JCSM
Journal of Clinical Sleep Medicine

EMERGING TECHNOLOGIES

Sleep Validity of a Non-Contact Bedside Movement and Respiration-Sensing Device

Margeaux M. Schade, PhD[1,2]; Christopher E. Bauer, PhD[1,3]; Billie R. Murray, BS[1]; Luke Gahan, MEng[4]; Emer P. Doheny, PhD[4]; Hannah Kilroy, MAI[4]; Alberto Zaffaroni, MEng[4]; Hawley E. Montgomery-Downs, PhD[1]

[1]Department of Psychology, West Virginia University, Morgantown, West Virginia; [2]Department of Biobehavioral Health, Pennsylvania State University, State College, Pennsylvania; [3]Department of Neuroscience, University of Kentucky, Lexington, Kentucky; [4]ResMed Sensor Technologies Ltd., Dublin, Ireland

Study Objectives: To assess the sleep detection and staging validity of a non-contact, commercially available bedside bio-motion sensing device (S+, ResMed) and evaluate the impact of algorithm updates.

Methods: Polysomnography data from 27 healthy adult participants was compared epoch-by-epoch to synchronized data that were recorded and staged by actigraphy and S+. An update to the S+ algorithm (common in the rapidly evolving commercial sleep tracker industry) permitted comparison of the original (S+V1) and updated (S+V2) versions.

Results: Sleep detection accuracy by S+V1 (93.3%), S+V2 (93.8%), and actigraphy (96.0%) was high; wake detection accuracy by each (69.6%, 73.1%, and 47.9%, respectively) was low. Higher overall S+ specificity, compared to actigraphy, was driven by higher accuracy in detecting wake before sleep onset (WBSO), which differed between S+V2 (90.4%) and actigraphy (46.5%). Stage detection accuracy by the S+ did not exceed 67.6% (for stage N2 sleep, by S+V2) for any stage. Performance is compared to previously established variance in polysomnography scored by humans: a performance standard which commercial devices should ideally strive to reach.

Conclusions: Similar limitations in detecting wake after sleep onset (WASO) were found for the S+ as have been previously reported for actigraphy and other commercial sleep tracking devices. S+ WBSO detection was higher than actigraphy, and S+V2 algorithm further improved WASO accuracy. Researchers and clinicians should remain aware of the potential for algorithm updates to impact validity.

Commentary: A commentary on this article appears in this issue on page 935.

Keywords: actigraphy, consumer device, sleep, validation

Citation: Schade MM, Bauer CE, Murray BR, Gahan L, Doheny EP, Kilroy H, Zaffaroni A, Montgomery-Downs HE. Sleep validity of a non-contact bedside movement and respiration-sensing device. *J Clin Sleep Med.* 2019;15(7):1051–1061.

INTRODUCTION

The consumer market for non-medical, sleep tracking technology and devices is highly competitive. Commercial sleep tracking technologies such as accelerometry-based wearable sensors, mattress-based sensors, mobile device-incorporated sensors (audio, vibration, camera), and hundreds of associated mobile sleep applications have become widely available.[1,2] North America and Europe account for approximately 65% of the sleep tracking device market, indicating an upward trend with the United States expected to dominate in purchasing thorough 2024.[3] However, this rapid market growth outpaces objective, impartial evaluation of device accuracy, which requires comparison against the physiologic gold-standard, polysomnography (PSG), and data evaluation using rigorous techniques such as epoch-by-epoch comparison for concordance. Although commercially available devices are marketed to health-conscious consumers and may not strive for clinical-level diagnostic utility, consumers nonetheless draw conclusions about their sleep from the feedback these devices provide. The current state of this industry is that the available devices overestimate sleep quality and quantity, and that they perform inconsistently, with poorer sleepers receiving less accurate feedback.[1]

PSG is the most consistent and physiologically accurate method of sleep identification currently available. It remains the best, and arguably only, quality control measure against which to test commercial device accuracy, yet it is not the only way that sleep is monitored in research and clinical science. Actigraphy (ie, scientific-grade accelerometry) is the primary alternative to PSG,[4] which developed because the latter can be cost-prohibitive, time-consuming, and—because it is usually laboratory-based—can disturb sleep compared to what would occur during a typical night at home.[5,6] Actigraphy is used clinically to evaluate patients' sleep/wake patterns[7,8] even though it has substantial limitations. Actigraphy alone is unable to distinguish sleep stages. Compared to PSG, actigraphy consistently shows diminished validity when the wearer is awake (ie, it has poor "specificity") because brief awakenings that produce little motion, and the physical stillness that precedes sleep onset, are often overlooked by motion-dependent algorithms. This inaccuracy results in overestimation of sleep time and quality; in contrast, actigraphy shows high accuracy for identification of true sleep (ie, it shows high "sensitivity").[8–10]

The fidelity of sleep estimation by consumer-marketed sleep tracking technology is not currently held to any regulatory standard. A dearth of published, systematic validation of

commercially available devices against PSG suggests that few of these devices have undergone independent assessment.[2,11] As of a 2017 review, no publicly available smartphone applications were in existence that had been successfully validated for sleep monitoring, although many are designed to accompany motion-sensitive devices.[12] A comprehensive review of all devices, their claims, and scarce empirical evidence about them was reviewed by Ko and colleagues in 2015[2] yet since that time many new products and updates have been released. Although validations adhering to best-practice standards are few, those that are draw the same basic conclusions about commercial device performance: available devices are limited in their specificity, particularly after sleep onset.[2,8–11,13–19] Marketing of wearable sleep tracking devices has now progressed to claim that these devices can identify sleep stages, despite the absence of empirical support for their ability to accurately detect sleep/wake.[20–22]

A unique category of sleep monitoring devices, non-contact bedside radiofrequency biomotion sensors (NRBS), use remote biosensing technologies, largely circumventing interference caused by even light instrumentation during sleep.[22,23] These devices may be a step forward in the consumer sleep tracking industry, but they too must undergo rigorous evaluation compared to the gold standard.

Our primary goal was to assess the validity of S+ (ResMed, San Diego, California), a non-contact NRBS device, compared to PSG. We also compared the S+ to actigraphy, in relation to PSG. During the study period, an updated algorithm was released by ResMed, with the potential to impact device performance. Thus, a secondary goal became evaluating whether, and how, the algorithm update changed device performance.

METHODS

Protocol

Participants were recruited using electronic, campus-wide advertisements. The study was approved by the West Virginia University Office of Research Integrity and Compliance (IRB). Following screening for inclusion criteria, informed consent was administered in person and signed by all participants. Participants arrived ~1 hour before their usual bedtime and spent one night in the sleep research laboratory, during which they underwent simultaneous monitoring by PSG, actigraphy, and S+. Participants were awakened and each study concluded at 6:00 AM (unless a different wake time was participant-requested). Participants were compensated for their time and travel with a $150 gift card.

Participant Sample

Inclusion criteria were age (18 years or older) and being a healthy sleeper. Sleep health was estimated using the Epworth Sleepiness Scale (cutoff score = 13),[24] and previous diagnosis with or current symptom(s) of a sleep disorder. Participants each had a body mass index (BMI) under 40 kg/m^2 (assessed by height and weight measured in the sleep laboratory). Women who were pregnant or within 6 weeks postpartum were also excluded. One participant was excluded from analyses, and referred for clinical evaluation, on the basis of having an apnea-hypopnea > 5 events/h. No participant exceeded 5 periodic limb movement sequences per hour.

