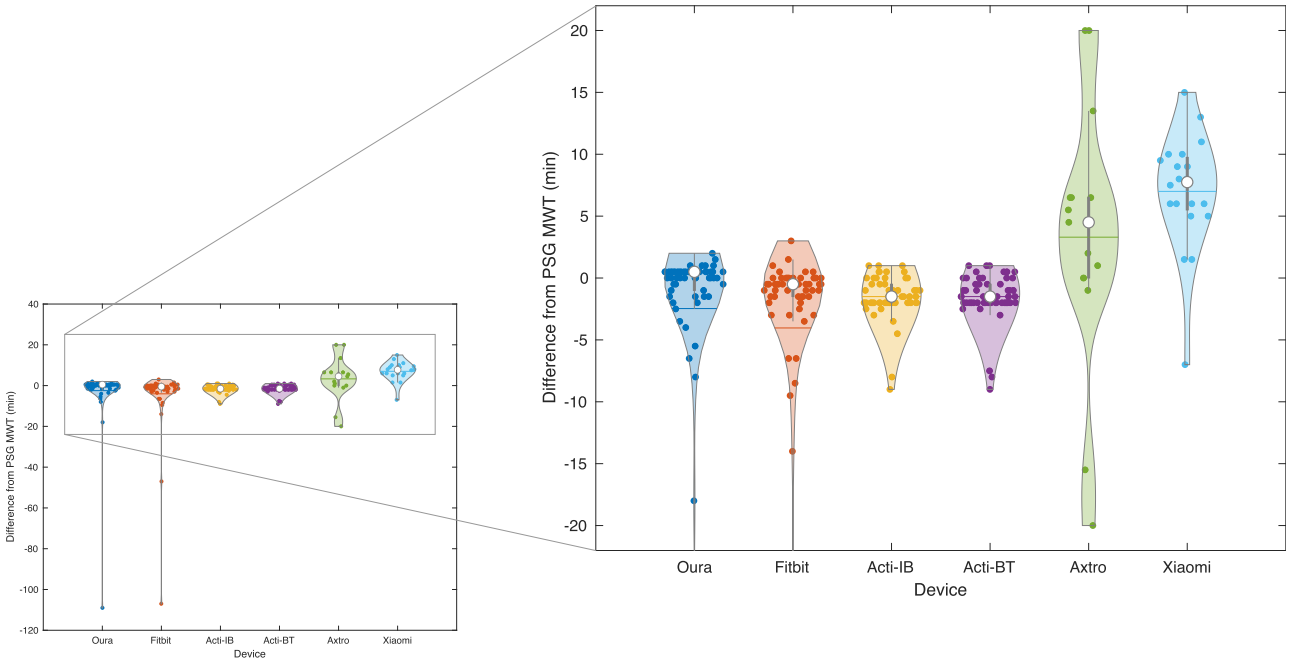

Fig. 11. Morning Wake Time accuracy for all devices and participants where a wake time was detected. The inset shows a zoomed-in version of the graph on the left

wider age range, and from people of different ethnicities. Actigraphs typically use variants of the Cole-Kripke[43] or Sadeh[44] algorithms, both of which were developed approximately 30 years ago. Using newer algorithms on the same actigraphy data can improve classification performance,[45,46] although the improvement afforded by new machine learning algorithms may not yet be as robust as originally envisioned.[47]

While high-quality CSTs can be improved further, their relative affordability, reliable performance and already extensive deployment provide valuable opportunities for sleep research. Large-scale studies on cross-country differences in sleep patterns have already been reported,[48-51] although they lack valuable phenotypic data that future studies should include.

