https://doi.org/10.5664/jcsm.8362

JCSM | Journal of Clinical Sleep Medicine

SCIENTIFIC INVESTIGATIONS

Validation of sleep measurement in a multisensor consumer grade wearable device in healthy young adults

Jennifer C. Kanady, PhD[1,2]; Leslie Ruoff, BS[1]; Laura D. Straus, PhD[1,2]; Jonathan Varbel, BA[1]; Thomas Metzler, MA[1]; Anne Richards, MD[1,2]; Sabra S. Inslicht, PhD[1,2]; Aoife O'Donovan, PhD[1,2]; Jennifer Hlavin, MS[1]; Thomas C. Neylan, MD[1,2,3]

[1]San Francisco Veterans Affairs Health Care System, San Francisco, California; [2]Department of Psychiatry, University of California, San Francisco, California; [3]Department of Neurology, University of California, San Francisco, California

Study Objectives: Our objective was to examine the ability of a consumer-grade wearable device (Basis B1) with accelerometer and heart rate technology to assess sleep patterns compared with polysomnography (PSG) and research-grade actigraphy in healthy adults.

Methods: Eighteen adults underwent consecutive nights of sleep monitoring using Basis B1, actigraphy, and PSG; 40 nights were used in analyses. Discrepancies in gross sleep parameters and epoch-by-epoch agreements in sleep/wake classification were assessed.

Results: Basis B1 accuracy was 54.20 ± 8.20%, sensitivity was 98.90 ± 2.70%, and specificity was 8.10 ± 15.00%. Accuracy, sensitivity, and specificity for distinguishing between the different sleep stages were 60–72%, 48–62%, and 57–86%, respectively. Pearson correlations demonstrated strong associations between Basis B1 and PSG estimates of sleep onset latency and total sleep time; moderate associations for sleep efficiency, duration of light sleep, and duration of rapid eye movement sleep; and a weak association for duration of deep sleep. Basis B1 significantly overestimates total sleep time, sleep efficiency, and duration of light sleep and significantly underestimates wake after sleep onset and duration of deep sleep.

Conclusions: Basis B1 demonstrated utility for estimates of gross sleep parameters and performed similarly to actigraphy for estimates of total sleep time. Basis B1 specificity was poor, and Basis B1 is not useful for the assessment of wake. Basis B1 accuracy for sleep stages was better than chance but is not a suitable replacement for PSG assessment. Despite low cost, ease of use, and attractiveness for patients, consumer devices are not yet accurate or reliable enough to guide treatment decision making in clinical settings.

Keywords: actigraphy, consumer wearable, photoplethysmography, polysomnogrrahy, sleep tracker, validation

Citation: Kanady JC, Ruoff L, Straus LD, et al. Validation of sleep measurement in a multisensor consumer grade wearable device in healthy young adults. *J Clin Sleep Med.* 2020;16(6):917–924.

BRIEF SUMMARY

Current Knowledge/Study Rationale: The popularity of consumer-grade wearable devices has increased, and many individuals use fitness trackers and smart watches to record health behaviors, including sleep. In clinical settings, consumer devices are attractive to patients because of their ease of use and integration with mobile and other devices. Despite the widespread popularity of consumer-grade wearable devices, relatively few studies have examined the accuracy of their sleep/wake assessment.

Study Impact: The ability of consumer-grade wearable devices to accurately assess sleep would be invaluable to the clinical and research communities because wearable devices are less expensive and more accessible than traditional sleep assessment and can be worn for extended periods of time in an individual's natural environment.

INTRODUCTION

Survey data indicate that nearly 75% of US consumers have access to a device that monitors sleep and that consumers find sleep patterns to be the most interesting behavior to track.[1,2] Despite the popularity of tracking sleep with a consumer device, few studies have examined how accurately consumer devices distinguish between sleep and wake. The ability of consumer devices to track sleep patterns can potentially benefit the clinical and research community by allowing for the easy collection of larger sample sizes at a substantially lower cost than traditional polysomnography (PSG) and actigraphy. Additionally, in clinical settings, consumer devices are attractive to patients because of their ease of use and integration with mobile and other devices. However, few validation studies have been performed, and the accuracy and reliability of these consumer devices are unclear. Thus, the goal of the present study was to compare sleep and wake measurements across a consumer wearable device, PSG, and actigraphy.