Polysomnography

Overnight PSG was recorded and analyzed using the REMbrandt system (Embla N7000; REMbrandt manager software [version 9.1]; Natus Medical Incorporated, Pleasanton, California). A standard 6-channel EEG montage was used, following the international 10–20 system for EEG placement (F3, F4, C3, C4, O1, and O2 with contralateral mastoid reference and a separate universal reference electrode).[25,26] Masseter and tibialis anterior electromyography, bilateral electrooculography, pulse oximetry, and electrocardiography were recorded; air flow and respiratory effort were measured using an oronasal thermistor and chest and abdominal respiratory inductance plethysmography; a reverberation sensor placed over the trachea was used to detect snoring. Continuous audio-video was recorded.

Full PSG with audio-video was scored for stage, arousals and events (respiratory-related and leg movement-related) by a Registered Polysomnographic Technologist (RPSGT) (M.S.), who was blind to actigraphy and S+ outcomes, according to American Academy of Sleep Medicine (AASM) standards.[26] All recorded epochs were included in calculations of sleep, sleep stages, and periods of wake, including equipment adjustment and restroom use time, except when data were considered unscorable by PSG (eg, poor signal quality, occurring for 26.2% of recorded data for 1 participant and for less than 2% of recorded data for 5 additional participants). Total sleep time (TST) and the percentage of each sleep stage relative to TST were calculated. Additional metrics were: wake after sleep onset (WASO), which included epochs scored as wake after initial sleep onset was scored, through either the first epoch of the final awakening or the end of the recording, whichever occurred first; and wake before sleep onset (WBSO), which included epochs scored as wake from the light's out (ie, the beginning of the recording) to the first epoch of sleep. We further calculated a metric representative of sleep quality that is independent of sleep onset latency and wake time at the end of the recording interval by dividing TST by the sleep period time, which we defined as the time from AASM-defined sleep onset through the participant's final awakening.

S+

The S+ is a non-contact sleep monitoring device that uses ultra-low power radiofrequency waves to monitor the user's movements while they are in bed. The average power emitted by the device is 1 mW (about 100 times lower than a standard mobile phone). The S+ uses a patented, range-gated sensing technology that ensures the range of operation is 1.5 meters, so that it monitors only the individual in closest proximity to the sensor. The sensor is capable of detecting sub-millimeter chest wall movements, which permits monitoring independent of sleeping position. The sensor's signal is designed to be unaffected by bedding or bed clothes. The system includes proprietary algorithms which are capable of estimating individual movement events' magnitude and duration as well as each breathing movement's amplitude and duration. The algorithm combines high resolution estimations of these parameters into 30-second

epoch features which are then used to map the bio-motion signals to different sleep stages. Bed presence is detected through an increase in the received signal power. User wakefulness (W) is associated with higher levels and longer duration of detected movement, and a higher degree of respiration rate variability. Rapid eye movement (REM) sleep is associated with the highest variability of detected respiratory parameters, while light sleep (intended to correspond to stage N1 and N2 sleep) and deep sleep (corresponding to stage N3 sleep) trend to increasing stability of the detected metrics. The system differs from previous non-contact technologies[22,23] insofar as its features were redesigned and the system was retrained to leverage insights obtained from new data acquired from both overnight PSG (ie, not part of the current study)[27] and home recordings. The S+ also leverages the strengths of a newer sensor release with higher signal to noise ratio. The S+ was initially launched in November 2014 (S+V1); in September 2015, an algorithm update designed to improve respiration and movement detection, by decreasing the system's sensitivity to spurious movements of short durations, which were found to have a very loose relationship with wakefulness, was released (S+V2).

The S+ works in conjunction with a mobile phone device, with which it is paired via Bluetooth; S+ streams data acquired by the sensor to a mobile phone app, where it is processed in real time. Summary information about sleep is transmitted to the cloud, where an advice engine combines trends in sleep parameters to provide the user with sleep-related recommendations. Both nightly and summary data are accessible to users through the mobile phone app or a dedicated website.

The S+ was placed in a standardized position on a bedside table, pointed toward the participant's chest. The S+ was connected to an iPod Touch (5th generation, iOS 8.1; Apple Inc. Cupertino, California) using the custom app (S+ by ResMed; 2015 ResMed Sensor Technologies Ltd), which stored and auto-scored the raw biomotion data. If S+ correctly identified a period when a participant was out of the bed, it was considered accurate in its scoring of wakefulness, as absence detection is an algorithm feature.

Actigraphy

Actiwatch 64 (AW-64; Mini Mitter, Inc., Bend, Oregon) was used with sleep analysis software (Actiware version 5.71.0, Philips Respironics, Murrysville, Pennsylvania). The device was worn on the non-dominant wrist. Data were recorded and auto-scored in 30-second epochs using "medium" wake threshold (the most commonly used threshold in other work utilizing this device)[28] and all other parameters were set to default for sleep/wake scoring using zero-crossing mode.

Device Synchronization

To allow epoch-by-epoch comparison between PSG and S+, and between PSG and actigraphy, the two computers and iPod Touch used for data collection among three methods were synchronized to within 1 second and then initiation of each recording was preprogrammed to begin simultaneously (removing the need for, and variation from, manual initiations). Each recording was analyzed using the simultaneous 30-second epochs from each device.

To compare PSG and S+ to actigraphy, which does not identify sleep stages, epochs scored as any sleep stage (N1, N2, N3, or R for PSG; light, deep and REM for S+) were converted to "sleep." For comparisons between PSG and S+, PSG epochs scored as stage N1 and N2 sleep were combined for comparisons to light sleep as measured by the S+. A summary of corresponding measures for each recording method/device is in **Figure S1** in the supplemental material.

Statistical Analyses

Agreement was calculated between each of the two S+ algorithm versions (S+V1 and S+V2) and PSG, and between actigraphy and PSG. Comparisons were then made between performance by S+ and actigraphy using repeated-measures ANOVA with *post hoc* pairwise comparisons. Performance for sleep stage identification was evaluated between S+V1 and S+V2 using repeated-measures *t* tests (n = 22). Data met normal distribution assumptions (skew and kurtosis < 3.2 times the standard error); no scores were 3 standard deviations beyond average. Greenhouse-Geisser correction was applied in cases without sphericity. Device agreement, sensitivity (stage-specific where appropriate), and specificity were evaluated. By convention, "sensitivity" is the percentage of sleep epochs (per PSG) that were accurately identified by S+ or actigraphy; "specificity" is the percentage of wake epochs (per PSG) that were accurately identified by S+ or actigraphy. We also calculated stage-specific sensitivity. Data are presented as mean ± standard deviation.

To display concordance between PSG and each device, we used a modified Bland-Altman plotting technique,[9] where PSG is considered the gold standard for comparison, rather than plotting the test devices against an average of the test device and PSG.[29] This approach highlights absolute differences between measures, at each magnitude, rather than whether they trend similarly.

Participant Characteristics

Thirty participants were recruited and data from 27 who were not excluded due to abnormal PSG (n = 1), total S+ data loss (n = 1), or key metric (WBSO; n = 1) loss were included in validity evaluation for actigraphy and S+V1 (**Figure S2** in the supplemental material). These 3 excluded participants did not differ significantly from the 27 included (χ^2 and independent samples *t* test analyses). Among the 27 participants available for inclusion in analyses, 40.7% were female; their mean age was 29.1 ± 11.7 years; mean years of education was 15.6 ± 2.5; median annual household income was $65,000 ± $71,606; one participant was Asian, and all others were white; 14.8% of the sample was married. Mean and median BMI were 26.8 ± 5.9 and 25.9 kg/m^2, respectively, corresponding to mildly overweight.[30] Eight participants were sent follow-up letters informing them of incidental findings during PSG (rare respiratory events, limb movements, hypnic jerks, and teeth grinding).