Table 5
Discrepancy analyses comparing Oura, Fitbit, and ActiGraph with PSG across both nights of the protocol and including/excluding peri-sleep periods (N = 50)

Measure	Dataset	PSG	Oura	Fitbit	ActiGraph
			Bias measures across device		
TST (min)	Night 1	365.06 (55.77)	**2.77 (37.40)$^{a\ X}$**	**9.97 (43.18)$^{ab\ X}$**	12.60 (46.82)$^{b\ X}$
	Night 2 BT	377.51 (42.62)	**−10.68 (26.70)$^{*ab\ Y}$**	**−12.65 (33.69)$^{*a\ Y}$**	−2.88 (41.50)$^{b\ Y}$
	Night 2 IB	377.51 (42.62)	−9.99 (25.98)$^{*a\ Y}$	−10.52 (35.64)$^{*a\ Y}$	21.36 (47.51)$^{*b\ X}$
SE (%)	Night 1	84.42 (10.14)^{X}	**0.60 (8.91)^{X}**	**2.04 (9.84)^{X}**	2.76 (10.84)^{X}
	Night 2 BT	87.57 (8.33)^{X}	**−2.59 (6.14)^{*Y}**	**−3.12 (7.57)^{*Y}**	−0.95 (9.39)^{Y}
	Night 2 IB	78.48 (8.05)^{Y}	−2.15 (5.19)$^{*a\ Y}$	−2.36 (7.09)$^{*a\ Z}$	4.11 (9.46)$^{*b\ X}$
SOL (min)	Night 1	15.52 (14.55)^{X}	10.99 (25.80)^{*a}	−1.39 (23.20)^{b}	−9.97 (15.22)$^{*c\ X}$
	Night 2 BT/IB	9.92 (12.15)^{Y}	16.07 (23.99)^{*a}	8.60 (21.98)^{*a}	−3.70 (12.39)$^{*b\ Y}$
WASO (min)	Night 1	52.81 (41.61)^{XY}	−13.76 (32.63)^{*a}	−8.58 (36.78)^{ab}	−2.64 (42.93)$^{b\ X}$
	Night 2 BT	45.32 (32.78)^{X}	−5.45 (21.99)^{a}	3.99 (27.00)^{b}	6.58 (34.62)$^{*b\ X}$
	Night 2 IB	60.13 (35.25)^{Y}	−3.80 (25.00)^{a}	3.80 (27.71)^{a}	15.91 (37.92)$^{*b\ Y}$
			Performance measures across device		
Accuracy (%)	Night 1	-	90.53 (7.13)$^{a\ X}$	89.04 (7.83)$^{b\ XY}$	86.92 (8.06)$^{c\ XY}$
	Night 2 BT	-	91.35 (5.20)$^{a\ XY}$	89.71 (6.29)$^{b\ X}$	88.45 (6.50)$^{b\ X}$
	Night 2 IB	-	92.12 (4.90)$^{a\ Y}$	90.39 (5.72)$^{b\ Y}$	84.69 (6.76)$^{c\ Y}$
Sensitivity (%)	Night 1	-	94.72 (4.16)	94.68 (3.99)^{X}	93.96 (4.43)
	Night 2 BT/IB	-	93.51 (4.75)	92.32 (5.17)^{Y}	92.94 (5.14)
Specificity (%)	Night 1	-	72.58 (19.30)$^{a\ X}$	66.12 (21.89)$^{b\ X}$	54.78 (22.52)$^{c\ X}$
	Night 2 BT	-	**75.15 (14.54)$^{a\ X}$**	**78.68 (17.71)$^{a\ Y}$**	**66.86 (22.27)$^{b\ Y}$**
	Night 2 IB	-	**88.07 (8.54)$^{a\ Y}$**	**86.40 (14.19)$^{a\ Z}$**	**58.23 (18.70)$^{b\ X}$**
Kappa	Night 1	-	0.63 (0.14)$^{a\ X}$	0.57 (0.14)$^{b\ X}$	0.49 (0.15)^{c}
	Night 2 BT	-	**0.61 (0.14)$^{a\ X}$**	**0.57 (0.15)$^{a\ X}$**	**0.50 (0.16)^{b}**
	Night 2 IB	-	**0.78 (0.10)$^{a\ Y}$**	**0.73 (0.12)$^{b\ Y}$**	**0.52 (0.15)^{c}**

