Original Research Article

DIGITAL HEALTH
Volume 10: 1–12
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20552076241277150
journals.sagepub.com/home/dhj

Evaluation of the Garmin Vivofit 4 for assessing sleep in youth experiencing sleep disturbances

Paul R Hibbing[1,2], Madison Pilla[2], Lauryn Birmingham[3], Aniya Byrd[3], Tumusifu Ndagijimana[3], Sara Sadeghi[3], Nedra Seigfreid[3], Danielle Farr[3], Baha Al-Shawwa[4,5], David G Ingram[4,5] and Jordan A Carlson[2,5]

Abstract

Objective: Wearable monitors are increasingly used to assess sleep. However, validity is unknown for certain monitors and populations. We tested the Garmin Vivofit 4 in a pediatric clinical sample.

Methods: Participants ($n = 25$) wore the monitor on their nondominant wrist during an overnight polysomnogram. Garmin and polysomnography were compared using 95% equivalence testing, mean absolute error, and Bland-Altman analysis.

Results: On average (mean ± SD), the Garmin predicted later sleep onset (by 0.84 ± 1.60 hours) and earlier sleep offset (by 0.34 ± 0.70 hours) than polysomnography. The resulting difference for total sleep time was -0.55 ± 1.21 hours. Sleep onset latency was higher for Garmin than polysomnography (77.4 ± 100.9 and 22.8 ± 20.0 minutes, respectively), while wake after sleep onset was lower (5.2 ± 9.3 and 43.2 ± 37.9 minutes, respectively). Garmin sleep efficiency averaged $3.3\% \pm 13.8\%$ lower than polysomnography. Minutes in light sleep and deep sleep (the latter including rapid eye movement) were within $\pm 3.3\%$ of polysomnography (both SDs = 14.9%). No Garmin means were significantly equivalent with polysomnography (adjusted $p > 0.99$). Mean absolute errors were 0.47 to 0.95 hours for time-based variables (sleep onset, offset, and latency, plus total sleep time and wake after sleep onset), and 8.9% to 21.2% for percentage-based variables (sleep efficiency and sleep staging). Bland-Altman analysis showed systematic bias for wake after sleep onset, but not other variables.

Conclusions: The Vivofit 4 showed consistently poor individual-level validity, while group-level validity was better for some variables (total sleep time and sleep efficiency) than others.

Keywords

Actigraphy, wearable electronic devices, methods

Submission date: 21 November 2023; Acceptance date: 5 August 2024

Introduction

Sleep is a crucial health behavior for children and adolescents, promoting brain health, weight management, and a host of other important functions.[1–5] It is therefore alarming that sleep disorders affect an estimated 25% to 50% of youth,[6,7] while only 3.7% of cases are diagnosed.[8] This creates a need to monitor sleep in a variety of clinical, scientific, and personal settings, but it is difficult to do so because of measurement challenges.

Polysomnography (PSG) is the gold standard method for measuring sleep, involving multimodal monitoring of

[1]Department of Kinesiology and Nutrition, University of Illinois Chicago, Chicago, IL, USA
[2]Center for Children's Healthy Lifestyles & Nutrition, Children's Mercy Kansas City, Kansas City, MO, USA
[3]STAR 2.0 Program, Children's Mercy Kansas City, Kansas City, MO, USA
[4]Department Pediatrics, Division of Pulmonary and Sleep Medicine, Children's Mercy Kansas City, Kansas City, MO, USA
[5]Department of Pediatrics, University of Missouri Kansas City, Kansas City, MO, USA

Corresponding author:
Paul R. Hibbing, Department of Kinesiology and Nutrition, University of Illinois Chicago, 1919 W. Taylor St, AHSB Rm 650, MC 517, Chicago, IL 60612, USA.
Email: phibbing@uic.edu

electroencephalography, eye and limb movements, and cardiovascular and respiratory parameters. However, there are notable limitations with PSG, including its cost, invasiveness, and labor intensiveness.[9] This has led to a surge of interest in wearable sleep monitors (especially wrist-worn devices) that may increase the feasibility and convenience of measuring sleep in natural environments over long periods of time.[10] Accordingly, it has become a high priority to test the accuracy of different monitors by comparison against PSG.[11] Consumer-grade sleep monitors have received a wealth of attention along this line,[12] with studies testing a range of monitors and showing a consistent trend toward low specificity (e.g. underestimation of wake time, leading to overestimation of sleep efficiency), along with variable trends for predicting sleep stages.[13–20] Despite these persistent difficulties, new sleep monitors are continually emerging, whose differing characteristics (e.g. unique sensor configurations and approaches to predicting sleep) may lead to improvements, thus creating a need for continual testing on a monitor- and population-specific basis.[21–23]

Currently, there is a testing gap for monitors sold by Garmin Ltd (Olathe, KS, USA), which is a major player in the wearable market and offers several monitors with sleep tracking capabilities.[24] Prior studies have tested three Garmin models against PSG in healthy adults (Fenix 5S, Vivosmart 3, and Vivosmart 4),[13–15] with results generally showing better performance in some areas (e.g. assessment of sleep onset latency) than others (e.g. assessment of sleep stages). These findings suggest Garmin monitors may have utility in certain situations, but there are evidence gaps related to validity of the most affordable and accessible Garmin monitor (the Vivofit 4, which takes a simplified approach to sleep monitoring) and validity in youth with sleep disorders. Such youth are an easily-overlooked and crucial group to study given the high prevalence and low diagnosis rates mentioned before. Therefore, the purpose of this study was to test the validity of the Garmin Vivofit 4 relative to PSG in a clinical sample of youth experiencing sleep disturbances. Based on the simplified design of the Vivofit 4, along with prior findings for the Garmin devices described above,[13–15] we hypothesized the Vivofit 4 would be equivalent to PSG for total sleep time (TST) and related variables, but not for wake after sleep onset (WASO) or sleep staging.

Methods

Recruitment and participants

Recruitment was conducted through the sleep clinic at Children's Mercy Kansas City. Eligible participants were 5- to 17-year-old patients for whom a sleep medicine specialist had ordered an overnight in-lab PSG due to concern for physiologic disruptors of sleep such as sleep apnea or limb movement disorders. A research team member invited these patients to participate in the research study. Patients were ineligible to participate if they had major craniofacial abnormalities or a prior diagnosis with Down syndrome, Prader-Willi syndrome, or other intellectual disability. Ethical approval for the study was obtained from the Children's Mercy Kansas City Institutional Review Board. Prior to beginning the study, participants gave signed informed assent to participate, and parents gave signed informed consent.