Technical malfunction in raw data transmission to back-end servers (which was required for analysis with updated algorithm) resulted in the exclusion of 5 additional participants from S+V2 algorithm analyses (**Figure S2** in the supplemental material). These 5 participants did not differ on any

Table 1—Sleep time and percentage from polysomnography.

Sleep Measure	n	Range (hours)	Hours	%TST	%SP
TST	27	4.0–7.0	5.6 (1.0)	–	85.1 (13.3)
Light sleep	27	1.7–4.2	3.3 (0.7)	58.8 (10.9)	49.2 (8.0)
Stage N1 sleep	27	0.2–1.8	0.8 (0.5)	14.9 (10.0)	11.7 (6.9)
Stage N2 sleep	27	1.4–3.7	2.5 (0.7)	44.0 (7.7)	37.5 (9.1)
Stage N3 sleep	27	0.4–2.5	1.5 (0.5)	26.9 (9.7)	23.3 (10.0)
Stage R sleep	27	0.0–2.1	0.8 (0.5)	14.3 (7.5)	12.7 (7.2)
WBSO	27	0.0–1.9	0.6 (0.5)	–	–
WASO	27	0.1–3.1	0.9 (0.9)	–	13.7 (12.3)

Number of hours and percentage of sleep time (out of TST, or relative to the SP) on the recording night. Values are presented as mean (standard deviation) unless otherwise indicated. SP = sleep period, TST = total sleep time, WASO = wake after sleep onset, WBSO = wake before sleep onset.

Table 2—Repeated-measure ANOVA comparing sleep/wake detection across actigraphy, S+V1, and S+V2 (n = 22).

Outcome Measure		Accuracy (%) Mean (SD)	*F*	*P*	η^2	*P*[b]
Overall sleep/wake	Actigraphy	85.1 (8.9)				vs. S+V1: .007
	S+V1	87.5 (8.3)				vs. S+V2: n.s.
	S+V2	87.6 (8.2)	7.72	.001	0.27	vs. Act: .005
Sleep (sensitivity)	Actigraphy	96.6 (2.6)				vs. S+V1: .049
	S+V1	94.8 (4.2)				vs. S+V2: .026
	S+V2	93.8 (3.7)	7.49[a]	.006	0.26	vs. Act: .003
WBSO (specificity)	Actigraphy	46.6 (26.6)				vs. S+V1: < .001
	S+V1	87.7 (16.9)				vs. S+V2: .326
	S+V2	90.4 (15.2)	45.00[a]	< .001	0.68	vs. Act: < .001
WASO (specificity)	Actigraphy	42.9 (18.1)				vs. S+V1: .237
	S+V1	48.3 (22.5)				vs. S+V2: .041
	S+V2	53.2 (21.6)	3.92[a]	.044	0.16	vs. Act: .017
Overall wake (specificity)	Actigraphy	47.6 (19.7)				vs. S+V1: < .001
	S+V1	69.5 (19.1)				vs. S+V2: .076
	S+V2	73.1 (20.0)	46.06[a]	< .001	0.69	vs. Act: < .001

Differences in sleep and wake epoch-by-epoch accuracy (%, relative to polysomnography) between each device or device algorithm. [a]Greenhouse-Geisser correction for Sphericity violation ($P < .05$). [b]*Post hoc* pairwise comparisons, least significant difference. Act = actigraphy, ANOVA = analysis of variance, S+V1 = algorithm version 1, S+V2 = algorithm version 2, SD = standard deviation, WASO = wake after sleep onset, WBSO = wake before sleep onset.

characteristics from the 22 who were included (χ^2 and independent samples *t* test analyses adjusted for unequal variances). Further, S+V1 agreement for these 5 participants did not differ from the 22 participants included in ANOVA and repeated-measures *t* tests. For actigraphy and S+V1, descriptive statistics are reported based on all participants available (n = 27).

RESULTS

Data from S+V1, S+V2, and actigraphy were compared epoch-by-epoch for agreement with scored PSG. The range of time in each sleep stage, the percentage of time in each sleep stage, and minutes of WBSO and WASO according to PSG are available in **Table 1**. Agreement accuracy for overall sleep and wake, and for WBSO and WASO, is available in **Table 2**. Agreement accuracy for sleep stages is available in **Table 3**. A separate evaluation of combined light sleep, deep sleep, REM sleep, and wake ("four-stage") accuracy was also performed; in this case, S+V1 had overall sensitivity of 60.6% ± 9.2% and S+V2 a sensitivity of 61.8% ± 7.0%. Bland-Altman concordance for the S+ device for overall sleep and wake is illustrated in **Figure 1A** and **Figure 2A**, and concordance for sleep stages is illustrated in **Figure 3**.

Figure 4 shows sensitivity and specificity values for S+ and actigraphy, as well as a previously-published[31] reference threshold for typical PSG inter-scorer reliability (ISR; ie, the consistency in epoch-by-epoch scoring among credentialed PSG scorers). Bland-Altman concordance for overall sleep and wake identified by actigraphy is illustrated in **Figure 1B** and **Figure 2B**. A summary of Bland-Altman concordance values across the three devices is in **Table 4**.

Magnitude of Device Discrepancies

Participants' TST according to PSG was 338 ± 58 minutes, or just over 5.5 hours. Average wake time was 1 hour 40 minutes ± 1 hour 11 minutes). Devices overestimated TST by at least 20 minutes for 9 (33%, S+V1), 6 (27%, S+V2), and 16 (59%,

Table 3—Repeated-measure *t* tests comparing stage detection between S+V1 and S+V2 (n = 22).

Outcome Measure		Accuracy (%) mean (SD)	*t*	*P*	*d*
Light sleep	S+V1	64.0 (10.8)			
	S+V2	65.1 (8.4)	0.61	.548	0.11
Stage N1 sleep	S+V1	60.8 (13.7)			
	S+V2	59.4 (14.7)	0.59	.562	0.10
Stage N2 sleep	S+V1	65.3 (12.2)			
	S+V2	67.6 (10.4)	1.31	.205	0.21
Stage N3 sleep	S+V1	61.1 (16.0)			
	S+V2	52.2 (15.4)	2.76	.012	0.56
Stage R sleep	S+V1	61.5 (30.9)			
	S+V2	61.6 (28.7)	0.02	.981	0.00

Differences in sleep stage epoch-by-epoch accuracy (%, relative to polysomnography) between S+ algorithms. S+V1 = algorithm version 1, S+V2 = algorithm version 2, SD = standard deviation.

Table 4—Summary of Bland-Altman concordance.

