SLEEP, 2023, **XX**, 1–10
https://doi.org/10.1093/sleep/zsad288
Advance access publication 8 November 2023
Original Article

Original Article

Can we rely on wearable sleep-tracker devices for fatigue management?

Jaques Reifman[1,*], Nikolai V. Priezjev[1,2] and Francisco G. Vital-Lopez[1,2]

[1]Department of Defense Biotechnology High Performance Computing Software Applications Institute, Telemedicine and Advanced Technology Research Center, United States Army Medical Research and Development Command, Fort Detrick, MD, USA and
[2]The Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc., Bethesda, MD, USA

[*]Corresponding author. Jaques Reifman, Senior Research Scientist and Director, Department of Defense Biotechnology High Performance Computing Software Applications Institute, Telemedicine and Advanced Technology Research Center, U.S. Army Medical Research and Development Command, ATTN: FCMR-TT, 504 Scott Street, Fort Detrick, MD 21702-5012, USA. Email: jaques.reifman.civ@health.mil.

Abstract

Study Objectives: Wearable sleep-tracker devices are ubiquitously used to measure sleep; however, the estimated sleep parameters often differ from the gold-standard polysomnography (PSG). It is unclear to what extent we can tolerate these errors within the context of a particular clinical or operational application. Here, we sought to develop a method to quantitatively determine whether a sleep tracker yields acceptable sleep-parameter estimates for assessing alertness impairment.

Methods: Using literature data, we characterized sleep-measurement errors of 18 unique sleep-tracker devices with respect to PSG. Then, using predictions based on the unified model of performance, we compared the temporal variation of alertness in terms of the psychomotor vigilance test mean response time for simulations with and without added PSG-device sleep-measurement errors, for nominal schedules of 5, 8, or 9 hours of sleep/night or an irregular sleep schedule each night for 30 consecutive days. Finally, we deemed a device error acceptable when the predicted differences were smaller than the within-subject variability of 30 milliseconds. We also established the capability to estimate the extent to which a specific sleep-tracker device meets this acceptance criterion.

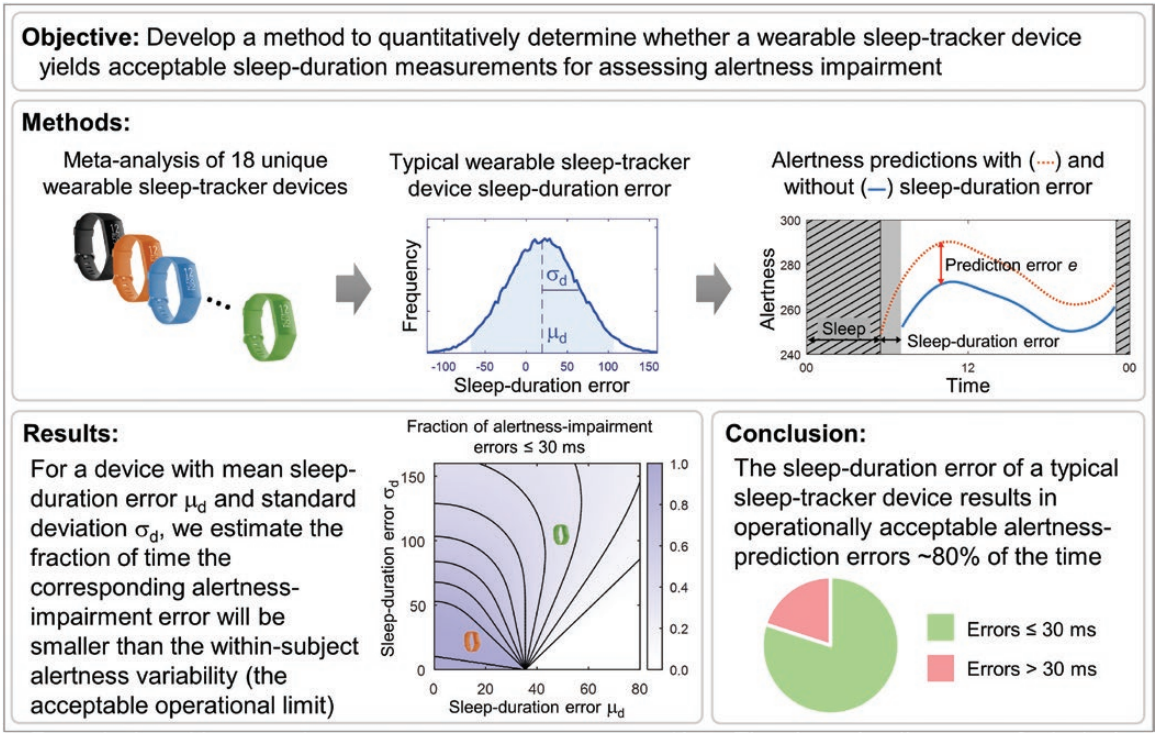
Results: On average, the 18 sleep-tracker devices overestimated sleep duration by 19 (standard deviation = 44) minutes. Using these errors for 30 consecutive days, we found that, regardless of sleep schedule, in nearly 80% of the time the resulting predicted alertness differences were smaller than 30 milliseconds.

Conclusions: We provide a method to quantitatively determine whether a sleep-tracker device produces sleep measurements that are operationally acceptable for fatigue management.

Key words: fatigue; sleep measures; sleep parameters; sleep-tracker device; wearables

Submitted for publication: August 11, 2023; **Revised:** October 30, 2023

Graphical Abstract



Statement of Significance

There is a general consensus that sleep parameters measured by wearable sleep-tracker devices are not error free. Moreover, to date, it is unclear to what extent we should trust their measurements within the context of a particular clinical or operational application. Here, we established a method to quantitatively assess whether sleep trackers are acceptable alternatives as part of a fatigue-management system for predicting alertness impairment. Importantly, for a specific device with its own sleep-measurement error characteristics, the method determines the extent to which these errors are smaller than the within-subject variability of alertness impairment and should be accepted. Future efforts should focus on generating new methods to assess the validity of sleep-tracker devices for other clinical or operational applications.

Introduction

The importance of sleep in daily life and advancements in measuring body motion and bio-signals have led to an unprecedented growth in the use of commercial wearable sleep-tracker devices [1]. Wearable sleep-tracker devices offer an attractive alternative for measuring sleep over the gold-standard polysomnography (PSG) because, in addition to their small size, comfort, ease of use, and low cost, they are suitable for prolonged recordings outside of the usual laboratory and clinical environments. As with other wearable devices [2], the key question is whether and for what applications these ubiquitous sleep trackers can be interchangeably used and equally interpreted for assessing sleep vis-à-vis PSG [3].

To date, there are no universally accepted standards for determining the validity of commercial wearable sleep-tracker devices in detecting sleep and wake states [1]. One widely used approach is to simultaneously collect sleep parameters, such as total sleep time (TST), sleep-onset latency (SOL), and sleep efficiency, from a device and PSG and use descriptive statistics of PSG-device differences for these parameters, as well as device sensitivity and specificity to sleep and wake states as compared to PSG [4–6]. By and large, the overall consensus is that sleep trackers have a very high sensitivity (>90%) but a relatively low specificity (~50%) for detecting sleep, because they cannot clearly distinguish motionless wake states from sleep, leading to overestimation of TST and sleep efficiency [1, 6, 7]. A limitation of such an approach for determining device validity is that it does not specify whether a sleep tracker can be interchangeably used for specific clinical or operational applications.