Protocol

The study followed a cross-sectional, single-night design that was consistent with clinical standards of care and comparable with other validation studies for wearable sleep-tracking devices.[15–17] Participants reported to the sleep clinic laboratory near bedtime, then underwent standard preparation for an overnight in-lab PSG with the help of a technologist. In addition to the standard procedures, the technologist placed a Garmin Vivofit 4 on the participant's nondominant wrist and recorded the time of attachment (which was typically about 30 minutes before lights out). PSG recording continued throughout the night until the participant was awakened, typically near 06:00. At that point, the technologist removed the PSG sensors and Garmin monitor, recording the time at which the monitor was removed. Across all participants, only three Vivofit 4 devices were used for testing, thereby limiting the potential influence of interdevice reliability on the results of the study.

Equipment

Polysomnography. Gold standard sleep data were collected via monitored in-lab PSG, consistent with longstanding conventions and recommendations.[25–27] The equipment and procedures were compliant with requirements of the American Academy of Sleep Medicine at the time of data collection,[28] including the use of a standard sensor array for data collection. Data scoring and cleaning were done in 30-second epochs by a certified PSG technologist. All procedures for data collection, reporting, and storage were managed using the BWAnalysis and BWCenter platforms (Neurovirtual USA, Fort Lauderdale, FL), which ultimately generated summary outputs that were integrated into the electronic medical record and extracted for the present analysis.

Garmin Vivofit 4. The Garmin Vivofit 4 is a commercially available monitor equipped with an accelerometer sensor. It is designed as an entry level monitor and thus does not collect photoplethysmography or heart rate data, distinguishing it from other Garmin monitors (e.g. the Vivosmart). A tradeoff of this design is that the Vivofit 4 can only predict "light" and "deep" sleep, as opposed to classifying rapid eye movement (REM) as a separate level of sleep (see https://support.garmin.com/en-US/?faq=mBRMf4ks7XqtsbI8J6). However, an advantage is that the approach parallels what is common in traditional sleep actigraphy (e.g. in the well-known accelerometer algorithms of Cole et al.[29] and Sadeh et al.[30]). Acceleration data from the Vivofit 4 are transmitted via Bluetooth® to the Garmin Connect application on a mobile device, where a proprietary algorithm then predicts sleep from the patterns of wrist movement. (The algorithm's architecture is unknown, and thus there is uncertain comparability to existing accelerometry algorithms.) For the present study, sleep estimates were accessed through the Garmin application programming interface (API) platform, which provided a convenient way to manage data from Garmin Connect. Table 1 lists the variables that were available or calculable from the API platform. Notably, epoch-level data were not made available by Garmin, meaning it was only possible to test night-level aggregate estimates.

Data processing and statistical analysis

Procedures and approach. Participant characteristics and diagnostic data were extracted from the scored PSG summary reports stored in the electronic medical record. Weight status was determined using body mass index (BMI) growth charts from the Centers for Disease Control and Prevention,[31] with severe obesity defined as ≥120% of the 95th percentile or BMI ≥ 35.[32] A physician reviewed all PSG reports and coded diagnoses as none, insomnia (i.e. chronic insomnia or behaviorally induced insufficient sleep syndrome), sleep-related breathing disorder (i.e. obstructive or central sleep apnea, hypoxemia, and hypoventilation), sleep-related movement disorder (i.e. restless leg syndrome, periodic limb movement disorder, and restless sleep disorder), or multiple (i.e. two or more of the above).

The PSG reports were also used to extract criterion sleep data. The focus was on the variables in Table 1, which were chosen to match what was available or calculable from the Garmin API platform. In the case of sleep stages, the PSG reports contained more detailed information than what was predicted by the Garmin monitor. Specifically, the PSG reports listed time in REM and three non-REM stages (N1, N2, and N3), while the Garmin monitor predicted only time in light and deep sleep, as noted previously. Therefore, the PSG values had to be recategorized as light and deep to match the output from the Garmin monitor. For the non-REM stages, this was done by coding N1 and N2 as light and N3 as deep, based on both clinical literature[33–35] and general Garmin documentation (see https://www.garmin.com/en-US/garmin-technology/health-science/sleep-tracking/). For REM, the corresponding Vivofit 4 category was unclear because no documentation was available to indicate whether REM was considered light or deep when developing the Vivofit 4 algorithm, nor was a clear choice implicated in the literature.[36]

Table 1. Sleep variables included in the main analysis comparing data from the Garmin Vivofit 4 against polysomnography (PSG).

Variable	Description	Equivalence Zone[a]
Sleep onset (bedtime)	Time of falling asleep	± 0.5 hours
Sleep offset (risetime)[b]	Time of waking up	± 0.5 hours
Sleep onset latency[c]	Amount of time in bed before falling asleep, in minutes	± 15 minutes
Total sleep time (TST)[d]	Amount of time spent sleeping, in hours	± 0.5 hours
Wake after sleep onset (WASO)	Amount of time spent awake after falling asleep, in minutes	± 15 minutes
Sleep efficiency[e]	TST as a percentage of total time in bed	± 5%
Light sleep[f,g]	Percent of sleep spent in light stages	± 5%
Deep sleep[f,h]	Percent of sleep spent in deep stages	± 5%

[a] For equivalence testing, this was used as the acceptable margin of error, within which a group-level estimate (i.e. mean) from the Garmin could be considered equivalent to PSG.
[b] For PSG, given as "lights on" time, also equal to the sum of bedtime, TST, and WASO; for Garmin, calculated using the latter sum.
[c] Garmin value calculated as the time from "lights out" (PSG) to bedtime, with negative values excluded from further analysis (n = 1).
[d] Garmin value calculated as the sum of light and deep sleep time.
[e] Garmin value calculated by dividing TST by total in-bed time (i.e. latency + TST + WASO).
[f] Raw PSG values were expressed in percentage units; Garmin percentages were calculated by dividing minutes in each stage by TST.
[g] For PSG, defined as either stages N1–N2 plus rapid eye movement (REM Light) or only stages N1–N2 (REM Deep).
[h] For PSG, defined as either stage N3 (REM Light) or stage N3 and rapid eye movement (REM Deep).

Therefore, we chose to report results from two sets of tests, one with REM classified as light sleep (REM Light) and another with REM classified as deep sleep (REM Deep).

Data cleaning and sensitivity analyses. For the main analysis, Garmin data were cleaned using two procedures. First, estimates of sleep onset (hereafter called bedtime) or offset (hereafter called risetime) were flagged if they occurred before the monitor was attached or after it was removed. The manual override feature in Garmin Connect was then used to replace any such bedtimes or risetimes with the corresponding times of attachment or removal, respectively. Second, data were cleaned by excluding nights that the Garmin algorithm labeled as "tentative" rather than "final," based on its internal checks. Although these data cleaning steps came with a known risk of inflating the observed validity (particularly for bedtime and risetime), the critical advantage was that derivative variables (e.g. WASO and sleep stages) would be recalculated using only wear-time data from "final" nights, ensuring a realistic picture of accuracy.