The first generation of consumer wearable devices introduced sleep measurement by way of accelerometry to the general public. Using proprietary algorithms, these devices provided estimates of sleep and wake using movement data. Validation studies comparing first-generation consumer wearable devices to PSG demonstrated that these devices performed similarly to research-grade actigraphy; they tend to overestimate sleep and underestimate wake.[3–9] Given these findings, consumer wearable devices may be a suitable replacement for traditional

research actigraphy when assessing gross sleep parameters. As actigraphy is the most widely used validated objective measure of sleep outside of the laboratory setting[10] and has been shown to be responsive to clinical interventions such as cognitive behavioral therapy for insomnia,[11] these results suggest consumer devices could also have the potential to provide helpful information about gross sleep parameters for clinical providers.

The more modern class of consumer wearable devices often includes the assessment of heart rate and heart rate variability in addition to accelerometry. The addition of heart rate and heart rate variability may not only increase the accuracy of sleep/wake distinction but may also allow for the measurement of different sleep stages. Heart rate and its variability vary as a function of sleep stage. For example, during non–rapid eye movement (REM) sleep, there is an increase in high-frequency power and a decrease in low-frequency power of heart rate variability; during REM sleep and wake, the opposite is true.[12–14] Studies using research-grade devices have demonstrated the benefits of heart rate and heart rate variability assessment for sleep classification.[12–14] However, it remains less clear whether the addition of heart rate and heart rate variability measurement adds to the accuracy of sleep detection in consumer wearable devices. Results from several studies suggest that consumer wearable devices with heart rate technology are comparable to actigraphy for the differentiation between sleep and wake, but accurate detection of sleep stages remains poor.[8,15,16]

The goal of the present study was to build on this literature by further examining the ability of a consumer wearable device with accelerometry and photoplethysmography (PPG)-based heart rate technology to accurately assess sleep patterns in healthy young adults. To achieve this goal, we assessed correspondence between a consumer wearable device with accelerometry and PPG (Basis B1), gold standard PSG, and research-grade actigraphy using both an automated scoring algorithm (ACT-auto) and a human-adjusted sleep period (ACT-human) in a sample of healthy young adults.

METHODS

Participants

Participants were recruited as part of a larger research study (clinicaltrials.gov identifier: NCT01243060). Eighteen clinically healthy, medication-free adults between the ages of 18 and 39 were included in the study (5 males, age: 26.8 ± 3.4 years). Participants had to be both medically and psychologically healthy as assessed by a clinical history, physical examination, and screening laboratory studies. Participants were also required to have a habitual bedtime between 2200 and 0000 hours and a habitual wake time between 0600 and 0800 hours in the last month to meet research diagnostic criteria for healthy sleep,[17] to have a body mass index of >18 or <28 kg/m^2, and be fluent in English. Exclusion criteria included the following: (1) a current psychiatric disorder, a lifetime history of any psychiatric disorder with psychotic features, or a history or alcohol or substance use disorder within the last 2 years as assessed by the Structured Clinical Interview for DSM-IV disorders[18]; (2) pregnancy; (3) neurologic disorders; (4) systemic illness affecting central nervous system function; (5) cardiovascular disease, hypertension, and/or high blood pressure; (6) current use of prescription or over-the-counter substances with psychoactive properties; (7) asthma or other reactive airways diseases; (8) self-reported history of caffeine use in excess of 400 mg/d on average; (9) working night shifts or extreme morning or evening tendencies; (10) sleep disorders as assessed by validated self-report measures[19–21] and PSG (participants with an apnea-hypopnea index ≥10 and/or ≥25 events/h were excluded); and (11) habitual long sleepers (>9 h/night) and short sleepers (<5 h/night).

Procedure

Data were collected as part of the larger research study referenced above. The Committee on Human Research at the University of California, San Francisco and the San Francisco Veterans Affairs Medical Center approved all study procedures. Before completing study procedures, informed consent was obtained.

Participants underwent 3 consecutive nights of simultaneous sleep monitoring using PSG, actigraphy (Micro Motionlogger Watch, Ambulatory Monitoring, Inc., Ardsley, NY), and a consumer wearable device (Basis B1, Basis Science, Inc., San Francisco, CA) in a controlled hospital setting at the University of California, San Francisco Parnassus Clinical Research Center. The actigraph and wearable device were placed side by side on the nondominant wrist. A total of 88 nights of concurrent PSG, actigraphy, and Basis B1 nights were collected. Before initiating data collection, Basis Science, Inc. agreed to provide us with 30-second epoch-by-epoch analyses from a random selection of 44 nights from these recordings. Basic Science, Inc. used the remainder of the data to further refine their algorithms. Following completion of data collection, Basic Science, Inc. provided us with a total of 41 nights of epoch-by-epoch analyses across 18 participants (8 participants with 3 nights of data, 7 participants with 2 nights of data, and 3 participants with 1 night of data); 3 nights were excluded because of irretrievable and/or corrupt data. For concordance analyses, the 3 devices were synced using PSG lights-out and lights-on times (eg, PSG lights out and lights on determined time in bed).