To overcome this limitation, the general approach is to define fixed thresholds of PSG-device differences to a priori set ranges of clinically satisfactory biases in sleep measures [1]. As originally proposed by Werner et al. [8], for a device to be clinically acceptable, the computed Bland and Altman [9] limits of agreement should be narrower (i.e. smaller) than 30 minutes. The Bland and Altman limits of agreement are defined as the range between the PSG-device mean difference (i.e. the bias of the device being investigated relative to PSG) and ± 1.96 standard deviations (SD) of the difference between the two measurement methods, where we would expect 95% of the differences to lie if they were normally distributed. However, the basis for the 30-minute threshold is unclear, other than the authors' own stated clinical experience in diagnosing and evaluating children's sleep difficulties [8]. Unfortunately, as noted by de Zambotti et al. [1], the ubiquitous citing of this 30-minute threshold as a clinically satisfactory

metric for determining the validity of sleep-tracker devices in estimating certain sleep measures (e.g. TST) created a vicious cycle, with each new study citing and relying on previously published studies to justify this arbitrary selection [3, 10–14]. Moreover, the threshold for clinical or operational acceptability of measurement errors should not be a fixed value but rather should depend on the intended use of the sleep measures.

Here, we propose a systematic method to quantitatively determine whether a commercial wearable sleep-tracker device yields operationally acceptable estimates of sleep measures for assessing alertness impairment as part of a fatigue-management system. We assume that PSG-device differences in TST that result in alertness-impairment levels smaller than the within-subject variability are operationally satisfactory. (Henceforth, we term PSG-device differences "sleep-measurement errors.") To develop this new method, we leveraged the well-validated unified model of performance (UMP), which accurately predicts the daily temporal variation of alertness of a group of individuals, as measured by the psychomotor vigilance test (PVT) mean response time (RT), for a given sleep schedule [15]. We used the UMP to simulate thousands of sleep schedules with and without sleep-measurements errors, from which we quantified the errors that led to alertness impairment smaller than ~30 milliseconds, the within-subject variability in terms of PVT mean RT, as estimated by Khitrov et al. [16]. To this end, for a specific device with known sleep-measurement errors as compared to the PSG (i.e. known mean difference in TST and SD of the difference), our method directly provides the fraction of times that a wearable device with these error characteristics is expected to yield sleep-alertness impairment errors smaller than the within-subject variability. We believe that such an approach provides a solid basis for identifying operationally valid wearable sleep-tracker devices, for the purpose of assessing their ability to affect our expectations of alertness levels.

Methods

Assessment of sleep-measurement errors

We reviewed the literature to identify studies that compared the performance of wearable sleep-tracker devices in detecting sleep in healthy adults and adolescents against those measured with the gold-standard PSG. Table 1 provides a brief description of 14 identified studies, which we used as the basis for our analysis, including information regarding age, number of sleep-recording nights, device name and model, TST per night, TST bias (i.e. mean sleep-duration error, with positive values indicating an overestimation of TST by the wearable device compared to PSG), and SOL

Table 1. Summary of the 14 Studies Used as the Basis for Our Analysis, Which Compared Sleep-Duration and Sleep-Onset Estimates From 18 Wearable Sleep-Tracker Devices Against the Gold-Standard Polysomnography

Study	Setting	# Participants (men)	Age,† years	# Nights	Device name and model	Mean TST (SD), minutes	Mean TST bias* (SD), minutes	Mean SOL bias†† (SD), minutes
V1	Home	40 (21)	18–30	1	Fitbit Flex	397 (64)	0 (16)	
				1	Withings Pulse O2		13 (37)	
				1	Misfit Shine		75 (49)	
				1	Basis Health Tracker		−2 (43)	
V2	Home	17 (6)	32.1 (7.4)	1	Fitbit Flex	387 (65)	7 (19)	1 (–)
V3	Home	15**	18–40	1	Withings Pulse O2	433 (71)	34 (34)	−1 (–)
				1	Up Move Jawbone		24 (42)	6 (–)
				1	SenseWear Pro Armband		10 (45)	−3 (–)
V4	Home	25 (15)	24.8 (4.4)	3	Fitbit Charge 2	351 (95)	−12 (32)	−11 (14)
V5	Lab	24 (14)	19–41	1	Fitbit	465 (48)	67 (53)	
V6	Lab	44 (18)	19–61	1	Fitbit Charge 2	379 (47)	9 (24)	−4 (9)
V7	Lab	28 (0)	50.1 (3.9)	1	Jawbone UP	367 (61)	27 (35)	5 (10)
V8	Lab	10**	18.3 (1.0)	3	Fitbit HR Charge	435 (56)	52 (152)	
V9	Lab	53 (25)	15–19	5	Oura ring	447 (58)	−44 (21)	
V10	Lab	34 (12)	28.1 (3.9)	3	Fatigue Science Readiband	416 (48)	13 (55)	−1 (9)
				3	Fitbit Alta HR	425 (33)	3 (22)	−3 (8)
				3	Garmin Fenix 5S	413 (53)	44 (46)	1 (14)
				3	Garmin Vivosmart 3	415 (48)	47 (44)	−1 (6)
V11	Lab	8 (4)	18–35	3	Zulu watch	408 (57)	6 (54)	3 (26)
V12	Lab	6 (3)	23.0 (2.2)	9	WHOOP 2.0	393 (61)	−18 (61)	
V13	Lab	19 (6)	19–64	2	Mi band 2	370 (104)	70 (67)	15 (34)
V14	Lab	41 (28)	14–22	1	Oura ring	392 (59)	1 (22)	0 (7)
	Total:	364			Average:		$\mu_d = 19$ ($\sigma_d = 44$)	$\mu_o = 0$ ($\sigma_o = 14$)

†Values are presented as mean age (standard deviation) or range.
*Mean sleep-duration error, with positive values denoting overestimation by the wearable device.
††Mean sleep-onset error, with positive values denoting overestimation by the wearable device.
**Sex information was not available in the original study. SD, standard deviation. SOL, sleep-onset latency. TST, total sleep time. References: V1 ([17]), V2 ([14]), V3 ([18]), V4 ([19]), V5 ([11]), V6 ([6]), V7 ([12]), V8 ([20]), V9 ([21]), V10 ([7]), V11 ([4]), V12 ([5]), V13 ([22]), V14 ([10]).

bias (i.e. mean sleep-onset error, with positive values denoting a delay in sleep-onset detection by the wearable device). Overall, these studies assessed the performance of 18 unique commercially available wearable sleep-tracker devices (including 22 different conditions), involving a total of 364 distinct individuals who had no history of sleep or neurological disorders. The PSG and wearable devices detected sleep–wake patterns from one to nine nights (mean = 2.2 nights and median = 1.0 nights), either at home using portable PSG devices (Studies V1–V4) or in a laboratory setting (Studies V5–V14). Studies V1 and V10 assessed four wearable devices, Study V3 tested three, and the remaining studies reported results for a single device.

Unified model of performance

Relying on the two-process model of sleep regulation originally proposed by Borbély [23], we previously developed the UMP to quantitatively predict the alertness impairment of a group of individuals spanning the continuum from total sleep deprivation to chronic sleep restriction [24, 25]. Using sleep-schedule history as its input, the UMP predicts alertness impairment $P_0(t) = S(t) + C(t)$ at a future time t, as measured by the PVT mean RT, where S denotes the sleep pressure by the homeostatic process, C represents the circadian process, and κ denotes the circadian amplitude [24]. Tables 2 and 3 summarize the governing equations and parameter values, respectively, used to predict the effect of sleep-schedule history on alertness impairment of a group of individuals as a function of time of day t.