To account for potential bias due to the data cleaning procedures, two sensitivity analyses were performed. The manual overrides were removed for both analyses, with the first continuing to exclude "tentative" nights while the second included both "final" and "tentative" nights. By comparing the results of the main analysis and sensitivity analyses, it was thus possible to examine the potential influence of nonwear and certitude ("tentative" or "final") on accuracy.

Statistical analyses. For each sleep variable, validity was tested at the group and individual levels. Group-level validity was assessed using equivalence testing, which is similar to standard t tests for mean difference, yet more appropriate for analyses for which the goal is to test similarity between measures.[37] Specifically, the null and alternative hypotheses are reversed in equivalence testing, so that rejection of the null hypothesis indicates significant equivalence (rather than significant difference) within a specified range of tolerance known as the equivalence zone. In the present analysis, the equivalence zones (see Table 1) were selected based on accepted margins of error whenever possible.[38,39] Individual level validity was summarized using mean absolute error ($\frac{1}{n}\sum_{i=1}^{n}|\mathrm{Garmin}_i - \mathrm{PSG}_i|$), and further examined via Bland-Altman analysis.[40,41] All statistical tests were performed with $\alpha = 0.05$. To account for multiple comparisons, p-values were adjusted using the false discovery rate correction.[42] Summary statistics are reported as mean ± SD.

Results

Data loss

A total of 42 participants completed the study protocol. Garmin sleep data were entirely missing in eight cases (i.e. no sleep time was predicted by the Garmin), which has also been observed elsewhere in laboratory[15] and free-living[43] studies. Thus, data were available from 34 participants prior to cleaning.

Data cleaning

No bedtime predictions occurred before the monitor was attached. However, one participant had undefined sleep onset latency and sleep efficiency because predicted bedtime occurred before "lights out" in the PSG report (see calculation footnotes of Table 1). There were nine risetimes that occurred after the monitor was removed. Eight of those values were replaced with the time of monitor removal, while the ninth (for which the removal time was missing) was replaced with the "lights on" time from PSG.

Of the 34 participants eligible for the main analysis, eight were labeled as "tentative" by the Garmin algorithm, resulting in exclusion. One additional participant was excluded because all sleep predictions from the Garmin occurred during nonwear (i.e. bedtime was predicted after removing the monitor). Thus, the final sample size for the main analysis was 25.

Main analysis

Table 2 provides a summary of participant characteristics. For risetime and TST, mean differences fell inside or close to the target range of ±30 minutes relative to PSG (Table 3). Strong group-level validity was also seen for sleep efficiency, with a mean difference of −3.3%. In contrast, mean differences had magnitudes of nearly an hour for bedtime and sleep onset latency, and more than half an hour for WASO. Mean sleep stage estimates differed considerably from PSG when using the REM Light classification scheme (difference magnitudes of ±20.7%), while they differed minimally when using the REM Deep classification scheme (difference magnitudes of ±3.3%). Standard deviations were large for all variables, contributing to nonsignificant equivalence tests (all adjusted $p > 0.99$). Raw data are shown in Figure 1.

Indicators of individual-level validity had greater magnitude than those for group-level validity, reflecting high variability and a tendency for over- and under-estimates to cancel out at the group level (see Table 3). This was especially true for TST (mean absolute error 1.6 times higher than the magnitude of mean bias) and sleep efficiency (2.7 times), as well as sleep stage estimates from the REM Deep classification scheme (3.7 times). Figure 2 shows Bland-Altman plots for each variable. There was limited systematic bias for most comparisons (adjusted $R^2 < 0.25$), with the exception of WASO (adjusted $R^2 = 0.94$). Consistent with the systematic bias for WASO, there was also evident heteroscedasticity for sleep

Table 2. Participant characteristics.

	Female (n = 9)	Male (n = 16)	Total (N = 25)
Age (years)	12.7 ± 5.2	10.5 ± 3.1	11.3 ± 4.0
Height (cm)	148.1 ± 22.9	144.4 ± 24.8	145.8 ± 23.7
Weight (kg)	75.3 ± 50.1	55.7 ± 41.8	62.7 ± 44.9
Weight status[a]			
Underweight	0 (0.0%)	1 (6.2%)	1 (4.0%)
Healthy weight	3 (33.3%)	9 (56.2%)	12 (48.0%)
Overweight	1 (11.1%)	1 (6.2%)	2 (8.0%)
Obese	1 (11.1%)	1 (6.2%)	2 (8.0%)
Severe obese	4 (44.4%)	4 (25.0%)	8 (32.0%)
Race[b]			
Black	0 (0.0%)	1 (6.2%)	1 (4.0%)
Hispanic	1 (11.1%)	0 (0.0%)	1 (4.0%)
White	8 (88.9%)	12 (75.0%)	20 (80.0%)
Multiracial	0 (0.0%)	3 (18.8%)	3 (12.0%)
Diagnosis			
None	0 (0.0%)	2 (12.5%)	2 (8.0%)
Insomnia	3 (33.3%)	2 (12.5%)	5 (20.0%)
Sleep-related breathing disorder	2 (22.2%)	6 (37.5%)	8 (32.0%)
Sleep-related movement disorder	1 (11.1%)	2 (12.5%)	3 (12.0%)
Multiple	3 (33.3%)	4 (25.0%)	7 (28.0%)

Values are mean ± SD for continuous variables, and n (%) for categorical.

[a] Body mass index (BMI) percentiles were calculated from Centers for Disease Control and Prevention growth charts. Weight status was binned into the following categories: < 5th percentile (underweight), 5th to 84.9th percentile (healthy weight), 85th to 94.9th percentile (overweight), 1.0–1.19 * 95th percentile (obese), and ≥ 1.2 * 95th percentile or BMI ≥ 35 (severe obese). For more information, see Kelly et al.[32]

[b] As recorded in electronic health record.

efficiency (i.e. greater spread for values < 85% than for values ≥ 85%).

Visual inspection of data in Figures 1 and 2 revealed no discernable trends along the lines of diagnosis, sex, or age. Although there were notable outliers in some cases (particularly for bedtime and risetime), the values did not tend to originate from the same participants in each panel except when common patterns would be expected (e.g. an erroneous estimate of bedtime corresponding to an erroneous estimate of sleep onset latency). Subsample sizes by sex and age were 4 (females ≤ 12 years), 11 (males ≤ 12 years), 5 (females > 12 years), and 5 (males > 12 years).

Sensitivity analyses

Results are presented in the Supplemental material. In the first sensitivity analysis (using uncorrected Garmin data from 26 nights with "final" designation), group-level

Table 3. Descriptive statistics of sleep-related variables from the criterion measure (polysomnography) and the Garmin Vivofit 4 monitor.