Basis B1

Each participant was fitted with a Basis B1 device, placed on the nondominant hand for the duration of the overnight recordings. Basis B1 devices consist of a sweat-proof, water-resistant wristband and are designed to be worn 24 h/d, 7 d/wk. The Basis B1 comes equipped with several sensors that provide PPG measurements of optical blood flow, 3-axis accelerometry, skin temperature, and galvanic skin response. Using proprietary algorithms, Basis B1 calculates sleep using input from both accelerometry and PPG heart rate technology. Basis Science, Inc. provided us with a 30-second epoch-by-epoch analysis of wake, light sleep (LS), deep sleep (DS), and REM sleep (RS) from Basis B1. These data were used in our epoch-by-epoch comparisons of sleep stage estimates across Basis B1 and PSG. For the epoch-by-epoch analysis of sleep versus wake, we collapsed Basis B1 epochs of LS, DS, and RS into a single all

sleep category. From these epoch-by-epoch data, we also calculated summary sleep variables of total sleep time (TST; number of minutes spent asleep after sleep onset), sleep onset latency (SOL; number of minutes to fall asleep after getting into bed), wake after sleep onset (WASO; number of minutes spent awake after sleep onset), sleep efficiency (SE; total sleep time divided by time in bed), duration of LS (TST-LS; number of minutes spent in LS following sleep onset), duration of DS (TST-DS; number of minutes in DS following sleep onset), and duration of RS (TST-RS; number of minutes in RS following sleep onset) using standard criteria within the field. Notably, some of the Basis B1 epochs were coded as unknown sleep (0.27% of total epochs), and some epochs contained no information and were left blank (1.06% of total epochs). These epochs were removed from statistical analyses.

Actigraphy

All participants were also fitted with a research-grade actigraph (Micro Motionlogger; Ambulatory Monitoring, Inc.), worn on the nondominant wrist throughout the duration of overnight recordings. Actigraphs were configured to collect data using the zero-crossing mode. Sleep was scored using the Cole-Kripke algorithm, and the down intervals (eg, time in bed) were established using PSG light-out and lights-on times. The sleep period (eg, sleep onset to sleep offset) was calculated in 2 different ways. The first was using the automatic down interval scoring algorithm provided by the software program, ActionW 2.7 (ACT-auto), which creates sleep periods during periods of consistent low activity when the participant appears to be sleeping. The second scoring method was a human-modified sleep period created by using information from activity output from the actigraph (ACT-human). More specifically, the start of the sleep period was modified to start at an activity level of 40 or less followed by at least 5 minutes of low activity and the end of the sleep period was modified to end at activity level of 200 or greater followed by at least 5 minutes of high activity.[10,22] Two independent and experienced raters created these human-modified sleep periods. When there were discrepancies between raters, the data were re-evaluated until a consensus was agreed on. The sleep variables examined from actigraphy were an epoch-by-epoch analysis of sleep versus wake and the sleep variables TST, SOL, WASO, and SE.

Polysomnography

Participants underwent 3 consecutive nights of ambulatory PSG sleep recording with the Embla Titanium (Natus Neurology, Middleton, WI) using the American Academy of Sleep Medicine standards.[23] The montage included scalp electroencephalogram electrodes (F3, F4, C3, C4, O1, O2) referenced to contralateral mastoids, 2 electrooculogram electrodes, 3 submental muscle tone electromyogram (EMG) electrodes, and 2-point electrocardiogram (EKG) electrodes. Electroencephalogram, EMG, and electrooculogram signals were sampled at 256 Hz; electroencephalogram and electrooculogram were filtered at 0.3–35 Hz; and EMG was filtered at 10–100 Hz. The first night of PSG recording included additional sensors, specifically 4 bilateral EMG sensors on the anterior tibialis, 2 respiratory effort belts using inductance plethysmography on the chest and abdomen, a thermistor and nasal cannula pressure transducer, and a pulse oximeter for detection of oxygen desaturation events, to screen for potential sleep apnea, periodic leg movement disorder, and/or restless legs syndrome. Two registered polysomnographic technologists visually scored the PSG data in 30-second epochs according to standard criteria.[23] To compare Basis B1 staging data of LS, DS, and RS to PSG staging data, PSG sleep stages N1 and N2 were combined into LS, PSG-measured stage 3 sleep was considered DS, and PSG-measured REM sleep was considered RS. To conduct an epoch-by-epoch comparison of PSG and actigraphy, 30-second epochs in PSG were time locked and transformed to 1-minute epochs. If a 1-minute epoch consisted of both sleep and wake, we scored the epoch as wake. The other PSG sleep outcome variables included TST, SOL, WASO, SE, TST-LS, TST-DS, and TST-RS. All available nights of data, including the screening night, were included in the analyses to increase the power of the study.