UMP simulations to assess the effect of sleep-measurement errors

To assess the effect of sleep-duration measurement errors (i.e. TST bias) from wearable devices on alertness impairment, we used the UMP to perform two sets of simulations: one using nominal sleep schedules and another using "device sleep schedules." We simulated three nominal sleep schedules with 5, 8, or 9 hours of fixed sleep per night for 30 consecutive days and a fourth schedule with irregular nominal sleep each night, which we randomly sampled from a uniform distribution ranging from 3 to 9 hours of sleep per night for 30 consecutive days. To create the device sleep schedules, we added different random sleep-duration errors ε_d to the nominal schedule for each of the 30 days of the simulation. To obtain ε_d for each day, we used the sleep-duration error statistics in Table 1, assumed that the error was normally distributed with mean μ_d and SD σ_d, and randomly sampled ε_d from this distribution. The simulation of the four nominal sleep schedules (5, 8, or 9 hours and irregular sleep from 3 to 9 hours), where we fixed the wake-up time each morning to 07:00, allowed us to assess whether the effect of ε_d on alertness impairment depended on the length of the daily sleep period. To create a device sleep schedule, we shifted the nominal sleep schedule by adding ε_d to the end of the schedule. For example, for the 8-hour nominal sleep schedule, sleep started at 23:00 and ended at 07:00, while for the corresponding device sleep schedule, sleep also started at 23:00 but ended at 07:00+ε_d (Figure 1A).

For each of the three fixed sleep schedules, we used the UMP to predict the time course of alertness impairment for the nominal sleep schedule and for 100 000 realizations of the device sleep schedule, where for each realization we randomly selected a different ε_d for each of the 30 days of the simulation and added the errors to the nominal schedule. We repeated these simulations by simultaneously adding the sleep-duration error with mean μ_d and SD σ_d and the sleep-onset error with mean μ_o and SD σ_o to the nominal sleep schedules, to create device sleep schedules that accounted for both types of sleep-measurement errors. For the fourth schedule with irregular nominal sleep each day, we performed a similar set of 100 000 realizations, where for each realization we first randomly selected a different nominal sleep duration for each of the 30 days and then, for each day, randomly selected a different ε_d to be added to the selected sleep duration to form the device sleep schedule for that day. Hence, each of the 100 000 realizations consisted of 30 distinct daily pairs of nominal sleep duration and sleep-duration error. For all simulations, we assumed that before the first day of a sleep schedule, individuals slept for 8 hours (from 23:00 to 07:00) and had no accumulated sleep debt [24, 25].

Table 2. Governing Equations of the Unified Model of Performance

Circadian process (C):

$$C(t) = \sum_{i=1}^{5} a_i \sin\left[i\tfrac{2\pi}{\tau}(t+\phi)\right] \quad (1)$$

where a_i, $i = 1, \ldots, 5$, denotes the amplitude of the five harmonics ($a_1 = 0.97$, $a_2 = 0.22$, $a_3 = 0.07$, $a_4 = 0.03$, and $a_5 = 0.001$); τ represents the period of the circadian oscillator (~24 h); and ϕ denotes the circadian phase.

Homeostatic process (S):

$$S(t) = \begin{cases} U - (U - S_0)e^{-t/\tau_w} & \text{during wakefulness} \\ L + (S_0 - L_0)e^{-t/\tau_{LA}} + (L_0 + 2U)\frac{\tau_S}{\tau_{LA} - \tau_S}(e^{-t/\tau_{LA}} - e^{-t/\tau_S}) & \text{during sleep} \end{cases} \quad (2)$$

where U and L denote the upper and lower asymptotes, respectively, of process S; τ_w and τ_s represent the time constants of the sleep pressure during wakefulness and sleep, respectively; and τ_{LA} denotes the time constant of the exponential decay of the effect of sleep history on alertness. [$S(0) = S_0$ and $L(0) = L_0$ correspond to the initial state values for S and L, respectively.]

Lower asymptote (L) of process S:

$$L(t) = Ud(t) \quad (3)$$

where d denotes sleep debt.

Sleep debt (d):

$$d(t) = \begin{cases} 1 - (1 - L_0/U)e^{-t/\tau_{LA}} & \text{during wakefulness} \\ -2 + (2 + L_0/U)e^{-t/\tau_{LA}} & \text{during sleep} \end{cases} \quad (4)$$

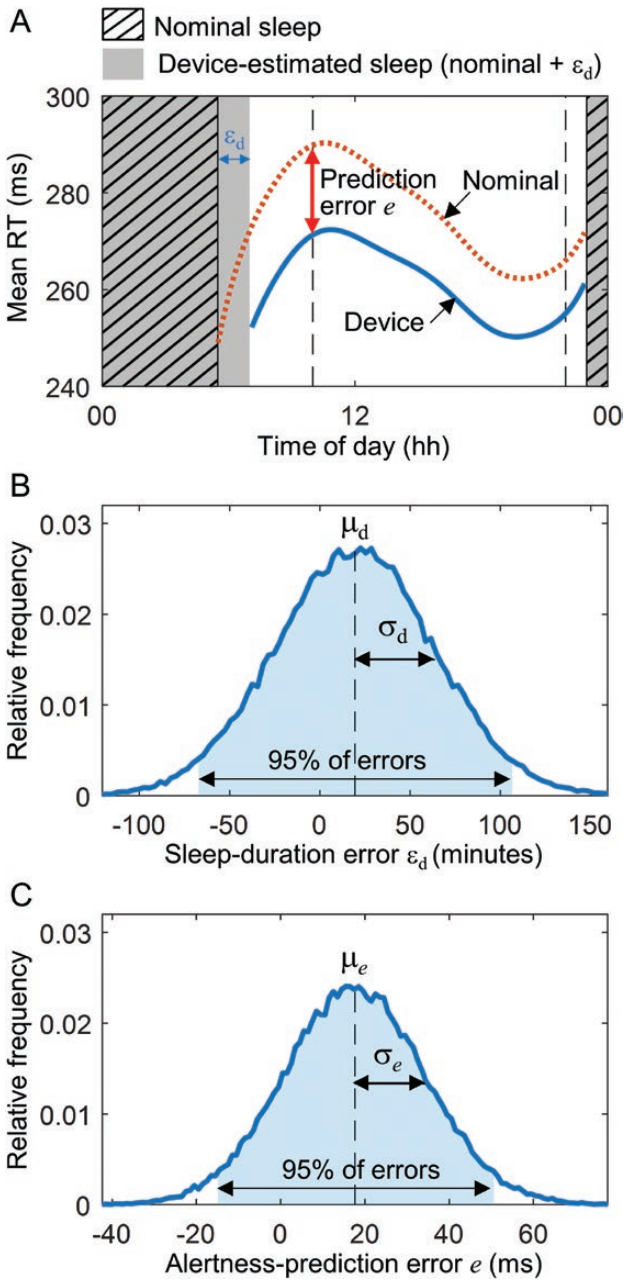


Figure 1. Simulations to assess the effect of sleep-measurement errors on alertness prediction. (A) Alertness predictions for one day as provided by the psychomotor vigilance test mean response time (RT) for a nominal sleep schedule (dotted line) and a device-estimated sleep schedule (solid line). The figure indicates the results for an 8-hour nominal sleep schedule from 23:00 to 07:00. We computed the alertness-prediction error for each day as the largest difference between the predictions for the two schedules between 10:00 and 22:00. Lower values of mean RT correspond to higher alertness levels. (B) Distribution of sleep-duration errors ε_d for one day over 100 000 simulations randomly sampled from a normal distribution with mean μ_d and standard deviation σ_d. (C) Distribution of alertness-prediction errors for one day over 100 000 simulations performed for different device-estimated schedules, with added sleep-duration errors ε_d corresponding to those in panel (B). From the distribution, we estimated the mean alertness-prediction error μ_e and its standard deviation σ_e as well as the interval containing 95% of the errors around μ_e.