Abbreviations: SE, sleep efficiency; SOL, sleep onset latency; TST, total sleep time; WASO, wake after sleep onset.
The PSG column presents mean (and standard deviation) of the sleep measures. Values in the Oura, Fitbit, and ActiGraph columns are means (and standard deviations) of device-PSG differences (bias), where positive and negative values indicate over- and underestimation compared with PSG, respectively. Dataset: Night 2 BT – Night 2 data excluding peri-sleep periods; Night 2 IB – Night 2 data including peri-sleep periods. Asterisks (*) indicate bias is significantly different from 0 ($p < .05$); a,b,c letters indicate significant differences between devices (read across rows; $p < .05$); X,Y,Z letters indicate significant differences between data sets (read down column for each measure; $p < .05$). All p-values are Bonferroni corrected for multiple comparisons. Comparisons discussed in the text are indicated in bold font.

Limitations and future perspectives

The technological ecosystem in which today's CSTs are being developed is significantly different from that in which the original actigraphs were created. This means that algorithm and software improvements can be deployed more quickly and at a much larger scale than before. The change in marker position for the ActiGraph data resulted in a significant improvement in performance, suggesting that CSTs could also exploit user input to improve sleep-tracking performance. For example, allowing users to indicate when they get into and out of bed, and when they try to fall asleep and wake up, would help disambiguate periods that are challenging for CSTs to classify accurately. While consistent user input of this kind is unlikely to be sustained in the long term, it could be used to give dedicated users more accurate sleep measures when they need it most.

Furthermore, with the wealth of data currently being collected, algorithms could be tuned to specific populations to optimize performance.[52] One feature that many research actigraphs have, which CSTs are yet to implement, is the ability to change sensitivity threshold levels. This may be particularly important in maintaining accuracy across different populations (e.g., children and adolescents, particularly boys, move more than adults during sleep, which requires different sensitivity thresholds to analyze appropriately[53,54]). Tuning sleep detection algorithms to accommodate aspects of physiology reliably associated with different sleep conditions as well as better handling of age effects are improvements to look forward to as more data is collected from increasingly diverse populations. However, this will require high-fidelity data labeling to fully realize, and it remains to be seen if this aspiration receives the necessary cooperation from device users and manufacturers.

Different devices provide sleep measures at different temporal resolutions. In order to compare these to PSG sleep measures, they need to be resampled to PSG epoch length (typically 30 seconds). This may, however, affect discrepancy measures, particularly for people with fragmented sleep who experience many stage transitions. While this is unlikely to have a significant impact on large effects found in performance evaluation studies in healthy sleepers, it should be considered when comparing device performance.

Core conclusions

Our results indicate that higher-quality CSTs provide robust sleep measurement in the face of a suboptimal bedtime routine and moderate sleep disruption. However, as accuracy decreased with lower sleep quality, these devices can be expected to be less accurate in those with insomnia or disturbed sleep. Future research, involving performance evaluation in different settings, during different user behaviors and on different populations will prove invaluable in delineating the boundaries within which CSTs provide acceptable sleep measures. The paradigm described in the present work should prove an important methodological advance for that purpose.

Funding

This work was supported by funds from Oura Health Oy (Finland), the Yong Loo Lin School of Medicine (National University of Singapore, Singapore), the Lee Foundation (Singapore), and the National Medical Research Council Singapore (STaR May2019-001) awarded to Dr Michael W.L. Chee.

Author contributions

A.W.: Conceptualization, Formal analysis, Visualization, Writing – original draft, Writing – review and editing. **H.G. and S.G.:** Formal analysis, Data curation, Visualization. **K.W. and N.C.:** Investigation, Data curation. **J.O.:** Conceptualization, Project administration, Supervision, Writing – review and editing. **M.C.:** Conceptualization, Supervision, Funding acquisition, Writing – original draft, Writing – review and editing.

Declaration of conflicts of interest

Oura Health Oy funded the data collection for the evaluation of its new sleep staging algorithm (OSSA 2.0), but the company did not influence the design of the study, analyses, its interpretation, or data presentation. All other equipment was contributed by the Sleep and Cognition Laboratory.

Acknowledgments

The authors would like to thank Yashmit Lepcha, Liang Tian, and Nicha Turton for their assistance in data collection and Kyra Chong for figure preparation.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.sleh.2024.02.007.