	Polysomnography	Garmin	Garmin difference	MAE
Bedtime (hours)[a]	22:14:40 ± 0.58	23:01:18 ± 1.16	0.84 ± 1.60	0.95 ± 1.53
Risetime (hours)[a]	05:59:45 ± 0.13	05:39:55 ± 0.47	−0.34 ± 0.70	0.47 ± 0.62
Sleep onset latency (minutes)[b]	22.8 ± 20.0	77.4 ± 100.9	54.6 ± 95.1	57.2 ± 93.6
Total sleep time (hours)	7.03 ± 0.78	6.49 ± 1.47	−0.55 ± 1.21	0.86 ± 1.01
Wake after sleep onset (minutes)	43.2 ± 37.9	5.2 ± 9.3	−38.0 ± 38.8	39.0 ± 37.7
Sleep efficiency (%)[b]	86.4 ± 8.5	83.1 ± 19.0	−3.3 ± 13.8	8.9 ± 10.9
Sleep stages (REM Light[c])				
Light sleep (%)	71.1 ± 10.3	50.4 ± 13.8	−20.7 ± 15.9	21.2 ± 15.2
Deep sleep (%)	28.9 ± 10.3	49.6 ± 13.8	20.7 ± 15.9	21.2 ± 15.2
Sleep stages (REM Deep[c])				
Light sleep (%)	53.7 ± 9.7	50.4 ± 13.8	−3.3 ± 14.9	12.2 ± 8.9
Deep sleep (%)	46.3 ± 9.8	49.6 ± 13.8	3.3 ± 14.9	12.1 ± 8.9

Values are mean ± SD.
[a]Summary values are circular mean ± SD, where SD was calculated using the mean shorter distance method (see paulhibbing.com/daytime); circular operations were not needed for paired calculations (mean difference and MAE), which is why the difference of circular means differs slightly from the values in the mean difference colum.n
[b]N of 24 rather than 25 because Garmin predicted bedtime before "lights out" for one participant, precluding calculation for these variables.
[c]Garmin monitor reported time in light and deep sleep. Polysomnography reported time in the REM stage and three non-REM stages (N1–N3). Polysomnography stages were coded in two ways, the first (REM Light) defining REM and N1–N2 as light sleep versus N3 as deep, while the second (REM Deep) defined only N1–N2 as light sleep and both N3 and REM as deep.
MAE: mean absolute error; REM: rapid eye movement.

performance was virtually unchanged for WASO. Mean difference for TST also had nearly identical magnitude compared to the main analysis, although the sign was reversed and the SD increased. Mean difference for sleep efficiency improved in comparison to the main analysis, dipping to −2.2% ± 13.3%. For the remaining variables, performance was worse in comparison to the main analysis, with mean differences exceeding an hour for time-based variables (bedtime, risetime, and sleep onset latency) and exceeding ±10% for sleep staging variables (regardless of how REM was classified). All equivalence tests were nonsignificant (adjusted $p > 0.99$). For individual-level validity, mean absolute errors increased 1.1 to 1.7 fold compared to the main analysis, except for risetime (3.9 fold increase) and sleep efficiency (0.9 fold decrease). Bland-Altman analysis continued to show limited evidence of systematic error (adjusted $R^2 < 0.30$, except for WASO with 0.92).

For the second sensitivity analysis (incorporating 8 "tentative" nights in addition to the 26 "final" nights from the first sensitivity analysis), the group-level results for TST and WASO changed very little relative to the first sensitivity analysis. There was also limited change for sleep staging estimates, regardless of how REM was classified. Mean difference for sleep efficiency improved substantially, reaching −0.82% ± 11.9%. For bedtime and sleep onset latency, mean differences returned to similar levels that were seen in the main analysis, while for risetime the mean difference remained much higher than what was seen for the main analysis (with slight improvement in comparison to the first sensitivity analysis). When comparing individual-level validity to the first sensitivity analysis, mean absolute error improved by 0.8 to 0.9 fold for bedtime, sleep onset latency, and sleep efficiency, while improvements were minimal for the remaining variables (0.96–1.00 fold). Indicators of systematic error remained negligible for all variables but WASO (adjusted $R^2 = 0.92$).

Discussion

In this study, we tested the validity of sleep estimates from the Garmin Vivofit 4 in a sample of youth experiencing sleep disturbances. Overall, there was mixed support for