In our previous development and validation of the UMP, we used time in bed as input to the model [15], and here, we used TST or TST+ε_d. However, because we are assessing differences in predicted-alertness impairment, where the only difference in the model inputs is the sleep-duration error ε_d, any discrepancy between time in bed and TST cancels out.

Table 3. Parameter Values (Standard Error) Used by the Unified Model of Performance for Predicting the Mean Response Time (RT) Statistics of the Psychomotor Vigilance Test

Parameter	Definition	Value (standard error)
U	Upper asymptote	497 (31) ms
τ_w	Time constant of the sleep pressure during wakefulness	23.0 (3.2) h
τ_s	Time constant of the sleep pressure during sleep	4.0 (1.0) h
S_0	Initial state value for process S	176 (15) ms
κ	Circadian amplitude	75 (7) ms
ϕ	Circadian phase	2.5 (0.2) h
τ_{LA}	Time constant of the exponential decay of the effect of sleep history on alertness	7.0 (2.6) d
L_0	Initial state value for the lower asymptote	140 (14) ms

Metrics to quantify the effect of sleep-measurement errors

To quantify the effect of sleep-measurement errors on alertness impairment, we computed the daily alertness-prediction error e for each of the four sleep schedules by comparing the alertness impairment predicted using the nominal sleep schedule against those predicted based on the device sleep schedule. To this end, for each of the 30 days, we computed the largest daily absolute difference between the predicted mean RT for the nominal schedule versus those of the device schedules for the period between 10:00 and 22:00 during wakefulness for all schedules, for each of the 100 000 realizations. We selected this time window for comparison purposes because it allowed us to compare all simulations over the same time span of the day. In the schedules with added random sleep-duration errors, the end of the sleep period occurred at 07:00+ε_d. Thus, we chose the 10:00 bound to ensure that the end of the sleep period occurred before this bound, and chose the 22:00 bound because the schedule with nominal sleep duration of 9 hours started at 22:00. Figure 1A illustrates the alertness-prediction error e for one realization of a sleep-duration error ε_d for one day. For each of the 30 days of these simulations, we also estimated the daily mean alertness-prediction error μ_e and SD σ_e, as well as the 2.5% and 97.5% quantiles, i.e. an interval containing 95% of the errors, over the 100 000 realizations. Figures 1, B and C illustrate the distribution of the sleep-duration errors and alertness-prediction errors, respectively, over these realizations, for one day.

To provide a benchmark to assess the magnitude of the alertness-prediction errors e and determine whether the device's sleep-measurement errors that resulted in these prediction errors were acceptable, we computed the fraction of absolute errors e smaller than the within-subject variability of alertness under well-rested conditions (~30 milliseconds, for PVT mean RT), as estimated by Khitrov et al. [16]. We estimated the within-subject variability by developing a linear mixed-effects model using data from a cross-over sleep-deprivation study in which the same sleep-satiated individuals performed PVTs between

10:00 and 20:00 during the baseline day of each arm of the study [26]. The mixed-effects model allowed us to directly estimate the within-subject variability, while accounting for between-subject variability and time of day. Hence, for the 5 hours of sleep per night schedule and potentially the irregular sleep schedule, this is a conservative estimate because it is known that the within-subject variability increases with sleep loss [26].

We also investigated the dependence of the alertness-prediction error e for a range of sleep-duration errors. Hence, instead of using a fixed value for μ_d and σ_d in our simulations as discussed above, we assessed μ_e and σ_e as a function of different values of μ_d and σ_d, for each of the three nominal schedules with fixed sleep durations each night. First, we defined a grid of μ_d versus σ_d, with μ_d varying from −80 to 80 minutes in 10-minute intervals and σ_d having three distinct values: 50%, 100%, or 150% of σ_d = 44 minutes in Table 1. Second, for each of the 51 pairs (17 × 3) of μ_d-σ_d combinations in the grid, we repeated the random sampling procedure discussed above using 20 000 realizations (instead of 100 000) to select a different ε_d to be added to the nominal schedule to form the device sleep schedule for each realization, for each of the 30 days of the simulation. Third, for each pair of μ_d-σ_d combinations in the grid, we computed μ_e and σ_e over the 20 000 realizations at day 30, for each of the three schedules. Finally, we used these simulation results to separately build two linear regression models to estimate μ_e and σ_e as a function of μ_d and σ_d, respectively, for the three sleep schedules. Using the linear regression models, we estimated the fraction of alertness-prediction errors that fell within the 30-millisecond within-subject variability threshold for a range of μ_d and σ_d values, and created a contour heat map to estimate this fraction of acceptable errors for a device with given values of sleep-measurement errors μ_d and σ_d.

Results

Estimation of sleep-duration and sleep-onset measurement errors

We observed a large variability in the sleep-duration measurement errors ε_d among the 18 unique wearable sleep-tracker devices in Table 1, the same device in different studies (Fitbit Flex in Study V1 and Study V2; Withings Pulse O2 in V1 and V3; Fitbit Charge 2 in V4 and V6; and Oura ring in V9 and V14), and different models of the same device (e.g. Fitbit Flex in V1, Fitbit Charge 2 in V4, Fitbit HR Charge in V8, and Fitbit Alta HR in V10). For example, while the Misfit Shine device reported in Study V1 yielded an average overestimation of sleep duration of 75 minutes, the Oura ring yielded a 1-minute overestimation in Study V14 and a 44-minute underestimation in Study V9. Similarly, while the Fitbit HR Charge in Study V8 yielded a large overestimation of sleep duration (52 minutes), the Fitbit Alta HR in Study V10 yielded a very small overestimation (3 minutes). In contrast, we observed considerably less variability in sleep-onset errors, with 12 of the 14 reported average values falling between −5 and 6 minutes (exceptions include Studies V4 and V13). To obtain representative descriptive statistics of sleep-duration measurement errors ε_d, we used the 22 conditions in Table 1 to estimate the mean error μ_d = 19 minutes and associated SD σ_d = 44 minutes, after using a Kolmogorov–Smirnov test to confirm that ε_d was normally distributed in 19 out of the 22 study conditions. We decided to equally weight each study condition in the estimation of μ_d and σ_d because of the large discrepancies in sleep-duration error between the same device in different study conditions (e.g. 21 minutes for Withings Pulse O2 in Studies V1/V3 and Fitbit Charge 2 in Studies V4/V6, and 45 minutes for Oura ring in Studies V9/V14) and because we could not characterize the source of the variance, that is, whether it was due to the device or the study condition. Similarly, for sleep-onset errors, we estimated the mean error μ_o = 0 minutes and associated SD σ_o = 14 minutes.