Statistical analyses

Consistent with previous validation studies,[3,5,7] we synchronized Basis B1, actigraphy, and PSG data collection by initializing the actigraph and Basis B1 on the same computer as PSG acquisition. Basis Science, Inc. provided us with a docking station that allowed the Basis B1 to connect to the PSG acquisition computer via USB. A total of 40 nights was used in data analyses comparing Basis B1 and PSG. We excluded 1 night of data because the PSG recording ended prematurely. The final selection of data included 3 nights of data from 8 participants, 2 nights of data from 6 participants, and 1 night of data from 4 participants (totaling 40 nights across 18 participants). Of the 40 nights, 37 nights were used to compare Basis B1, PSG, and actigraphy because a single participant declined to wear both wrist devices at the same time for all 3 nights.

Accuracy, sensitivity, and specificity

An epoch-by-epoch analysis for the time in bed period was conducted to determine global sensitivity and specificity of the Basis B1 device compared with PSG, ACT-auto compared with PSG, and ACT-human compared with PSG. To statistically account for the clustering of epochs within nights of sleep from the same participant, we used a mixed effect logistic regression model with random night effects to estimate sensitivity and specificity.[24] This was done by calculating the probability of an epoch being scored as sleep by Basis B1 or actigraphy given that it was scored as sleep by PSG (sensitivity) and the probability of an epoch being scored as wake by Basis B1 or actigraphy given that it was scored as wake by PSG (specificity). Other validation studies have often also included an assessment of accuracy (the agreement rate between PSG and Basis B1 or PSG and actigraphy, calculated as [true positive + true negative]/[true positive + true negative + false negative + false positive]). We argue that accuracy assessments can often be misleading and often indicate that devices are performing better than they actually are because of accuracy being calculated during a sleep period when a person is asleep for most of the time (eg, if the device reports sleep 100% of the time, the device will have high accuracy). Thus, we included an assessment of area under the curve (AUC) to measure device accuracy. An AUC of

1 indicates perfect agreement; an AUC of 0.50 indicates no better than chance. We also calculated sensitivity, specificity, and accuracy (AUC) for LS, DS, and RS sleep to examine the ability of Basis B1 to determine epoch-by-epoch sleep stages compared with PSG. Because these calculations require binary variables, we collapsed these different states into 2 categories: (1) the sleep stage of interest and (2) everything else (eg, other sleep stages and wake).

Basis B1, actigraphy, and PSG sleep variables

Pearson correlations were used to compare means for PSG-recorded TST, SOL, WASO, and SE to Basis B1, ACT-auto, and ACT-human measurements of these constructs. Pearson correlations were also used to examine correspondence between Basis B1 and PSG means for TST-LS, TST-DS, and TST-RS.

The Bland-Altman

The agreement between Basis B1 and PSG sleep parameter estimates was calculated using the Bland-Altman technique.[25,26] We used the Bland-Altman technique to plot the difference between Basis B1 and PSG (Basis B1 minus PSG) against the gold standard PSG measurement for each sleep variable (ie, TST, SOL, WASO, SE, TST-LS, TST-DS, TST-RS) to determine whether there was a bias in Basis B1. The mean difference, standard deviation, 95% confidence intervals of the bias, and lower and upper agreement limits were calculated for each sleep variable. The Basis B1 bias is represented as the mean difference between Basis B1 and PSG, with a negative mean difference representing an underestimation and a positive mean difference representing an overestimation. Bland-Altman analyses were also calculated to compare agreement between PSG and ACT-auto and PSG and ACT-human for TST, SOL, WASO, and SE using the same methods highlighted above.