Measure	S+V1	S+V2 [a]	Actigraphy
Total sleep	17.7 (61.4)	13.2 (51.4)	43.4 (53.5)
Light (stage N1 + N2) sleep	6.4 (52.2)	12.8 (41.2)	–
Deep (stage N3) sleep	1.0 (50.0)	−10.5 (40.3)	–
REM (stage R) sleep	10.2 (26.6)	10.7 (28.5)	–

Average number of minutes' disagreement (concordance) between a device or device algorithm and polysomnography. Values are presented as mean (standard deviation). [a]n = 22; other measures n = 27. REM = rapid eye movement, S+V1 = algorithm version 1, S+V2 = algorithm version 2.

actigraphy) participants. Instances of TST underestimation by at least 20 minutes were fewer, but occurred for 5 (19%, S+V1), 5 (27%, S+V2), and 2 (7%, actigraphy) participants (**Figure 1A** and **Figure 1B**). The influence of sleep quality on device estimates of sleep is illustrated in the supplemental material (**Figure S3**). In all cases of sleep quality below 90% (n = 13 for S+V1 and actigraphy; n = 10 for S+V2), actigraphy overestimated sleep time. Below 90% sleep quality, S+V1 overestimated sleep time for 8 (62%) and underestimated for 5 (38%) of participants, while S+V2 overestimated and underestimated sleep time equally.

In all cases where wakefulness exceeded 2.5 hours, S+ underestimated wake (5 participants for S+V1 and 2 participants for S+V2). When wake time was under 2.5 hours, total wake time estimation by both S+V1 and S+V2 was more equivocal. Actigraphy rarely overestimated wake (4 participants, in two cases by fewer than 5 minutes) and, as wake time increased, actigraphy's wake underestimation worsened (**Figure 2A** and **Figure 2B**).

S+V1 overestimated light sleep by at least 20 minutes for 10 participants (37%) and underestimated light sleep by at least 20 minutes for 9 participants (33%); S+V2 overestimated light sleep for 6 (27%) and underestimated for 3 (14%). When light sleep was in excess of 4 hours, the S+ consistently underestimated light sleep time (**Figure 3A**).

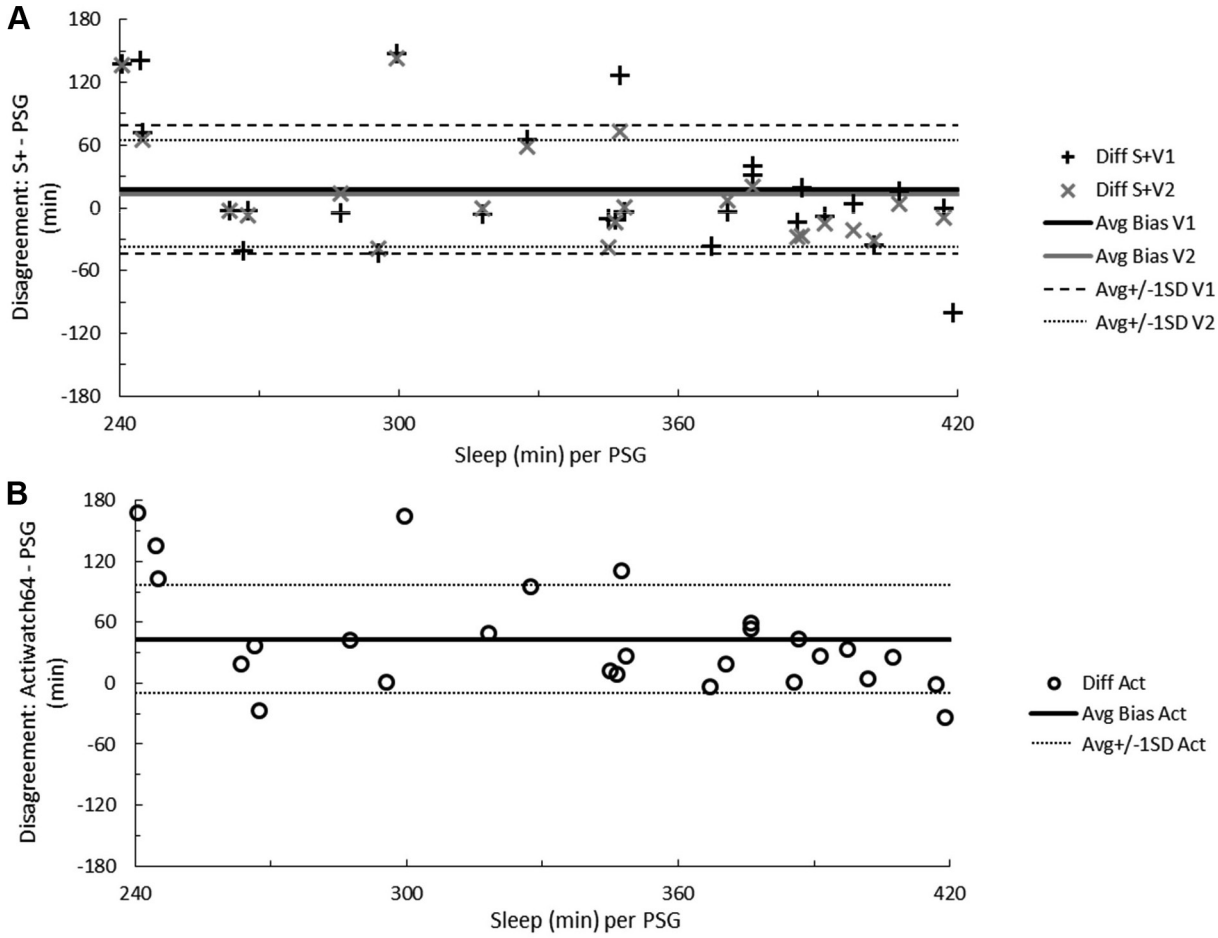
For deep sleep, S+V1 overestimated by at least 20 minutes for 8 participants (30%) and underestimated by at least 20 minutes for 9 participants (33%). S+V2 overestimated by at least 20 minutes for 4 participants (18%) and underestimated by at least 20 minutes for 8 participants (36%). Below 1.5 hours deep sleep the S+ tended to overestimate, and above 1.5 hours deep sleep the S+ tended to underestimate, deep sleep time (**Figure 3B**).

Both S+V1 and S+V2 overestimated REM sleep, each by at least 20 minutes for 8 participants (30% and 36%, respectively), while REM sleep underestimation by S+V1 occurred for 2 participants (7%) and underestimation by S+V2 occurred for 4 participants (18%). When REM sleep exceeded 1 hour, S+V2 underestimation followed an increasing trend (**Figure 3C**)

Device and Algorithm Performance Comparisons

Performance of S+V1, S+V2, and actigraphy were compared to each other. Percent accuracy accomplished by each device is available in **Table 2** and **Table 3**. Both S+V1 and S+V2 had significantly higher overall agreement with PSG than did actigraphy (pairwise actigraphy versus S+V1 $P = .007$; pairwise actigraphy versus S+V2 $P = .005$); S+V1 and S+V2 did not differ significantly (**Table 3**). Within-groups, S+V1 and S+V2 four-stage sensitivity did not differ. S+ sensitivity was lower than actigraphy for both S+V1 ($P = .049$) and S+V2 ($P = .003$); S+V1 four-stage sensitivity was higher than S+V2 ($P = .026$). S+V1 and S+V2 did not differ significantly on overall specificity, and specificity for both was significantly higher than actigraphy (both, $P < .001$). Both S+V1 and S+V2 had significantly higher specificity than actigraphy for detecting WBSO ($P < .001$, both), but did not differ significantly from one another. S+V2 had higher specificity for WASO than both S+V1 ($P = .041$) and actigraphy ($P = .017$); S+V1 and actigraphy did not differ significantly for WASO specificity. Only deep sleep

Figure 1—Total sleep time disagreement between measurements.