Alertness-prediction error e for a typical sleep-duration measurement error

To quantify the effects of daily sleep-duration measurement errors ε_d on alertness prediction, we computed the daily mean alertness-prediction error μ_e for the 5-, 8-, and 9-hour fixed schedules of sleep per night as well as for the irregular nightly sleep-duration schedule, for 30 consecutive nights. Figure 2 shows the daily values of μ_e (circles) and the intervals around μ_e containing 95% of the errors (shaded areas), for the 100 000 realizations with sleep-duration errors sampled randomly from a normal distribution with μ_d = 19 minutes and σ_d = 44 minutes. For all four schedules, the daily values of μ_e gradually increased over time and reached an asymptote by day 17, changing by < 1 milliseconds in the last 13 days of the simulations. By day 30, μ_e approached ~18 milliseconds for all four schedules, suggesting that an average overestimation of sleep duration of 19 minutes/night resulted in an average overprediction of alertness that was smaller than the within-subject variability of ± 30 milliseconds, illustrated by the dashed horizontal lines in Figure 2.

To investigate the effect of sleep-onset errors on alertness impairment, we simultaneously added the sleep-duration error (μ_d = 19 minutes and σ_d = 44 minutes) and the sleep-onset error (μ_o = 0 minutes and σ_o = 14 minutes) to each of the three fixed nominal sleep schedules and repeated the 100 000 simulations. Because at the end of 30 days of simulations the largest difference in μ_e and σ_e between these simulations and those where we only considered the sleep-duration error was ≤ 3 milliseconds, we restricted our analysis to sleep-duration errors.

Figure 2 also shows that the alertness-prediction error μ_e was largely insensitive to the nominal sleep schedule, with mean errors and associated 95% intervals at day 30 differing by less than 2 and 7 milliseconds, respectively, between the four schedules. These are illustrated by the distribution of alertness-prediction errors at day 30 on the right-hand side of the plots. The largest difference occurred between the 5- and 8-hour schedules, with μ_e of 16 versus 18 milliseconds and 95% error intervals of 59 versus 66 milliseconds, respectively.

Although the mean alertness-prediction error μ_e remained at ~18 milliseconds, the 95% intervals in Figure 2 (shaded areas) indicated that a fraction of the errors exceeded the within-subject variability of 30 milliseconds. As in the case of the mean error, the fraction of errors that exceeded the 30-millisecond threshold gradually reached an asymptote at ~23% at day 30, for each of the four schedules. Therefore, on any given day, there was a ~77% probability that the device sleep-duration error would lead to alertness errors smaller than the within-subject variability.

Alertness-prediction error e for a range of sleep-duration measurement errors

Because of the large variability in the statistics of sleep-duration errors among the devices reported in Table 1, we extended our analysis and considered cases with μ_d ranging from −80 to 80 minutes (in 10-minute intervals) and σ_d set to 22, 44, or 66 minutes. For each of the 51 (17 × 3) pairs of μ_d-σ_d combinations, we performed 20 000 simulations for each of the three fixed sleep schedules and estimated μ_e and σ_e at the end of the 30 days of simulations. Figures 3, A and B show the mean alertness-prediction

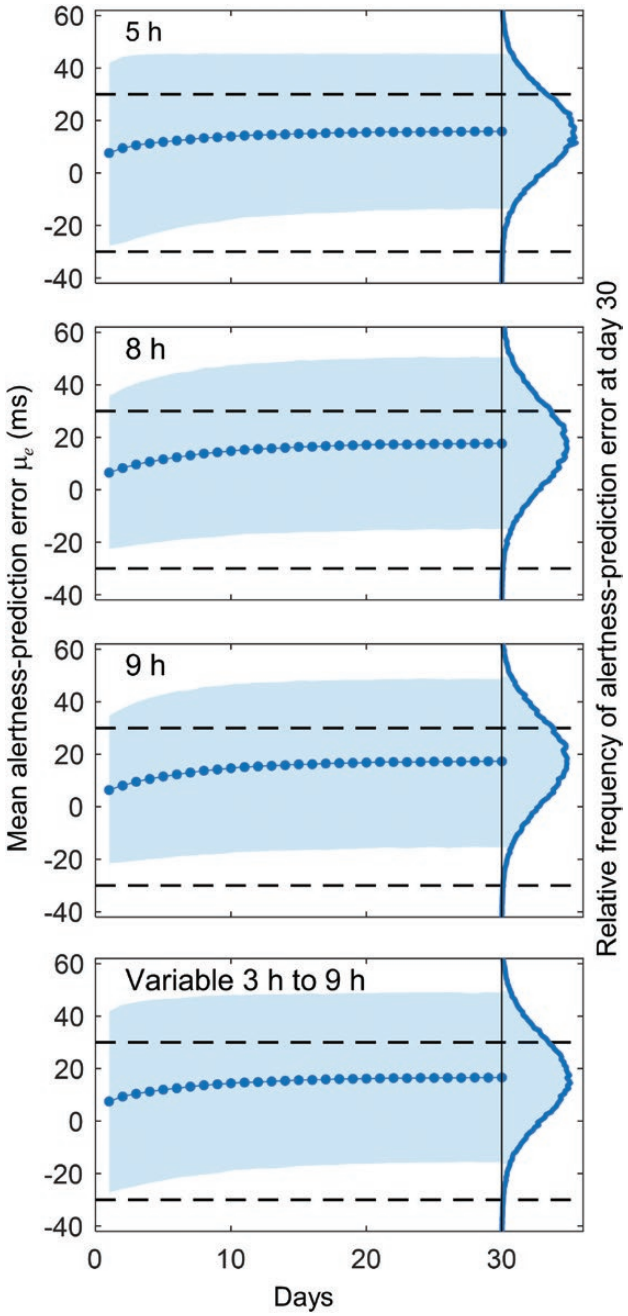


Figure 2. Daily alertness-prediction errors resulting from sleep-duration measurement errors of a wearable sleep-tracker device for 30 consecutive days. The plots show the mean alertness-prediction error μ_e (circles) and the intervals containing 95% (shaded areas) of the daily alertness-prediction errors for nominal sleep schedules of 5, 8, or 9 hours of sleep per night, or for daily irregular sleep schedules (bottom panel) randomly sampled from a uniform distribution ranging from 3 to 9 hours of sleep per night. For each panel, we performed 100 000 simulations with different sleep-duration errors each day randomly sampled from a normal distribution with mean $\mu_d = 19$ minutes and standard deviation $\sigma_d = 44$ minutes, derived from the 22 study conditions listed in Table 1. The solid lines on the right-hand side of each panel represent the distribution of the alertness-prediction errors at day 30 of the simulation. The horizontal dashed lines indicate the ± 30 millisecond threshold of the within-subject variability for an average individual sleeping 8 hours per night ([16]).

error μ_e and its SD σ_e (dots), respectively, for the 153 simulations (51 times the three sleep schedules) as a function of μ_d and σ_d. Note that μ_e changed linearly with the mean sleep-duration error μ_d (Figure 3A), where the three different values of σ_d and sleep schedules (illustrated by the multiple dots for a given μ_d) had a minimal effect on μ_e. In fact, the linear regression model $\mu_e = 0.85\ \mu_d$ captured 99% of the variance of μ_e (Figure 3A, solid line), indicating that, in general, the mean alertness-prediction error changed by 0.85 milliseconds for each minute of sleep-duration error. Similarly, the SD of the alertness-prediction error σ_e depended mainly on the SD of the sleep-duration error σ_d (Figure 3B, dots), with the linear regression model $\sigma_e = 0.35\ \sigma_d$ (Figure 3B, solid line) capturing 95% of the variance of σ_e and indicating that, in general, the SD of the alertness-prediction error changed by 0.35 milliseconds for each minute of the SD of the sleep-duration error. Thus, using these models, we can estimate the mean alertness-prediction error and its SD for a given sleep-tracker device. For example, we estimated that the WHOOP 2.0 device (Table 1, V12), with $\mu_d = -18$ minutes and $\sigma_d = 61$ minutes, would yield an absolute mean alertness-prediction error $\mu_e = 15$ milliseconds and an associated SD $\sigma_e = 21$ milliseconds.