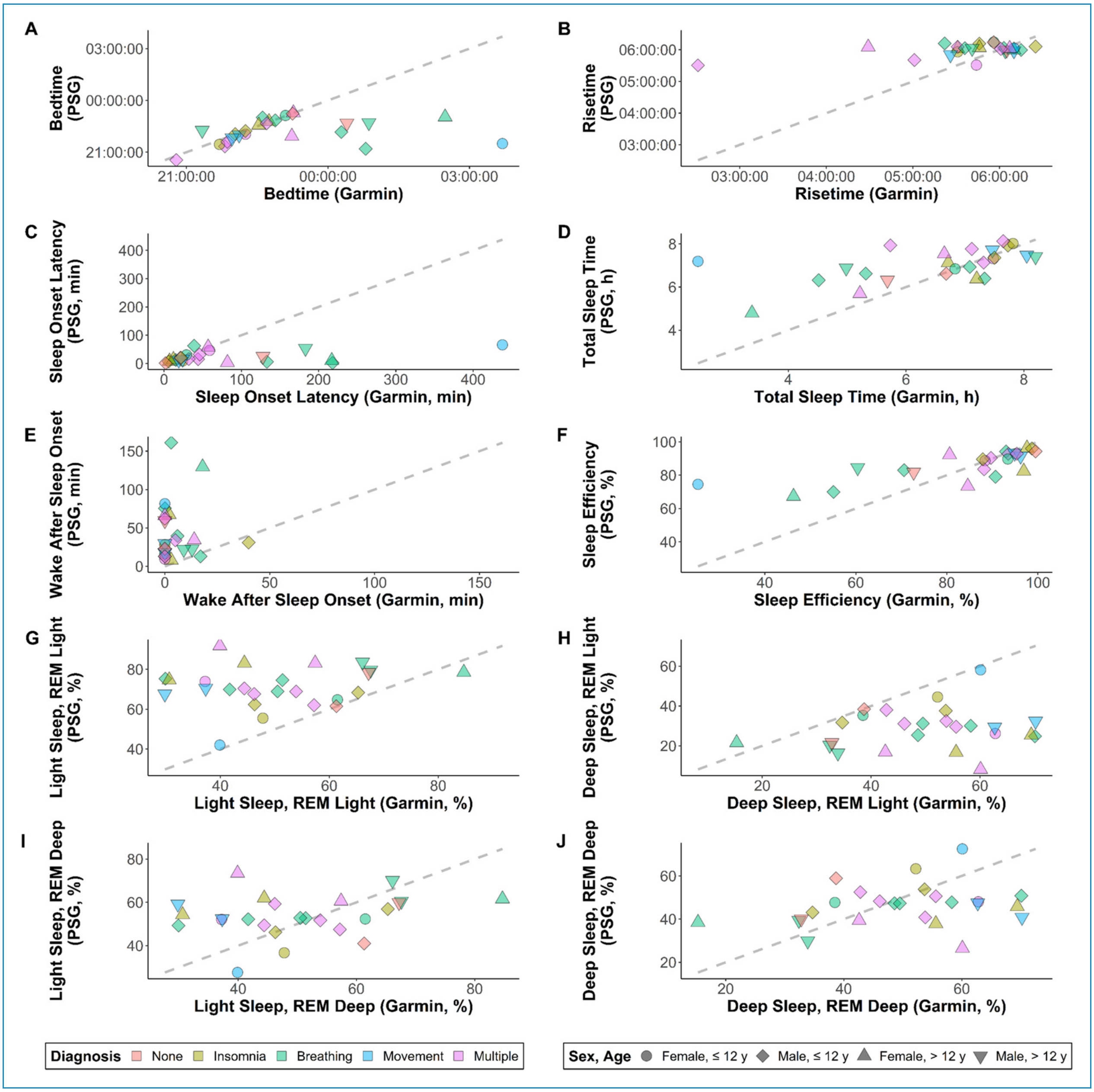

Figure 1. Scatterplots showing raw data for Garmin Vivofit 4 on the x-axis and polysomnography (PSG) on y-axis, with color- and shape-coding to show trends among diagnostic and demographic subgroups. The dashed line is the line of identity, representing perfect agreement. Variables shown are (a) bedtime, (b) risetime, (c) sleep onset latency, (d) total sleep time, (e) wake after sleep onset, (f) sleep efficiency, (g) percent of time in light sleep when classifying rapid eye movement as light (REM Light), (h) percent of time in deep sleep for REM Light, (i) percent of time in light sleep when classifying REM as deep (REM Deep), and (j) percent of time in deep sleep for REM Deep. For sleep onset latency and sleep efficiency, N = 24 rather than 25, due to the Garmin monitor predicting bedtime before "lights out" for one participant.

our initial hypotheses, except with respect to WASO, which was consistently underestimated. Individual-level findings were generally poor for all variables, but group-level findings fell inside or close to accepted margins of error for TST (mean values within 33–36 minutes of PSG across the main and sensitivity analyses) and sleep efficiency (within 0.82–3.3%).[38,39] This reflects cancelation of over- and under-estimates for those variables (indicated by small mean differences relative to larger mean absolute errors), which is important to consider when deciding whether to use the Vivofit 4 in future research. Although no results were significantly equivalent with PSG, numerous factors contributed to this, including not only prediction error, but also the small sample size, adjustments for

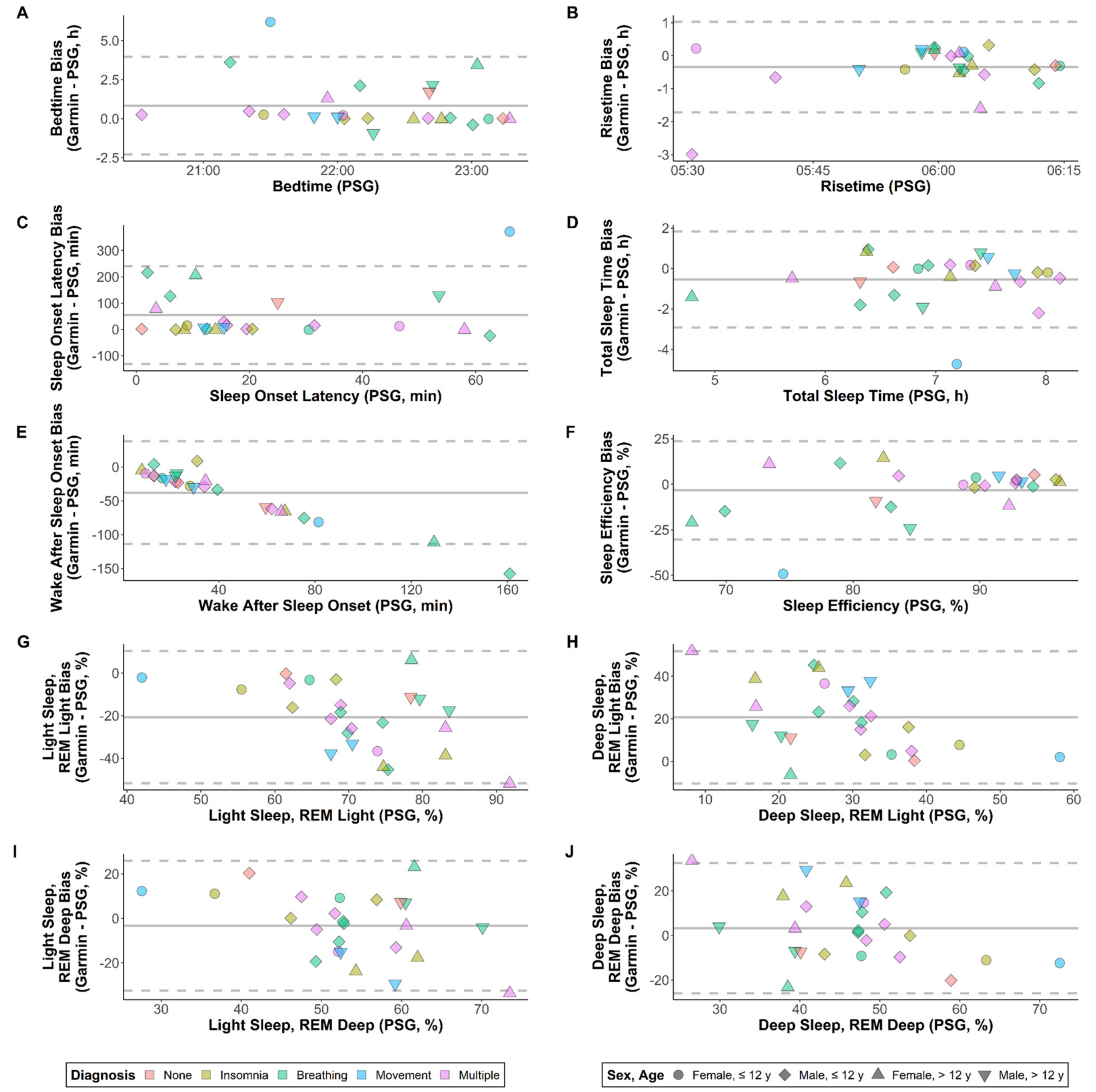

Figure 2. Bland-Altman plots showing patterns of bias (y-axis) relative to measured values from polysomnography (PSG; x-axis), with color- and shape-coding to show trends among diagnostic and demographic subgroups. Solid grey lines are mean bias, and dashed grey lines are limits of agreement. Variables shown are (a) bedtime, (b) risetime, (c) sleep onset latency, (d) total sleep time, (e) wake after sleep onset, (f) sleep efficiency, (g) percent of time in light sleep when classifying rapid eye movement as light (REM Light), (h) percent of time in deep sleep for REM Light, (i) percent of time in light sleep when classifying REM as deep (REM Deep), and (j) percent of time in deep sleep for REM Deep. For sleep onset latency and sleep efficiency, N = 24 rather than 25, due to the Garmin monitor predicting bedtime before "lights out" for one participant.

multiple comparison (30 total tests), and high standard deviation of the differences. Thus, larger studies are warranted to confirm our findings.