Bland-Altman concordance between total sleep time as measured by PSG and the S+ device **(A)** or actigraphy **(B)**. Solid lines indicate average discrepancy, dashed and dotted lines indicate ± 1 standard deviation, and symbols correspond to the number of minutes differing for each participant. On the ordinate, a positive difference indicates overestimation and a negative difference underestimation relative to PSG. The range of participant sleep time in minutes, according to RPSGT-scored PSG, is on the abscissa. Act = actigraphy, Avg = average, Diff = difference, PSG = polysomnography, S+V1 = algorithm version 1, S+V2 = algorithm version 2, SD = standard deviation.

sensitivity significantly differed between devices; deep sleep was lower for S+V2 than S+V1 (P = .012; **Table 3**).

DISCUSSION

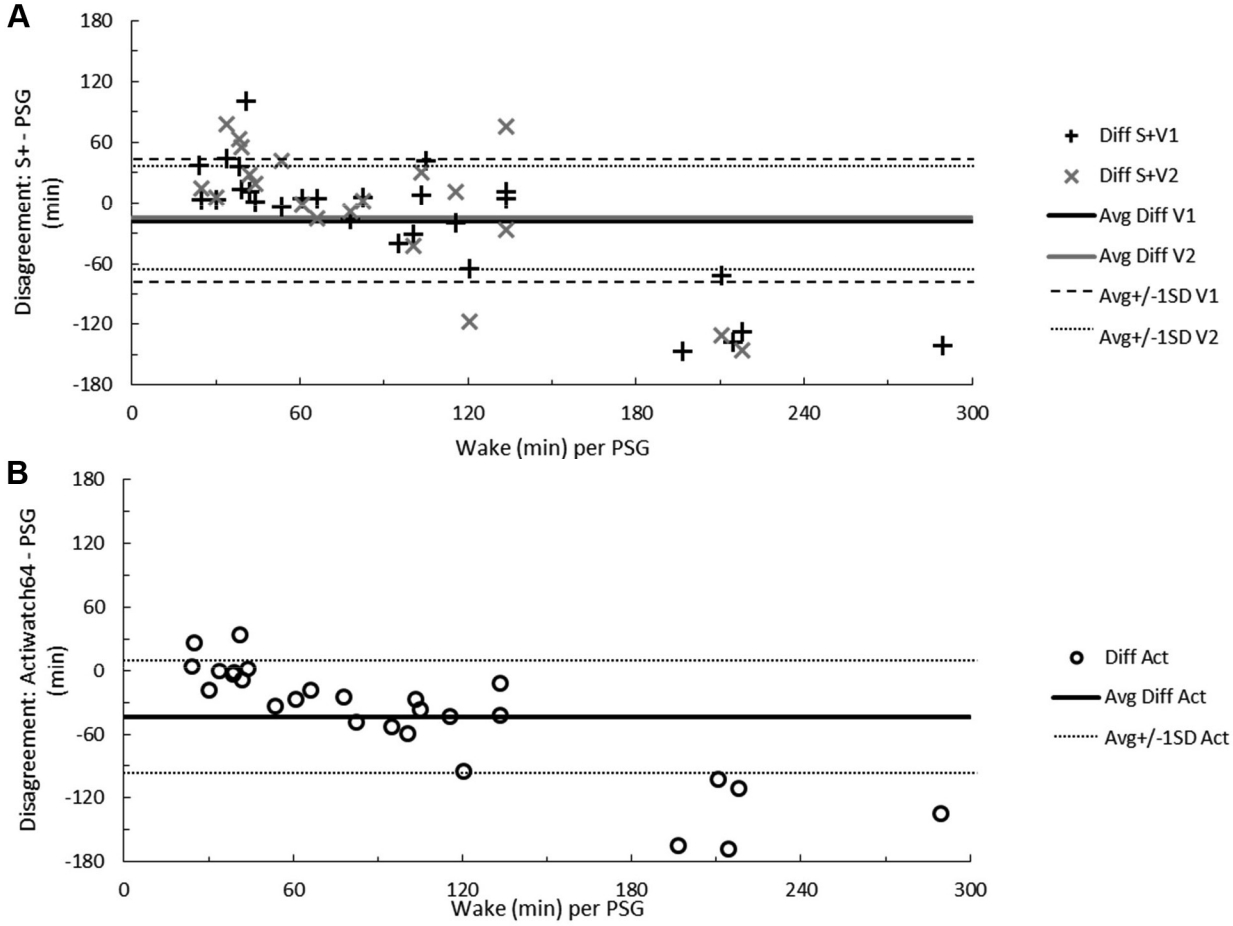
The respiratory and motion sensing S+ device was able to identify sleep and wake with an accuracy of about 87% relative to PSG, regardless of algorithm. Its sensitivity to sleep, over 90%, was higher than its specificity, between 70% and 75%. S+ was weakest in the same area as most comparable devices; however, in WASO detection (where it correctly identified epochs scored as wake by a credentialed technologist about 50% of the time), the S+V2 algorithm detected WASO significantly better than actigraphy. Sleep staging accuracy by the S+ did not exceed 68% for any stage.

S+ Performance Summary Relative to PSG

Wake detection agreement with the RPSGT scorer by the S+ was lower than two experienced RPSGT scorers might be expected to achieve relative to one another (80% wake agreement).[31] However, 70% epoch-by-epoch agreement is an improvement over actigraphy, which is on average under 50%. When identifying wakefulness up to about 1.5 hours, both versions of the S+ algorithm were prone to bias similar to or better than the ISR error rate. This contrasts with other actigraph-like devices that aim to identify sleep and wake without distinguishing stage: the Fitbit (prior to release of Alta HR) and Jawbone UP have both tested with lower specificity than actigraphy, especially concerning WASO detection.[9,32] Relative to other non-contact biomotion sensors of its type, S+ also performed well overall: sensitivity, specificity, and overall sleep/wake agreement were similar to the SleepMinder (87% to 95% sensitivity, 42% to 50% specificity, and 75% to 86% overall agreement with PSG)[22] sensitivity was similar to the SleepDesign HSL-101 (96%) but S+ specificity was notably higher (versus 38%).[23]

ISR disagreement for light sleep nears 14%,[31] while the S+ disagreed with the scoring technologist about 35% of the time. For both S+V1 and S+V2, identifying stage N1 sleep as sleep was more challenging than identifying stage N2 sleep

Figure 2—Total wake time disagreement between measurements.



Bland-Altman concordance between total wake time as measured by PSG and the S+ device **(A)** or actigraphy **(B)**. Solid lines indicate average discrepancy, dashed and dotted lines indicate ± 1 standard deviation, and symbols correspond to the number of minutes differing for each participant. On the ordinate, a positive difference indicates overestimation and a negative difference underestimation relative to PSG. The range of participant wake time in minutes, according to RPSGT-scored PSG, is on the abscissa. Act = actigraphy, Avg = average, Diff = difference, PSG = polysomnography, S+V1 = algorithm version 1, S+V2 = algorithm version 2, SD = standard deviation.