From a practical standpoint, it is also useful to estimate the likelihood that sleep-duration measurement errors from a wearable device would result in acceptable errors in alertness predictions, i.e. deviations that would not exceed the within-subject variability threshold of 30 milliseconds. Thus, using the linear regression models, we estimated μ_e and σ_e for a range of values of μ_d (0 to 80 minutes) and σ_d (0 to 160 minutes) and, for each μ_d-σ_d pair combination, computed the fraction of alertness-prediction errors $\mu_e < 30$ milliseconds. Figure 4 shows the contour lines of a heat map, which indicates the values of μ_d and σ_d that would result in a given fraction of $\mu_e < 30$ milliseconds, after using the wearable device for at least 20 consecutive days. Accordingly, we can use this contour heat map to estimate the likelihood that a given sleep-tracker device would lead to an acceptable alertness-prediction error. For example, for the Fitbit Charge 2, with $\mu_d = 9$ minutes and $\sigma_d = 24$ minutes (Study V6 in Table 1), we would expect that > 90% of the alertness-prediction errors would be < 30 milliseconds (Figure 4, star), whereas for the WHOOP 2.0 (Study V12) and the Mi band 2 (Study V13), respectively, >70% and > 10% of the alertness-prediction errors would be < 30 milliseconds.

Discussion

A large body of work has investigated the validity of commercially available wearable sleep-tracker devices as a low-cost and more practical alternative to measure sleep parameters than the gold-standard PSG. While these wearable devices are inadequate alternatives for capturing nuanced sleep patterns during rapid-eye-movement and non-rapid-eye-movement sleep [4, 7, 17, 19], they offer the capability to measure more basic sleep parameters, such as TST and SOL [4, 17, 19, 22]. Nevertheless, their high sensitivity for sleep detection comes with the cost of a relatively low specificity, as these devices cannot always accurately identify motionless awake periods, leading to overestimation of sleep duration and errors in detecting sleep onset [4–7, 22]. Here, we sought to provide an approach to determine the extent to which such measurement errors are operationally acceptable by assessing how they affect estimates of fatigue and alertness impairment. We believe that such an approach is complementary to summary statistics [27] and offers the means to assess the practical utility of wearable sleep-tracker devices. Our approach also supports the notion that the required validity of wearable-device measurements is highly dependent on the effect that their inaccuracies may have on the desired endpoint [2], alertness impairment in our case.

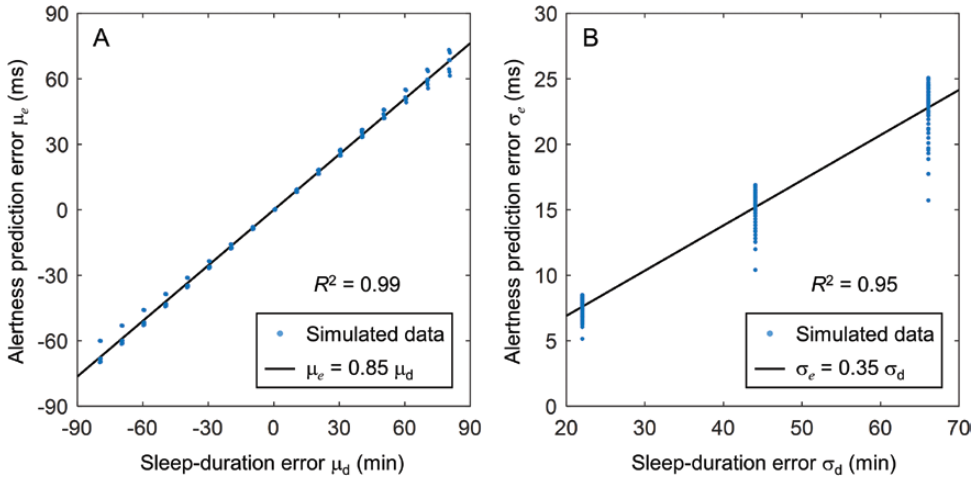


Figure 3. Mean alertness-prediction error μ_e (A) and its standard deviation σ_e (B) as a function of sleep-duration measurement errors μ_d and σ_d, respectively. Each plot shows 153 dots (17 μ_d values × 3 σ_d values × 3 nominal sleep schedules) corresponding to the values at day 30 estimated from 20 000 simulations. The solid lines correspond to the best-fit linear model with zero intercept. R^2, coefficient of determination.

To this end, we first characterized measurement errors in sleep duration (i.e. TST) and sleep onset (i.e. SOL) by performing a meta-analysis based on 14 studies, involving 18 unique commercially available wearable sleep-tracker devices in 22 different conditions, where we compared the reported measurement errors between these devices and PSG. Then, to gauge the effect of sleep-duration and sleep-onset errors on estimates of alertness impairment, we used the well-validated UMP [15] to perform computer simulations and compared alertness predictions at the end of 30 consecutive days of nominal sleep schedules with 5, 8, or 9 hours of sleep per night or irregular sleep-duration each night against those with added errors representing the corresponding device sleep schedules. Based on the simulation results, we constructed linear regression models to estimate the expected alertness-prediction error for a specific wearable device. Finally, we used the within-subject variability in alertness impairment under well-rested conditions, as measured by the PVT mean RT (i.e. 30 milliseconds) [16], as a conservative benchmark [26] to assess device errors that could be tolerated before inducing a substantial error in the estimation of fatigue. Our working hypothesis is that devices are acceptable when their sleep-measurement errors lead to daily errors in alertness estimates of less than 30 milliseconds.

The meta-analysis indicated that the sleep-duration errors were more pronounced than the sleep-onset errors, with mean duration errors ranging from –44 to 70 minutes and mean onset errors of –11 to 15 minutes (Table 1). Thus, when we used the UMP to perform simulations with only the sleep-duration errors or with both types of errors (using the representative statistics for the 22 study conditions in Table 1, last row), we observed negligible differences (≤3 milliseconds) in alertness-prediction errors after 20 consecutive days of daily measurement errors, in each of the three simulated schedules with fixed sleep durations (Figure 2, top three panels). Consequently, we focused our analysis solely on the effects of sleep-duration errors.