The sensitivity analyses demonstrated a potential impact of nonwear on observed validity of the Garmin algorithm, while there was limited impact of "final" versus "tentative" designation. These caveats have implications not only for the Vivofit 4 and the interpretation of our findings, but also for other monitors and research studies that may be similar in various respects. Accordingly, our discussion covers both the specific implications of this study and the broader implications for ongoing research.

Specific implications of this study

The Vivofit 4 is part of a continually expanding market of consumer-grade monitors that have developed substantial reach in the population for both personal use and applied research.[22] To date, however, only a few studies have evaluated sleep estimates from such monitors in pediatric samples.[17–19] To our knowledge, the current study was the first to test the popular Vivofit 4 in a clinical sample of youth, filling an important gap given the strengths of the device (e.g. scalability and cost-effectiveness) and the high prevalence and underdiagnosis of sleep disorders in the youth population.[6–8]

Other pediatric evaluations have been focused on Fitbit monitors, with the Ultra model being tested by Meltzer et al.[17] and the Charge HR model being tested by de Zambotti et al.[18] and Godino et al.[19] Only the sample of Meltzer et al.[17] was comparable to ours, whereas healthy youth were recruited for the other studies.[18,19] The results of Meltzer et al.[17] showed generally poor performance of the Fitbit Ultra compared to PSG, and they also observed a tendency for systematic error along lines of age and condition severity. Specifically, they noted that older children and adolescents exhibited more motionless wake time, which may have factored into greater underestimation of WASO in that group. Similarly, their findings suggested that movement was more prominent for youth with sleep disordered breathing than those without. In the present study, we did not observe comparable systematic error along lines of age, sex, or diagnosis, although we did observe error being higher for some individuals than others. In particular, several variables in Figure 1 followed a pattern in which most predictions were fairly accurate while a small number were highly inaccurate. The lack of systematic pattern in our study may be attributable to our contrasting sample characteristics (e.g. inclusion of many different disorders) or the small sample size, the latter of which was also acknowledged as a limitation in the study by Meltzer et al.[17]

Consistent with broader research on other monitors across populations,[14,16–20] our findings showed dramatic underestimation of WASO when using the Vivofit 4. We also experienced a high incidence of missing data, which was consistent with what has been reported for other Garmin monitors.[15,43] In the free-living study by Kubala et al.,[43] results could not be reported from the Vivosmart HR due to missing data, whereas sufficient data were obtained from six other monitors. Kainec et al.[15] also reported a high rate of data loss for the Garmin Vivosmart 4 in their laboratory-based study (5.6% of cases), but they reported even higher rates of data loss for all but one other monitor (5.6–18.8% of cases). Thus, while missing data is clearly a prevalent issue with Garmin monitors, it appears not to be entirely unique among consumer-grade monitors, and there is no clear pattern to suggest a single cause (e.g. device-related, protocol-related, or something else).

Before moving on from the specific implications of the present study, it is important to note that we only addressed the night-level criterion validity of the Vivofit 4. There may be different implications when looking at other measurement attributes such as sensitivity to change or the motivational impact of wearing the monitor itself. Thus, the appropriateness of the Vivofit 4 in clinical, scientific, or personal use may vary depending on the objective, and more research is needed to determine utility across the full range of potential uses.

Broader implications for future research

The present study followed a common design for clinical testing of sleep estimates from consumer-grade wearable monitors,[15–17] yet also had unique characteristics that may be instructive for future research. One such characteristic was the need for customized procedures to address the division of sleep stages into only two categories (light and deep) and the classification of nights as “tentative” or “final.” While these specific characteristics may not apply to other monitors, there may be analogous traits in other monitors that require similar nuance and customization when designing the analyses for a performance evaluation. These customizations can be approached in the manner we demonstrated for the present study, based on scientific and clinical guidance wherever possible, along with any available information from the manufacturer. However, it is notable that manufacturer information is often sparse, due to proprietary restrictions that result in a “black box” conundrum.[12] We strove to make this a transparent part of our reporting, and future analyses should do the same so that potential users can factor it into their considerations about using a given monitor.

Apart from issues that affect individual monitors, we also encountered general issues that may affect many monitors and the design of study protocols for testing them. These issues included familiar ones such as the lack of access to epoch-level data[11,15] and specific design tradeoffs of the in-lab PSG protocol (e.g. the balance of internal and external validity during a single-night assessment in an artificial laboratory environment).[44–46] A novel issue we addressed was related to how much data were collected before bedtime and after risetime. To our knowledge, this issue has not been addressed or discussed in prior work, and we sought to address it through our use of manual overrides and sensitivity analyses. Therefore, we offer detailed comments below.

Regarding data collected before bedtime, Garmin suggests that some of their other monitors should be worn for at least 2.0 hours before going to bed (see https://support.garmin.com/en-US/?faq=qvzNMwxuTb9NxZ6Ce2a9z9). While there is no such guideline for the Vivofit 4, it remains possible that predictions in the current study were impacted by nonwear in the period before bedtime. Some contrary evidence may be visible in Figure 1, where most bedtime predictions fell close to the line of identity. This could indicate the Garmin algorithm normally predicts bedtime effectively, regardless of the amount of data collected beforehand. Nevertheless, there were also clear outliers, which may represent cases where there was unique need for more pre-bedtime data. This could also explain why the outliers did not follow clear patterns across sex, age, or diagnosis. Overall, there is a clear need for additional research to explore whether the amount of data collected before bedtime has a strong impact on predictions, not only for the Garmin Vivofit 4, but also for other monitors.