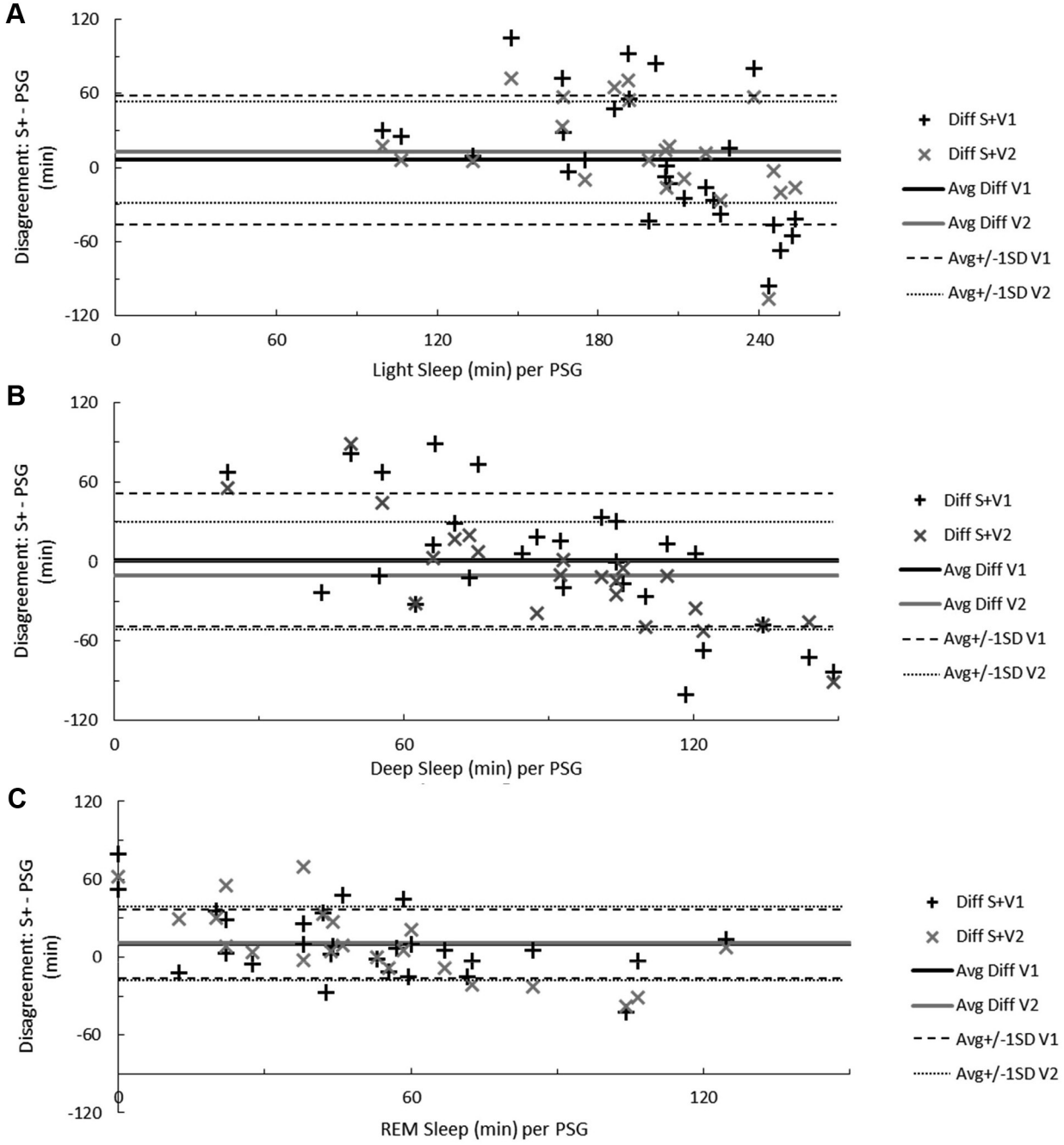
as sleep; this trend is also applicable among human scorers, where stage N1 sleep disagreement is much higher than stage N2 sleep (differing by about 40%).[31] Deep sleep (stage N3 sleep) bias (**Figure 3B**) was lowest between 1 and 2 hours of deep sleep time. Among technologists, deep sleep is agreed upon about 70% of the time; the higher-performing S+V1 algorithm achieved about 60% agreement with our technologist. Although similar for both S+V1 and S+V2, epoch-by-epoch REM sleep agreement with the RPSGT was about 20% lower by the S+ than the roughly 80% expected agreement among credentialed scorers.

S+ Performance Summary Relative to Actigraphy

Although S+ had significantly higher overall specificity and significantly lower overall sensitivity than actigraphy, its overall sleep/wake agreement with PSG was higher. Both algorithms also had higher WBSO epoch-by-epoch agreement than actigraphy, with performance differences between 40% and 50%. Further, from a clinical perspective, the approximately 10% performance difference observed in favor of S+V2 relative to actigraphy is robust enough to potentially impact therapeutic decision-making.

Wake detection and sleep overestimation are consistent limitations of movement-based algorithms. As such, limitations of this device type tend to be most severe with clinical populations who struggle with sleep continuity. Referring to Bland-Altman **Figure 1B** and **Figure 2B**, and consistent with extant literature, actigraphy consistently overestimated TST and underestimated wake time. Wake underestimation by actigraphy in this study was almost universal and appeared to be further related to sleep quality (**Figure S3**); S+ was also likely to underestimate wake time in the context of very large amounts (> 3 hours) of wakefulness, but garnered a performance edge when participant wake time was less extreme. The most dramatic underestimations for both actigraphy and S+ occurred in these instances of excessive wakefulness, as might be expected given the known limitations of wake detection, with several underestimations approaching 2 hours discrepancy with PSG. Nonetheless, devices should aim to be reliable even when nocturnal sleep is of low quality.

Figure 3—Sleep staging disagreement between measurements.



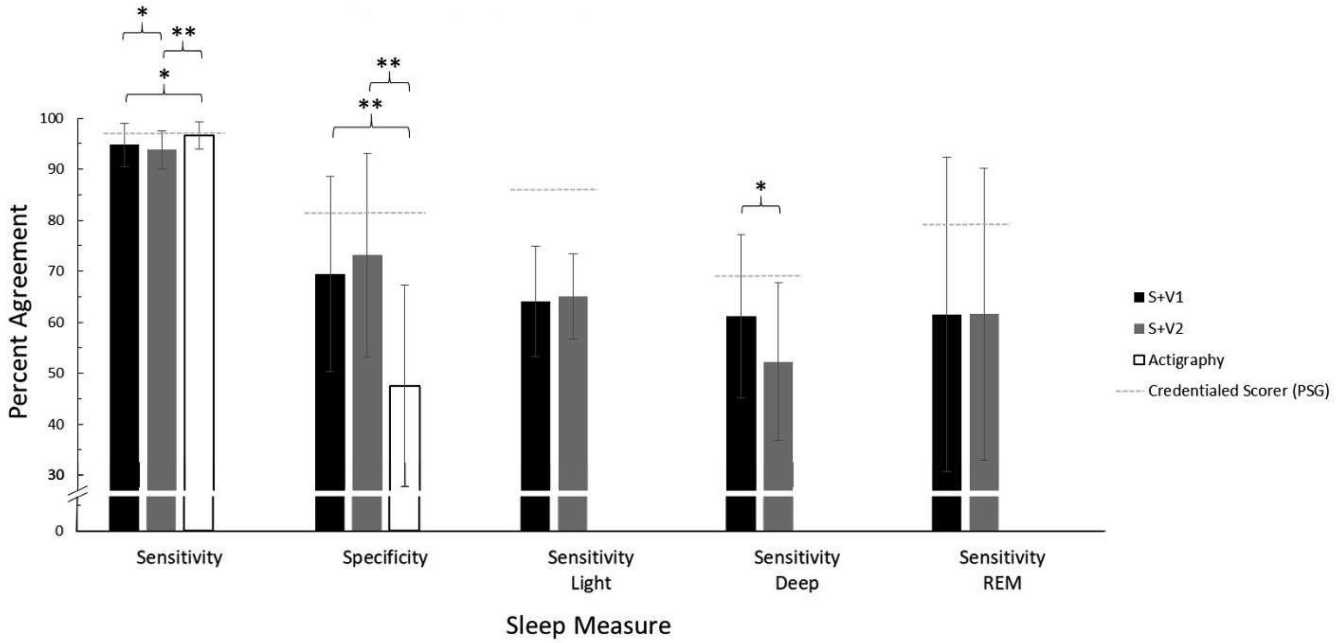
Bland-Altman concordance between the S+ device and PSG when recognizing both stage N1 and N2 sleep as light sleep **(A)**, stage N3 sleep as deep sleep **(B)**, and REM sleep **(C)**. Solid lines indicate average discrepancy, dashed and dotted lines indicate ± 1 standard deviation, and symbols correspond to the number of minutes differing for each participant. On the ordinate, a positive difference indicates overestimation and a negative difference underestimation relative to PSG. The range of participant light, deep, or REM sleep time in minutes, according to RPSGT-scored PSG, is on the abscissa. Avg = average, Diff = difference, PSG = polysomnography, S+V1 = algorithm version 1, S+V2 = algorithm version 2, SD = standard deviation.