For the average sleep overestimation (μ_d = 19 minutes and associated σ_d = 44 minutes, in Table 1), the simulation results showed that the mean alertness-prediction error μ_e remained below the 30-millisecond threshold for each of the four conditions in Figure 2. However, for this level of acceptable overestimation, the probability of achieving a prediction error > 30 milliseconds on any given day was ~23%. For example, after

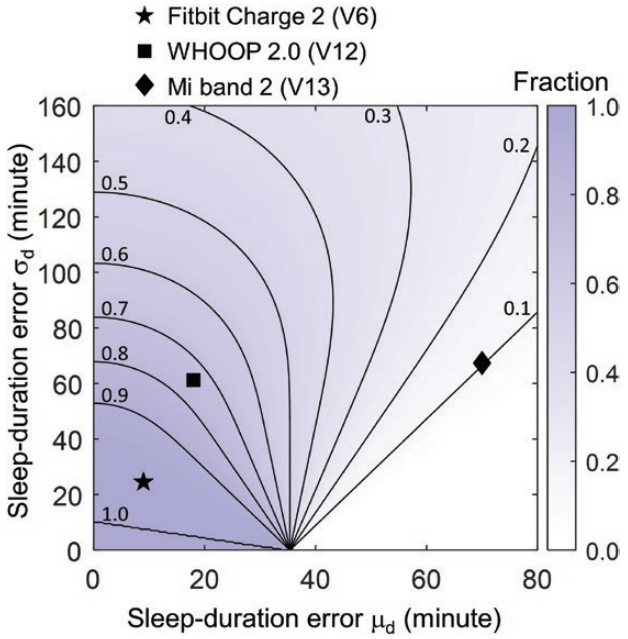


Figure 4. Fraction of alertness-prediction errors < 30 milliseconds as a function of sleep-duration errors μ_d and σ_d. The lines indicate the values of μ_d and σ_d that result in the corresponding fraction of alertness-prediction errors < 30 milliseconds. The plot also shows the values for three devices listed in Table 1. If a wearable device has a negative μ_d, use its absolute value to read the plot (see WHOOP 2.0, square, which has a μ_d = –18 minutes).

20 days of daily sleep overestimations, alertness-prediction errors > 30 milliseconds would occur on 1 of every 4 days, with a mean error of 39 milliseconds and SD of 12 milliseconds on those days. Hence, we conclude that for this level of device-measurement errors, the effects on expected alertness levels are negligible.

Reid and Dawson showed that a wearable sleep-tracker device has similar sleep-duration error characteristics for both diurnal and nocturnal sleep [28]. Thus, using the average sleep overestimation results (μ_d = 19 minutes and σ_d = 44 minutes) in Table 1, we repeated our simulations for the case of a more operationally relevant scenario, such as nightshift work with restricted diurnal

sleep (4 hours of daily sleep from 08:00 to 12:00) [29], which we previously used to validate the UMP predictions [15, 30]. Our simulations yielded a mean alertness-prediction error $\mu_e \leq 15$ milliseconds for each of the 30 simulated days (Supplementary Figure S1), suggesting that our approach is also applicable to nightshift work with diurnal sleep.

Analyses of our simulation results suggest that the alertness-prediction error e is largely insensitive to the nominal sleep schedule and that its distribution could be derived from the distribution of the sleep-duration error ε_d (Figure 3). In fact, this conclusion holds even when sleep is restricted to 3 hours per night (mean alertness-prediction error of 13 milliseconds, Supplementary Figure S2). This conclusion stems from the fact that, although the time course of alertness varies non-linearly throughout the day, the effect of the circadian process cancels out because it is the same for both the nominal and the device sleep schedules, and the effect of the homeostatic process is mainly linear for the range of sleep-duration errors of the sleep-tracker devices.

Although the results based on the average sleep overestimation provided a general idea of the effects of daily use of sleep-tracker devices on predicted alertness, it would be useful to have the ability to assess the effect of a particular wearable device, with specific measurement errors μ_d and σ_d. To this end, we performed thousands of simulations for a range of μ_d-σ_d combinations covering the devices in Table 1 and used these data to construct linear regression models that estimated the expected alertness-prediction errors μ_e and σ_e as a function of specific values of μ_d and σ_d, respectively (Figure 3). For example, for Fitbit Charge 2 (Table 1, V6), with normally distributed errors $\mu_d = 9$ minutes and $\sigma_d = 24$ minutes, the models yielded a mean prediction error of $\mu_e = 8$ milliseconds and SD $\sigma_e = 8$ milliseconds, which are comparable to the estimates obtained using the UMP ($\mu_e = 8$ milliseconds and $\sigma_e = 9$ milliseconds). In fact, the estimates obtained using the linear regression models were very similar to those obtained using the UMP for the 22 study conditions in Table 1, with an $R^2 \geq 0.96$.

We also used the linear regression models to estimate the fraction of alertness-prediction errors < 30 milliseconds for a given wearable device. The heat contour map in Figure 4 provides the means to graphically obtain this fraction as a function of a device's μ_d and σ_d. For example, the Fitbit Charge 2 (V6) results in > 90% of the prediction errors lower than the within-subject variability (Figure 4, star), suggesting that, on any given day, the probability that a random sleep-duration error would lead to an alertness-level error greater than the within-subject variability is relatively small (<10%). Overall, nine devices in eight studies (Fitbit Flex, V1 and V2; Withings Pulse O2, V1; Basis Health Tracker, V1; SenseWear Pro Armband, V3; Fitbit Charge 2, V4 and V6; Fatigue Science Readiband, V10; Fitbit Alta HR, V10; Zulu watch, V11; and Oura ring, V14) resulted in > 80% of the prediction errors below 30 milliseconds, indicating that these sleep trackers could be used to measure TST as part of a fatigue-management system. Thus, the heat map provides the means to transform sleep-duration error characteristic of any given device into a quantitative metric with practical ramifications for fatigue management.

Our work has limitations. First, given the considerable variability in the reported sleep-duration measurement errors, it is likely that certain wearable sleep-tracker devices may have error characteristics different from those used in our simulations. Nevertheless, we expect that the overall results reported here will remain valid and serve as a general guideline for various devices. Second, we focused our primary analysis on nocturnal sleep with fixed and irregular sleep schedules ≥3 hours per night, and a secondary analysis of 4 hours of daily diurnal sleep associated with a simulated nightshift work scenario. Thus, we do not know the extent to which our results would hold for short daytime naps, which may be more challenging to measure with wearable devices, as well as sleep under circadian misalignment. Third, the studies used to develop the UMP and assess the sleep-tracker devices were based on groups of individuals with no history of sleep or neurological disorders. Therefore, we do not know whether the main conclusions reported herein would be applicable for groups of individuals with sleep-related disorders. Fourth, our numerical analysis is valid for an "average" individual, as in our simulations we used the group-average feature of the UMP, which provides population-averaged predictions of alertness impairment and does not take into account an individual's resilience or vulnerability to sleep loss. Fifth, we based the potential operational applicability of our results for fatigue management on a limited number of simulations, which would need to be broadened and further validated for other clinical and operational applications. Finally, the UMP predicts alertness as measured by the PVT. Therefore, the extent to which its predictions can be generalized to other aspects of neurobehavioral performance remains unknown.

In summary, the results of our numerical analysis indicate that while commercially available wearable sleep-tracker devices, on average, overestimate sleep duration by 19 minutes (SD = 44 minutes), they do provide an acceptable low-cost alternative to measuring sleep duration for assessing fatigue. We found that in nearly 80% of the time, the resulting mean RT predicted-alertness error would be smaller than the within-subject variability of 30 milliseconds. We also provided the means to use the sleep-measurement error characteristic of a particular sleep-tracker device to determine whether it is operationally acceptable for fatigue management. We conclude that the sleep-duration errors observed in half of the sleep-tracker devices reviewed here are acceptable when the objective is to assess discrepancies in alertness level for fatigue management.