References

1. de Zambotti M, Goldstein C, Cook J, et al. State of the science and recommendations for using wearable technology in sleep and circadian research. Published online *Sleep.* 2023:zsad325. https://doi.org/10.1093/sleep/zsad325
2. 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. https://doi.org/10.1093/sleep/zsz254
3. 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. https://doi.org/10.1093/sleep/zsaa170
4. Chee NI, Ghorbani S, Golkashani HA, Leong RL, Ong JL, Chee MW. Multi-night validation of a sleep tracking ring in adolescents compared with a research actigraph and polysomnography. *Nat Sci Sleep.* 2021;13:177–190. https://doi.org/10.2147/NSS.S286070
5. Chinoy ED, Cuellar JA, Huwa KE, et al. Performance of seven consumer sleep-tracking devices compared with polysomnography. *Sleep.* 2021;44(5):zsaa291. https://doi.org/10.1093/sleep/zsaa291
6. 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. https://doi.org/10.2147/NSS.S348795
7. 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. https://doi.org/10.1111/jsr.12931
8. 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. https://doi.org/10.5664/jcsm.7932
9. Ong JL, Golkashani HA, Ghorbani S, et al. Selecting a sleep tracker from EEG-based, iteratively improved, low-cost multisensor, and actigraphy-only devices. *Sleep Health J Natl Sleep Found.* 2024;10(1):9–23. https://doi.org/10.1016/j.sleh.2023.11.005
10. Blood ML, Sack RL, Percy DC, Pen JC. A comparison of sleep detection by wrist actigraphy, behavioral response, and polysomnography. *Sleep.* 1997;20(6):388–395. Accessed October 2, 2023. ⟨https://academic.oup.com/sleep/article-abstract/20/6/388/2732123⟩.
11. de Zambotti M, Goldstone A, Claudatos S, Colrain IM, Baker FC. A validation study of Fitbit Charge 2™ compared with polysomnography in adults. *Chronobiol Int.* 2018;35(4):465–476. https://doi.org/10.1080/07420528.2017.1413578
12. Sadeh A, Hauri PJ, Kripke DF, Lavie P. The role of actigraphy in the evaluation of sleep disorders. *Sleep.* 1995;18(4):288–302. https://doi.org/10.1093/sleep/18.4.288