Extant literature also suggests that actigraphy suffers bias (underestimation) after about 30 minutes of WASO, and our data appear to corroborate this observation.[10] There were more participants with overestimated wake time by S+ algorithms than by actigraphy, such that S+ may overestimate when detecting wake time of less than 60 minutes. Between 60 and 120 minutes of wake, the S+ provided a more accurate overall estimate of wake time.

S+V1 and S+V2 Performance Comparison

S+V1 and S+V2 did not differ on overall epoch-by-epoch accuracy, using either dichotomous sleep/wake (for like-comparison with actigraphy) or four-stage (light, deep, REM, and wake) categorizations. Nor did the algorithm versions differ on overall specificity, WBSO specificity, light sleep (combined stage N1 and N2 sleep) sensitivity, individual stage N1 or N2 sleep sensitivity, or REM sleep sensitivity. However, S+V2

Figure 4—Epoch-by-epoch agreement with PSG.



Percent agreement either between a sleep monitoring device (S+V1, S+V2, or actigraphy) and a corresponding PSG record (scored by an RPSGT) in this study, or between PSG records scored by multiple registered technologists (indicated by dashed gray lines; inter-scorer reliability).[31] Error bars indicate standard deviation (n = 22 in all device groups). Overall sensitivity and specificity results reflect post hoc pairwise outcomes after significant omnibus repeated-measures ANOVA. Stage-specific sensitivity results reflect outcomes of within-subjects *t* tests. * $P < .05$, ** $P < .01$. PSG = polysomnography, REM = rapid eye movement sleep, S+V1 = algorithm version 1, S+V2 = algorithm version 2.

changes did result in lower overall sleep sensitivity, lower stage N3 sleep sensitivity, and higher WASO specificity compared to S+V1. This algorithm adjustment therefore appears to have correctly targeted a common limitation of movement-based devices: wake detection (or sleep overestimation).

S+V2 WASO specificity was apparently at the expense of sleep detection sensitivity. The clinical significance of a statistical change in sensitivity around 1% is not as compelling as the potential for up to 5% improvement in WASO accuracy, particularly given that WASO is so consistently evasive of algorithm detection among sleep-estimating devices. Nonetheless, a both statistically and clinically sizeable specificity reduction (of about 9%) with the implementation of S+V2 highlights the importance of sustaining sleep identification accuracy while algorithms aim to improve wake detection.

Limitations

In this work, we used a relatively small sample of only healthy sleepers, so extrapolation to clinical groups cannot be inferred. Our sample's sleep architecture also deviated from our expectations of normal sleepers, in both stage and TST. REM sleep time was lower than typically reported for young healthy adults (14.3% versus about 25%), as was stage N2 sleep time (44.0% versus about 50%). Stage N1 sleep time was higher than expected (14.9% versus < 5%), as was stage N3 sleep time (26.9% versus about 20%). Low TST in the laboratory (5.6 hours on average) may have contributed to skewed architecture distributions, especially because participants may have had more REM sleep time if permitted to sleep beyond scheduled morning awakening in the laboratory. Further, low TST may also have artificially inflated device accuracy because it decreased the opportunity for more naturally occurring awakenings; however, this alarm clock-based context is also ecologically valid. The impact of a novel sleeping environment (ie, the laboratory and monitoring equipment) is likely to have also contributed to elevated time spent in stage N1 sleep. A first-night effect[33] may have negatively affected the quality of participant sleep and, given that wake and stage N1 sleep are typically low-performing stages for devices, may have resulted in an underestimation of in-home device performance on the whole. Nonetheless, a device that is valid and reliable under more sleeping circumstances than only the most ideal offers greater translational value, clinically.

Future Directions

Future work should evaluate the S+ sleep onset latency accuracy for the purpose of clinical translation, as our analysis of WBSO cannot be considered an equivalent representation of device accuracy in detecting this measure; the clinically significant measure of latency from "lights out" to sleep onset.

CONCLUSIONS

Although the S+ was quantitatively worse than the ISR for PSG in all standard metrics, it offers the advantages of automated staging and wireless, non-contact sleep data collection in the standard home environment. Relative to other published

evaluation of commercially available sleep-tracking devices that do not incorporate neurocortical data, this bedside device better identifies WBSO—a major challenge in this industry. There is still room to improve WASO specificity and sleep stage detection of the S+ and other devices. Consumers should be wary, however, that their sleep data from S+ and other devices may change in the wake of ongoing algorithm adaptations—not all of which are improvements in every respect—and data may not reflect an actual change in consumer sleep quality or quantity while manufacturers strive to achieve their highest-performing algorithm.

ABBREVIATIONS

AHI, apnea-hypopnea index
ANOVA, analysis of variance
BMI, body mass index
EEG, electroencephalography
IRB, Institutional Review Board
NRBS, non-contact bedside radiofrequency biomotion sensors
PSG, polysomnography
REM, rapid eye movement
RPSGT, Registered Polysomnographic Technologist
TST, total sleep time
WASO, wake after sleep onset
WBSO, wake before sleep onset