Supplementary Material

Supplementary material is available at *SLEEP* online.

Funding

This work was sponsored by the Military Operational Medicine Program Area Directorate of the U.S. Army Medical Research and Development Command (USAMRDC), Fort Detrick, MD. The Henry M. Jackson Foundation was supported by the USAMRDC under Contract No. W81XWH20C0031.

Disclosure Statements

This was not an industry-supported study. JR and FGVL receive royalties for the licensing of the *2B-Alert* technology to Distritec. Nonfinancial disclosure: The authors indicate no other conflicts of interest. The opinions and assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the U.S. Army, the U.S. Department of Defense, or The Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc. This paper has been approved for public release with unlimited distribution.

Author Contributions

JR designed the study. NVP and FGVL performed the computations. JR, NVP, and FGVL analyzed the results and wrote the manuscript. All authors have reviewed the manuscript and approved the submitted version.

Data Availability

All data will be made available following a written request to the corresponding author, along with a summary of the planned research.

References

1. de Zambotti M, Cellini N, Goldstone A, Colrain IM, Baker FC. Wearable sleep technology in clinical and research settings. *Med Sci Sports Exerc*. 2019;**51**(7):1538–1557. doi: 10.1249/MSS.0000000000001947
2. Laxminarayan S, Hornby S, Belval LN, *et al*. Prospective validation of *2B-Cool*: Integrating wearables and individualized predictive analytics to reduce heat injuries. *Med Sci Sports Exerc*. 2023;**55**(4):751–764. doi: 10.1249/MSS.0000000000003093
3. Meltzer LJ, Walsh CM, Traylor J, Westin AM. Direct comparison of two new actigraphs and polysomnography in children and adolescents. *Sleep*. 2012;**35**(1):159–166. doi: 10.5665/sleep.1608
4. Devine JK, Chinoy ED, Markwald RR, Schwartz LP, Hursh SR. Validation of Zulu watch against polysomnography and actigraphy for on-wrist sleep-wake determination and sleep-depth estimation. *Sensors*. 2020;**21**(1):76. doi: 10.3390/s21010076
5. 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
6. 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. doi: 10.1080/07420528.2017.1413578
7. 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
8. Werner H, Molinari L, Guyer C, Jenni OG. Agreement rates between actigraphy, diary, and questionnaire for children's sleep patterns. *Arch Pediatr Adolesc Med*. 2008;**162**(4):350–358. doi: 10.1001/archpedi.162.4.350
9. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;**1**(8476):307–310.
10. de Zambotti M, Rosas L, Colrain IM, Baker FC. The sleep of the ring: Comparison of the OURA sleep tracker against polysomnography. *Behav Sleep Med*. 2019;**17**(2):124–136. doi: 10.1080/15402002.2017.1300587
11. 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
12. de Zambotti M, Claudatos S, Inkelis S, Colrain IM, Baker FC. Evaluation of a consumer fitness-tracking device to assess sleep in adults. *Chronobiol Int*. 2015;**32**(7):1024–1028. doi: 10.3109/07420528.2015.1054395
13. Toon E, Davey MJ, Hollis SL, Nixon GM, Horne RS, Biggs SN. Comparison of commercial wrist-based and smartphone accelerometers, actigraphy, and PSG in a clinical cohort of children and adolescents. *J Clin Sleep Med*. 2016;**12**(3):343–350. doi: 10.5664/jcsm.5580
14. 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. doi: 10.1016/j.jpsychores.2017.03.009
15. Priezjev NV, Vital-Lopez FG, Reifman J. Assessment of the unified model of performance: accuracy of group-average and individualised alertness predictions. *J Sleep Res*. 2023;**32**(2):e13626. doi: 10.1111/jsr.13626
16. Khitrov MY, Laxminarayan S, Thorsley D, *et al*. PC-PVT: A platform for psychomotor vigilance task testing, analysis, and prediction. *Behav Res Methods*. 2014;**46**(1):140–147. doi: 10.3758/s13428-013-0339-9
17. Mantua J, Gravel N, Spencer RM. Reliability of sleep measures from four personal health monitoring devices compared to research-based actigraphy and polysomnography. *Sensors*. 2016;**16**(5):646. doi: 10.3390/s16050646
18. Gruwez A, Libert W, Ameye L, Bruyneel M. Reliability of commercially available sleep and activity trackers with manual switch-to-sleep mode activation in free-living healthy individuals. *Int J Med Inform*. 2017;**102**:87–92. doi: 10.1016/j.ijmedinf.2017.03.008
19. Liang Z, Chapa Martell MA. Validity of consumer activity wristbands and wearable EEG for measuring overall sleep parameters and sleep structure in free-living conditions. *J Healthc Inform Res*. 2018;**2**:152–178. doi: 10.1007/s41666-018-0013-1
20. 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
21. Chee N, Ghorbani S, Golkashani HA, Leong RLF, Ong JL, Chee MWL. Multi-night validation of a sleep tracking ring in adolescents compared with a research actigraph and polysomnography. *Nat Sci Sleep*. 2021;**13**:177–190. doi: 10.2147/NSS.S286070
22. Ameen MS, Cheung LM, Hauser T, Hahn MA, Schabus M. About the accuracy and problems of consumer devices in the assessment of sleep. *Sensors*. 2019;**19**:4160. doi: 10.3390/s19194160
23. Borbély AA. A two process model of sleep regulation. *Hum Neurobiol*. 1982;**1**(3):195–204.
24. Rajdev P, Thorsley D, Rajaraman S, *et al*. A unified mathematical model to quantify performance impairment for both chronic sleep restriction and total sleep deprivation. *J Theor Biol*. 2013;**331**:66–77. doi: 10.1016/j.jtbi.2013.04.013
25. Ramakrishnan S, Wesensten NJ, Balkin TJ, Reifman J. A unified model of performance: validation of its predictions across different sleep/wake schedules. *Sleep*. 2016;**39**(1):249–262. doi: 10.5665/sleep.5358
26. Rupp TL, Wesensten NJ, Balkin TJ. Trait-like vulnerability to total and partial sleep loss. *Sleep*. 2012;**35**(8):1163–1172. doi: 10.5665/sleep.2010
27. Nguyen QNT, Le T, Huynh QBT, Setty A, Vo TV, Le TQ. Validation framework for sleep stage scoring in wearable sleep trackers and monitors with polysomnography ground truth. *Clocks Sleep*. 2021;**3**(2):274–288. doi: 10.3390/clockssleep3020017
28. Reid K, Dawson D. Correlation between wrist activity monitor and electrophysiological measures of sleep in a simulated shiftwork environment for younger and older subjects. *Sleep*. 1999;**22**(3):378–385. doi: 10.1093/sleep/22.3.378
29. Wesensten NJ, Reichardt RM, Balkin TJ. Ampakine (CX717) effects on performance and alertness during simulated night shift work. *Aviat Space Environ Med*. 2007;**78**(10):937–943. doi: 10.3357/asem.2055.2007
30. Vital-Lopez FG, Doty TJ, Reifman J. Optimal sleep and work schedules to maximize alertness. *Sleep*. 2021;**44**(11):zsab144. doi: 10.1093/sleep/zsab144