13. Van De Water ATM, Holmes A, Hurley DA. Objective measurements of sleep for non-laboratory settings as alternatives to polysomnography - a systematic review: a systematic review of objective sleep measures. *J Sleep Res.* 2011;20(1pt2):183–200. https://doi.org/10.1111/j.1365-2869.2009.00814.x
14. Lefkowitz C. Staff Perspective: Why We Don't Recommend Wearable Sleep Trackers for the Assessment and Treatment of Insomnia Disorder. Published 2021. Available at: ⟨https://deploymentpsych.org/blog/staff-perspective-why-we-don%E2%80%99t-recommend-wearable-sleep-trackers-assessment-and-treatment⟩. Accessed January 10, 2024.
15. Chung SJ, An H, Suh S. What do people do before going to bed? A study of bedtime procrastination using time use surveys. *Sleep.* 2020;43(4):zsz267. https://doi.org/10.1093/sleep/zsz267
16. Exelmans L, Gradisar M, Van den Bulck J. Sleep latency versus shuteye latency: prevalence, predictors and relation to insomnia symptoms in a representative sample of adults. *J Sleep Res.* 2018;27(6):e12737. https://doi.org/10.1111/jsr.12737
17. Exelmans L, Van den Bulck J. Bedtime, shuteye time and electronic media: sleep displacement is a two-step process. *J Sleep Res.* 2017;26(3):364–370. https://doi.org/10.1111/jsr.12510
18. Kang SG, Kang JM, Ko KP, Park SC, Mariani S, Weng J. Validity of a commercial wearable sleep tracker in adult insomnia disorder patients and good sleepers. *J Psychosom Res.* 2017;97:38–44. https://doi.org/10.1016/j.jpsychores.2017.03.009
19. Paquet J, Kawinska A, Carrier J. Wake detection capacity of actigraphy during sleep. *Sleep.* 2007;30(10):1362. https://doi.org/10.1093/sleep/30.10.1362
20. Johns MW. A new method for measuring daytime sleepiness: the Epworth sleepiness scale. *Sleep.* 1991;14(6):540–545. https://doi.org/10.1093/sleep/14.6.540
21. Netzer NC, Stoohs RA, Netzer CM, Clark K, Strohl KP. Using the Berlin Questionnaire to identify patients at risk for the sleep apnea syndrome. *Ann Intern Med.* 1999;131(7):485. https://doi.org/10.7326/0003-4819-131-7-199910050-00002
22. Siclari F, LaRocque J, Postle B, Tononi G. Assessing sleep consciousness within subjects using a serial awakening paradigm. *Front Psychol.* 2013;4. https://doi.org/10.3389/fpsyg.2013.00542
23. Stephan AM, Lecci S, Cataldi J, Siclari F. Conscious experiences and high-density EEG patterns predicting subjective sleep depth. *Curr Biol.* 2021;31(24):5487–5500.e3. https://doi.org/10.1016/j.cub.2021.10.012
24. AZSoft Technology Inc. App Usage. Published online 2023. Available at: ⟨https://play.google.com/store/apps/details?id=com.a0soft.gphone.uninstaller&hl=en⟩.
25. Perslev M, Darkner S, Kempfner L, Nikolic M, Jennum PJ, Igel C. U-Sleep: resilient high-frequency sleep staging. *NPJ Digit Med.* 2021;4(1):72. https://doi.org/10.1038/s41746-021-00440-5
26. Iber C., Ancoli-Israel S., Chesson A. Quan S.F. The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology, and Technical Specification. American Academy of Sleep Medicine; 2007.
27. Rechtschaffen A, Kales A. *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects.* Bethesda, MD: US Department of Health, Education and Welfare. U.S. Public Health Service, U.S. Government Printing Office; 1968.
28. The MathWorks Inc. MATLAB version: 9.13.0 (R2022b). Published online 2022. Available at: ⟨https://www.mathworks.com⟩.
29. R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing; 2022. ⟨https://www.R-project.org/⟩.
30. IBM Corp. IBM SPSS Statistics for Macintosh. Published online 2020.
31. Ohayon MM, Carskadon MA, Guilleminault C, Vitiello MV. Meta-analysis of quantitative sleep parameters from childhood to old age in healthy individuals: developing normative sleep values across the human lifespan. *Sleep.* 2004;27(7):1255–1273. https://doi.org/10.1093/sleep/27.7.1255
32. Scott H, Lack L, Lovato N. A systematic review of the accuracy of sleep wearable devices for estimating sleep onset. *Sleep Med Rev.* 2020;49:101227. https://doi.org/10.1016/j.smrv.2019.101227
33. Bechtold B. Violin Plots for Matlab, Github Project. Published online 2016. doi:10.5281/zenodo.4559847.
34. Cooke JR, Ancoli-Israel S. Normal and abnormal sleep in the elderly. *Handb Clin Neurol.* 2011;98:653–665. https://doi.org/10.1016/B978-0-444-52006-7.00041-1