REFERENCES

1. Shelgikar AV, Anderson PF, Stephens MR. Sleep tracking, wearable technology, and opportunities for research and clinical care. *Chest.* 2016;150(3):732–743.
2. Ko PR, Kientz JA, Choe EK, Kay M, Landis CA, Watson NF. Consumer sleep technologies: a review of the landscape. *J Clin Sleep Med.* 2015;11(12);1455–1461.
3. Smart sleep tracking device market: global demand analysis & opportunity outlook 2024. Research Nester website. https://www.researchnester.com/reports/smart-sleep-tracking-device-market-global-demand-analysis-opportunity-outlook-2024/373. Published October 3, 2017. Accessed June 21, 2019.
4. Sadeh A. The role and validity of actigraphy in sleep medicine: an update. *Sleep Med Rev.* 2011;15(4):259–267.
5. Lorenzo J-L, Barbanoj M-J. Variability of sleep parameters across multiple laboratory sessions in healthy young subjects: the "very first night effect". *Psychophysiology.* 2002;39(4):409–413.
6. Le Bon O, Staner L, Hoffmann G, et al. The first-night effect may last more than one night. *J Psychiatr Res.* 2001;35(3);165–172.
7. Ancoli-Israel S, Cole R, Alessi C, Chambers M, Moorcroft W, Pollak CP. The role of actigraphy in the study of sleep and circadian rhythms. American Academy of Sleep Medicine review paper. *Sleep.* 2003;26(3):342–392.
8. Morgenthaler TI, Lee-Chiong T, Alessi C, et al., Standards of Practice Committee of the AASM. Practice parameters for the clinical evaluation and treatment of circadian rhythm sleep disorders: an American Academy of Sleep Medicine report. *Sleep.* 2007;30(11):1445–1459.
9. Montgomery-Downs, HE, Insana, SP, Bond, JA. Movement toward a novel activity monitoring device. *Sleep Breath.* 2012;16(3):913–917.
10. Marino M, Li Y, Rueschman MN, et al. Measuring sleep: accuracy, sensitivity, and specificity of wrist actigraphy compared to polysomnography. *Sleep.* 2013;36(11):1747–1755.
11. Van de Water ATM, Holmes A, Hurley DA. Objective measurements of sleep for non-laboratory settings as alternatives to polysomnography – a systematic review. *J Sleep Res.* 2011;20(1–2):183–200.
12. Lorenz, CP, Williams AJ. Sleep apps: what role do they play in clinical medicine? *Curr Opin Pulm Med.* 2017;23(6):512–516.
13. Toon E, Davey M, Hollis S, Nixon G, Horne R, Biggs S. Validation of two popular commercial devices for the assessment of sleep in children. *Sleep Med.* 2015;16(suppl 1):S29.
14. Choi BH, Seo JW, Choi JM, et al. Non-constraining sleep/wake monitoring system using bed actigraphy. *Med Bio Eng Comput.* 2007;45(1):107–114.
15. Walsh L, McLoone S, Ronda J, Duffy JF, Czeisler C. Noncontact pressure-based sleep/wake discrimination. *IEEE Trans Biomed Eng.* 2017;64(8);1750–1760.
16. Shambroom JR, Fábregas SE, Johnstone J. Validation of an automated wireless system to monitor sleep in healthy adults. *J Sleep Res.* 2012;21(2):221–230.
17. Griessenberger H, Heib DPJ, Kunz AB, Hoedlmoser K, Schabus M. Assessment of a wireless headband for automatic sleep scoring. *Sleep Breath.* 2013;17(2);747–752.
18. Berthomier C, Drouot X, Herman-Stoïca M, et al. Automatic analysis of single-channel sleep EEG: validation in healthy individuals. *Sleep.* 2007;30(11):1587–1595.
19. Senny F, Maury G, Cambron L, Leroux A, Destiné J, Poirrier R. The sleep/wake state scoring from mandible movement signal. *Sleep Breath.* 2012;16(2):535–542.
20. Beattie Z, Oyang Y, Statan A, et al. Estimation of sleep stages in a healthy adult population from optical plethysmography and accelerometer signals. *Physiol Meas.* 2017;38(11):1968–1979.
21. Hedner J, Pillar G, Pittman SD, Zou D, Grote L, White DP. A novel adaptive wrist actigraphy algorithm for sleep-wake assessment in sleep apnea patients. *Sleep.* 2004;27(8):1560–1566.
22. De Chazal P, Fox N, O'Hare E, et al. Sleep/wake measurement using a non-contact biomotion sensor. *J Sleep Res.* 2011;20(2):356–366.
23. O'Hare E, Flanagan D, Penzel T, Garcia C, Frohberg D, Heneghan C. A comparison of radio-frequency biomotion sensors and actigraphy versus polysomnography for the assessment of sleep in normal subjects. *Sleep Breath.* 2015;19(1):91–98.
24. Johns MW. A new method for measuring daytime sleepiness: the Epworth Sleepiness Scale. *Sleep.* 1991;14(6):540–545.
25. Klem GH, Lüders HO, Jasper HH, Elger C. The ten-twenty electrode system of the International Federation. The International Federation of Clinical Neurophysiology. *Electroencephalogr Clin Neurophysiol Suppl.* 1999;52:3–6.
26. Berry RB, Brooks R, Gamaldo CE, et al., for the American Academy of Sleep Medicine. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications.* Version 2.2. Darien, IL: American Academy of Sleep Medicine; 2015.
27. Zaffaroni A, Gahan L, Collins L, et al. Automated sleep staging classification using a non-contact biomotion sensor. Paper presented at: European Sleep Research Society Meeting; September 2014; Tallinn, Estonia.
28. Meltzer LJ, Montgomery-Downs HE, Insana SP, Walsh CM. Use of actigraphy for assessment in pediatric sleep research. *Sleep Med Rev.* 2012;16(5):463–475.
29. Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. *Statistician.* 1983;32(3):307–317.
30. Gallagher D, Heymsfield SB, Heo M, Jebb SA, Murgatroyd PR, Sakamoto Y. Healthy percentage body fat ranges: an approach for developing guidelines based on body mass index. *Am J Clin Nutr.* 2000;72(3):694–701.
31. Norman, RG, Pal, I, Stewart, C, Walsleben, JA, Rapoport, DM. Interobserver agreement among sleep scorers from different centers in a large dataset. *Sleep.* 2000;23(7):901–908.
32. 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.
33. Agnew HW, Webb WB, Williams RL. The first night effect: an EEG study of sleep. *Psychophysiology.* 1966;2(3):263–266.

ACKNOWLEDGMENTS

The authors thank the study participants for their contributions to the study.

SUBMISSION & CORRESPONDENCE INFORMATION

Submitted for publication August 22, 2018
Submitted in final revised form March 19, 2019
Accepted for publication April 1, 2019
Address correspondence to: Hawley E. Montgomery-Downs, PhD, 2218 Life Sciences Building, West Virginia University, Morgantown, WV 26506-6040; Email: hawley.montgomery-downs@mail.wvu.edu

DISCLOSURE STATEMENT

All authors have reviewed and approved this manuscript. Funding for this research was provided by a grant from ResMed, Inc. and included summer stipends for graduate and undergraduate research staff (Schade, Bauer, and Murray), two S+ devices, and funds for purchasing disposable polysomnography supplies. Dr. Montgomery-Downs received no compensation from ResMed, Inc. Gahan and Doheny were employed by ResMed Sensor Technologies at the time of the manuscript drafting and analysis. Kilroy and Zaffaroni are employees of ResMed Sensor Technologies.

EDITOR'S NOTE

The Emerging Technologies section focuses on new tools and techniques of potential utility in the diagnosis and management of any and all sleep disorders. The technologies may not yet be marketed, and indeed may only exist in prototype form. Some preliminary evidence of efficacy must be available, which can consist of small pilot studies or even data from animal studies, but definitive evidence of efficacy will not be required, and the submissions will be reviewed according to this standard. The intent is to alert readers of *Journal of Clinical Sleep Medicine* of promising technology that is in early stages of development. With this information, the reader may wish to (1) contact the author(s) in order to offer assistance in more definitive studies of the technology; (2) use the ideas underlying the technology to develop novel approaches of their own (with due respect for any patent issues); and (3) focus on subsequent publications involving the technology in order to determine when and if it is suitable for application to their own clinical practice. The *Journal of Clinical Sleep Medicine* and the American Academy of Sleep Medicine expressly do not endorse or represent that any of the technology described in the Emerging Technologies section has proven efficacy or effectiveness in the treatment of human disease, nor that any required regulatory approval has been obtained.