Regarding data collected after risetime, the concern is inverse of what was described before. That is, it deals with excess data collected after the study rather than insufficient data collected before. In the current study, the Garmin monitors continued to collect data after being removed, similar to what might occur in other studies (e.g. anytime there is a lag between removing the monitor and retrieving the data) or in personal use (e.g. if removing the monitor to bathe immediately after waking). This additional data may register with a low-movement profile that is classified as a continuation of sleep rather than a transition to nonwear, especially for devices like the Vivofit 4 that include only an accelerometer sensor. We were able to partially address this issue by comparing results from manually corrected versus uncorrected data in the main and sensitivity analyses. Results showed that nonwear after risetime did influence some variables (especially risetime itself), while others were less affected (e.g. bedtime, sleep onset latency, and WASO). Therefore, it seems likely that the findings for the latter variables were legitimate and not heavily influenced by the present study design. This highlights an important point, namely, that monitor performance and study design can be stronger in some areas than others. Nuanced interpretation of the evidence is therefore crucial for understanding the strengths, weaknesses, and unknowns of consumer-grade sleep monitors (especially those for which epoch-level data are unobtainable).

Lastly, the present findings should be considered alongside estimates that 25% to 50% of typically developed children and adolescents have a sleep disorder,[6,7] while only 3.7% of cases are diagnosed.[8] This suggests that the measurement issues observed in the present study could affect large numbers of unidentified youth, even in studies of apparently healthy individuals with no known sleep disorders. A related implication is that comparisons against healthy control samples must be made with caution (i.e. only after clinical confirmation that no undetected sleep disturbance is present in the healthy controls). Our recruitment strategy was designed to enroll participants with sleep disorders, but there were a small number ($n = 2$) whose sleep study did not point to a specific diagnosis. Although this is far too few to support definitive comments on accuracy for those with versus without sleep disorders, we observed no clear differences in the limited data we were able to present. Future studies should examine this more closely while understanding that rigorous verification of healthy control status will be required.

Strengths and limitations

This study had strengths and limitations. A key strength was the combined focus on a widely used monitor (Vivofit 4) and an underrepresented population in prior research (youth with sleep disorders). Furthermore, our sample included participants experiencing a range of disturbances. While the latter diversity was a strength of the sample, the small sample size was a limitation that led to insufficient power for detecting subgroup differences. The issue of sample size was further complicated by the inherent limitations of a single-night PSG protocol, along with monitor failures and “tentative” sleep classifications, as discussed previously. Epoch-level data may have made these issues easier to address, but such data were not available through the Garmin API platform. Our analyses also did not account for medication status of the participants or potential intermonitor differences that may have added noise to the assessment (although we were able to limit the latter possibility by using only three different devices across all participants). The sample had limited racial and ethnic diversity, which is especially important to consider given known disparities in prevalence of sleep disorders across these lines.[2,47] In general, the characteristics of our sample make it difficult to generalize results to broader pediatric populations, particularly since individuals experiencing sleep disturbances may pose unique measurement challenges that are not as prevalent among sound sleepers. Additional research is needed to evaluate consumer monitors in other and broader population groups. Nevertheless, the present study provides important insights into sleep monitoring for pediatric patient populations, including to highlight protocol-related issues in validation research.

Conclusions

The Garmin Vivofit 4 has uncertain utility for assessing sleep in youth with sleep disturbances. Group-level estimates of TST and sleep efficiency may be acceptable in some settings, while performance is weaker for other variables and individual-level estimates. Validity appears to be strong for many individuals and weak for a select few, whose characteristics follow unclear patterns. These issues could be protocol-related, and may therefore also affect other validation studies that use a similar protocol. Overall, caution is warranted when using the Garmin Vivofit 4 in future research where criterion validity is a primary need. The need for caution is especially important considering the high prevalence and underdiagnosis of sleep disturbances in youth. Future studies should test performance in larger and more diverse samples with multiple days of data collection, and explore reliability and sensitivity to change, with attention to the potential accuracy differential between those with and without a sleep disorder.

Acknowledgements: The Garmin Vivofit 4 brand and product name are used in accordance with "fair-use" provisions of the United States Copyright Act (section 107). Results from this study have not been published elsewhere or duplicated in any form, and no wearable device companies were involved in the conduct of the study or in the interpretation and publication of the results. This manuscript was prepared as part of the Summer Training in Academic Research (STAR) 2.0 program through Children's Mercy Kansas City. The authors wish to thank the Children's Mercy Sleep Center and its staff for their assistance with data collection for the study.

Contributorship: PRH, MP, BAS, DGI, and JAC designed the study. PRH and MP processed the data. PRH, LB, AB, TN, SS, NS, and DF analyzed the data, performed the literature review, and drafted the manuscript. BAS and DGI provided clinical oversight. All authors reviewed and edited the manuscript and approved the final version of the manuscript.

Declaration of conflicting interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval: The ethics committee of Children's Mercy Kansas City approved this study (REC number: STUDY00001890).

Funding: The authors received no financial support for the research or authorship of this article. The publication of the article was supported by a minigrant through the STAR 2.0 program.

Guarantor: PRH

ORCID iD: Paul R. Hibbing https://orcid.org/0000-0002-1469-5160

Supplemental material: Supplemental material for this article is available online.

References

1. Colrain IM and Baker FC. Changes in sleep as a function of adolescent development. *Neuropsychol Rev* 2011; 21: 5–21.
2. Meltzer LJ, Williamson AA and Mindell JA. Pediatric sleep health: it matters, and so does how we define it. *Sleep Med Rev* 2021; 57: 101425.
3. Alrousan G, Hassan A, Pillai AA, et al. Early life sleep deprivation and brain development: insights from human and animal studies. *Front Neurosci* 2022; 16: 833786.
4. Landhuis CE, Poulton R, Welch D, et al. Childhood sleep time and long-term risk for obesity: a 32-year prospective birth cohort study. *Pediatrics* 2008; 122: 955–960.
5. Matricciani L, Paquet C, Galland B, et al. Children's sleep and health: a meta-review. *Sleep Med Rev* 2019; 46: 136–150.
6. Carter K, Hathaway N and Lettieri C. Common sleep disorders in children. *Am Fam Physician* 2014; 89: 368–377.
7. Stores G. Sleep disorders in children and adolescents. *BJPsych Adv* 2015; 21: 124–131.
8. Meltzer LJ, Johnson C, Crosette J, et al. Prevalence of diagnosed sleep disorders in pediatric primary care practices. *Pediatrics* 2010; 125: e1410–8.
9. Gozal D and Kheirandish-Gozal L. New approaches to the diagnosis of sleep-disordered breathing in children. *Sleep Med* 2010; 11: 708–713.
10. de Zambotti M, Cellini N, Goldstone A, et al. Wearable sleep technology in clinical and research settings. *Med Sci Sports Exercise* 2019; 51: 1538–1557.
11. Menghini L, Cellini N, Goldstone A, et al. A standardized framework for testing the performance of sleep-tracking technology: step-by-step guidelines and open-source code. *Sleep* 2021; 44: zsaa170.
12. Schutte-Rodin S, Deak MC, Khosla S, et al. Evaluating consumer and clinical sleep technologies: an American Academy of Sleep Medicine update. *J Clin Sleep Med* 2021; 17: 2275–2282.
13. Mouritzen NJ, Larsen LH, Lauritzen MH, et al. Assessing the performance of a commercial multisensory sleep tracker. Ferri R, editor. *PLoS ONE* 2020; 15: e0243214.
14. Chinoy ED, Cuellar JA, Huwa KE, et al. Performance of seven consumer sleep-tracking devices compared with polysomnography. *Sleep* 2021; 44: zsaa291.
15. Kainec KA, Caccavaro J, Barnes M, et al. Evaluating accuracy in five commercial sleep-tracking devices compared to research-grade actigraphy and polysomnography. *Sensors (Basel)* 2024; 24: 635.
16. Montgomery-Downs HE, Insana SP and Bond JA. Movement toward a novel activity monitoring device. *Sleep Breath* 2012; 16: 913–917.
17. Meltzer LJ, Hiruma LS, Avis K, et al. Comparison of a commercial accelerometer with polysomnography and actigraphy in children and adolescents. *Sleep* 2015; 38: 1323–1330.
18. de Zambotti M, Baker FC, Willoughby AR, et al. Measures of sleep and cardiac functioning during sleep using a multisensory commercially-available wristband in adolescents. *Physiol Behav* 2016; 158: 143–149.
19. Godino JG, Wing D, Zambotti Md, et al. Performance of a commercial multi-sensor wearable (Fitbit charge HR) in measuring physical activity and sleep in healthy children. *PLoS ONE* 2020; 15: e0237719.
20. Roberts DM, Schade MM, Mathew GM, et al. Detecting sleep using heart rate and motion data from multisensor consumer-grade wearables, relative to wrist actigraphy and polysomnography. *Sleep* 2020; 43: zsaa045.
21. Russo K, Goparaju B and Bianchi M. Consumer sleep monitors: is there a baby in the bathwater? *NSS* 2015; 7: 147–157.
22. Evenson KR, Goto MM and Furberg RD. Systematic review of the validity and reliability of consumer-wearable activity trackers. *Int J Behav Nutrit Physical Activ* 2015; 12: 159.