35. Peter-Derex L, Yammine P, Bastuji H, Croisile B. Sleep and Alzheimer's disease. *Sleep Med Rev.* 2015;19:29–38. https://doi.org/10.1016/j.smrv.2014.03.007
36. Gera A, Comella C. Sleep in Parkinson's disease. In: Gupta R, Neubauer DN, Pandi-Perumal SR, eds. *Sleep and Neuropsychiatric Disorders.* Singapore: Springer Nature Singapore; 2022:633–645. https://doi.org/10.1007/978-981-16-0123-1_31
37. Agnew HW, Webb WB, Williams RL. The first night effect: an EEG study of sleep. *Psychophysiology.* 1966;2(3):263–266. https://doi.org/10.1111/j.1469-8986.1966.tb02650.x
38. Spina MA, Andrillon T, Quin N, Wiley JF, Rajaratnam SM, Bei B. Does providing feedback and guidance on sleep perceptions using sleep wearables improves insomnia? Findings from Novel Insomnia Treatment Experiment ("NITE"), a randomised controlled trial. *Sleep.* 2023;46(9):zsad167. https://doi.org/10.1093/sleep/zsad167
39. 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;3:721919https://doi.org/10.3389/fdgth.2021.721919
40. Altini M, Kinnunen H. The promise of sleep: a multi-sensor approach for accurate sleep stage detection using the oura ring. *Sensors.* 2021;21(13):4302. https://doi.org/10.3390/s21134302
41. Haghayegh S, Khoshnevis S, Smolensky MH, Diller KR, Castriotta RJ. Deep neural network sleep scoring using combined motion and heart rate variability data. *Sensors.* 2020;21(1):25. https://doi.org/10.3390/s21010025
42. Massar SA, Chua XY, Soon CS, et al. Trait-like nocturnal sleep behavior identified by combining wearable, phone-use, and self-report data. *NPJ Digit Med.* 2021;4(1):90. https://doi.org/10.1038/s41746-021-00466-9
43. Cole RJ, Kripke DF, Gruen W, Mullaney DJ, Gillin JC. Automatic sleep/wake identification from wrist activity. *Sleep.* 1992;15(5):461–469. https://doi.org/10.1093/sleep/15.5.461
44. Sadeh A, Sharkey M, Carskadon MA. Activity-based sleep-wake identification: an empirical test of methodological issues. *Sleep.* 1994;17(3):201–207. https://doi.org/10.1093/sleep/17.3.201
45. Palotti J, Mall R, Aupetit M, et al. Benchmark on a large cohort for sleep-wake classification with machine learning techniques. *NPJ Digit Med.* 2019;2(1):50. https://doi.org/10.1038/s41746-019-0126-9
46. van Hees VT, Sabia S, Jones SE, et al. Estimating sleep parameters using an accelerometer without sleep diary. *Sci Rep.* 2018;8(1):12975https://doi.org/10.1038/s41598-018-31266-z
47. Patterson MR, Nunes AA, Gerstel D, et al. 40 years of actigraphy in sleep medicine and current state of the art algorithms. *NPJ Digit Med.* 2023;6(1):51. https://doi.org/10.1038/s41746-023-00802-1
48. Jonasdottir SS, Minor K, Lehmann S. Gender differences in nighttime sleep patterns and variability across the adult lifespan: a global-scale wearables study. *Sleep.* 2021;44(2):zsaa169. https://doi.org/10.1093/sleep/zsaa169
49. Kuula L, Gradisar M, Martinmäki K, et al. Using big data to explore worldwide trends in objective sleep in the transition to adulthood. *Sleep Med.* 2019;62:69–76. https://doi.org/10.1016/j.sleep.2019.07.024
50. Ong JL, Tandi J, Patanaik A, Lo JC, Chee MWL. Large-scale data from wearables reveal regional disparities in sleep patterns that persist across age and sex. *Sci Rep.* 2019;9(1):3415. https://doi.org/10.1038/s41598-019-40156-x
51. Willoughby AR, Alikhani I, Karsikas M, Chua XY, Chee MW. Country differences in nocturnal sleep variability: observations from a large-scale, long-term sleep wearable study. *Sleep Med.* 2023;110:155–165. https://doi.org/10.1016/j.sleep.2023.08.010
52. Khademi A, EL-Manzalawy Y, Master L, Buxton OM, Honavar VG. Personalized sleep parameters estimation from actigraphy: a machine learning approach. *Nat Sci Sleep.* 2019;11:387–399. https://doi.org/10.2147/NSS.S220716
53. Grandner MA, Rosenberger ME. *Actigraphic sleep tracking and wearables: historical context, scientific applications and guidelines, limitations, and considerations for commercial sleep devices. Sleep and Health.* Elsevier; 2019:147–157. https://doi.org/10.1016/B978-0-12-815373-4.00012-5
54. Meltzer LJ, Short M, Booster GD, et al. Pediatric motor activity during sleep as measured by actigraphy. *Sleep.* 2019;42(1):zsy196. https://doi.org/10.1093/sleep/zsy196