23. Depner CM, Cheng PC, Devine JK, et al. Wearable technologies for developing sleep and circadian biomarkers: a summary of workshop discussions. *Sleep* 2020; 43: zsz254.
24. Evenson KR and Spade CL. Review of validity and reliability of Garmin activity trackers. *J Measur Physical Behav* 2020; 3: 170–185.
25. Ancoli-Israel S, Cole R, Alessi C, et al. The role of actigraphy in the study of sleep and circadian rhythms. *Sleep* 2003; 26: 342–392.
26. Marino M, Li Y, Rueschman MN, et al. Measuring sleep: accuracy, sensitivity, and specificity of wrist actigraphy compared to polysomnography. *Sleep* 2013; 36: 1747–1755.
27. de Zambotti M, Menghini L, Grandner MA, et al. Rigorous performance evaluation (previously, "validation") for informed use of new technologies for sleep health measurement. *Sleep Health* 2022; 8: S2352721822000171.
28. Berry R, Quan S, Abreu A, et al. *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications, version 2.6*. Darien, IL: American Academy of Sleep Medicine, 2020.
29. Cole RJ, Kripke DF, Gruen W, et al. Automatic sleep/wake identification from wrist activity. *Sleep* 1992; 15: 461–469.
30. Sadeh A, Sharkey KM and Carskadon MA. Activity-based sleep-wake identification: an empirical test of methodological issues. *Sleep* 1994; 17: 201–207.
31. Kuczmarski R. 2000 CDC growth charts for the United States: methods and development. *Vital Health Stat* 2002; 11: 1–190.
32. Kelly AS, Barlow SE, Rao G, et al. Severe obesity in children and adolescents: identification, associated health risks, and treatment approaches: a scientific statement from the American Heart Association. *Circulation* 2013; 128: 1689–1712.
33. Carskadon MA and Dement WC. Chapter 2 – normal human sleep: an overview. In: Kryger MH, Roth T and Dement WC (eds) *Principles and practice of sleep medicine (fourth edition)*. Philadelphia: W.B. Saunders, 2005 [cited 2024 Apr 5], pp.13–23. Available from: https://www.sciencedirect.com/science/article/pii/B0721607977500094.
34. Malhotra RK and Avidan AY. Sleep stage scoring. In: Thomas RJ, Bhat S and Chokroverty S (eds) *Atlas of sleep medicine*. Cham: Springer International Publishing, 2023 [cited 2024 Apr 5], pp.125–163. Available from: https://doi.org/10.1007/978-3-031-34625-5_7.
35. Patel AK, Reddy V, Shumway KR, et al. Physiology, sleep stages. In: *StatPearls* [internet]. Treasure Island (FL): StatPearls Publishing, 2024 [cited 2024 Apr 5]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK526132/
36. Blumberg MS, Lesku JA, Libourel PA, et al. What is REM sleep? *Curr Biol* 2020; 30: R38–R49.
37. Dixon PM, Saint-Maurice PF, Kim Y, et al. A primer on the use of equivalence testing for evaluating measurement agreement. *Med Sci Sports Exercise* 2018; 50: 837–845.
38. Werner H, Molinari L, Guyer C, et al. Agreement rates between actigraphy, diary, and questionnaire for children's sleep patterns. *Arch Pediatr Adolesc Med* 2008; 162: 350–358.
39. Driller MW, O'Donnell S and Tavares F. What wrist should you wear your actigraphy device on? Analysis of dominant vs. non-dominant wrist actigraphy for measuring sleep in healthy adults. *Sleep Sci* 2017; 10: 132–135.
40. Bland J and Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; 327: 307–310.
41. Bland JM and Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res* 1999; 8: 135–160.
42. Benjamini Y and Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Royal Stat Soc Series B* 1995; 57: 289–300.
43. Kubala AG, Barone Gibbs B, Buysse DJ, et al. Field-based measurement of sleep: agreement between six commercial activity monitors and a validated accelerometer. *Behav Sleep Med* 2020; 18: 637–652.
44. Roomkham S, Lovell D, Cheung J, et al. Promises and challenges in the use of consumer-grade devices for sleep monitoring. *IEEE Rev Biomed Eng* 2018; 11: 53–67.
45. Bianchi MT. Sleep devices: wearables and nearables, informational and interventional, consumer and clinical. *Metabolism* 2018; 84: 99–108.
46. Chaudhry FF, Danieletto M, Golden E, et al. Sleep in the natural environment: a pilot study. *Sensors* 2020; 20: 1378.
47. Williamson AA, Johnson TJ and Tapia IE. Health disparities in pediatric sleep-disordered breathing. *Paediat Resp Rev* 2022; 45: 2–7. 2022 Jan 28 [cited 2022 Aug 2]; Available from: https://www.sciencedirect.com/science/article/pii/S1526054222000057