RESULTS

Accuracy, sensitivity, and specificity

Results from the global epoch-by-epoch analyses comparing Basis B1, ACT-auto, and ACT-human to PSG are presented in **Table 1**. Global accuracy of the Basis B1 device for distinguishing all sleep and wake epochs was only slightly better than chance at 54.2% and was approximately 10% less than the accuracy of ACT-auto (64%) and ACT-human (66%). Global sensitivity of the Basis B1 was high (99%) and was comparable to sensitivity values for ACT-auto and ACT-human. Global specificity of the Basis B1 was extremely low at 8.10% and was substantially lower than actigraphy specificity (29.8–30.1%)

Results from the epoch-by-epoch analyses between Basis B1 and PSG assessments of LS, DS, and RS sleep are presented in **Table 2**. Basis B1 accuracy values for LS, DS, and RS were modest and ranged from 60 to 72%. Accuracy was highest for RS (72%); followed by DS (68%), and then LS (60%). Basis B1 sensitivity values for the different sleep stages fell within the poor to moderate range, with the highest sensitivity ratings for LS (62.10%), followed by RS (57.90%), and then DS (48.80%). Specificity was high for RS and DS at 85.90% and 86.70%, respectively. Specificity for LS was poor at 57.50%.

Basis B1, actigraphy, and PSG sleep variables

Results from Pearson correlations examining associations between PSG and Basis B1, PSG and ACT-auto, and PSG and ACT-human sleep variable estimates are presented in **Table 3**. Pearson correlations demonstrated a strong association between Basis B1 and PSG estimates of TST ($r = .82$) and SOL ($r = .71$), weak-to-moderate associations for estimates of SE ($r = .27$) and TST-LS ($r = .29$), and weak associations for TST-RS ($r = .19$), TST-DS ($r = .15$), and WASO ($r = .01$). The associations between Basis B1 and PSG estimates of TST and SOL are comparable to the association between PSG and ACT-auto and PSG and ACT-human. Associations between Basis B1 and PSG estimates of WASO and SE are numerically weaker than the associations between actigraphy and PSG for the same sleep variables.

The Bland-Altman

Table 4 contains means and standard deviations for the sleep variables of interest across devices and results from Bland-Altman analyses. Bland-Altman analyses revealed that Basis B1 significantly overestimated TST (mean difference = 15.11 minutes, $P < .01$), SE (mean difference = 4.00%, $P < .01$), and TST-LS (mean difference = 16.14 minutes, $P = .01$) and significantly underestimated WASO (mean difference = −15.11 minutes, $P < .01$). Although not statistically significant, Basis B1 also underestimated TST-DS (mean difference = −8.64 minutes, $P = .11$) and SOL (mean difference = −2.28 minutes, $P = .12$) and overestimated RS (mean difference = 5.18 minutes, $P = .20$). These findings parallel biases found when using actigraphy. Both ACT-auto and ACT-human significantly overestimated TST and SE and underestimated WASO compared with PSG. Complementary Bland-Altman plots for Basis B1 and PSG can be found in **Figure 1**.

Table 1—Epoch-by-epoch sleep-wake classification agreement between Basis B1 and PSG and actigraphy and PSG.

Global Sleep/Wake Assessment	Basis B1 vs PSG	ACT-auto vs PSG	ACT-human vs PSG
Global accuracy (AUC; % ± SD)	54.2 ± 8.2%	63.7 ± 12.8%	65.8 ± 15.1%
Global sensitivity (% ± SD)	98.9 ± 2.7%	98.6 ± 1.1%	98.6 ± 1.2%
Global specificity (% ± SD)	8.1 ± 15.0%	30.1 ± 26.1%	29.8 ± 26.8%

Accuracy is represented by AUC; sensitivity is the ability to detect true sleep compared with PSG; and specificity is the ability to detect true wake compared with PSG. ACT-auto = actigraphy automatically scored by the software program, ACT-human = human-adjusted actigraphy scoring, AUC = area under the curve, PSG = polysomnography.

Table 2—Accuracy, sensitivity, and specificity percentages for Basis B1 measurements of LS, DS, and RS compared with PSG.

Basis B1 Sleep Stages	Accuracy (AUC) (% ± SD)	Sensitivity (% ± SD)	Specificity (% ± SD)
Basis B1 LS	60.00 ± 5.50%	62.10 ± 7.40%	57.50 ± 7.50%
Basis B1 DS	67.80 ± 8.30%	48.80 ± 15.90%	86.70 ± 5.20%
Basis B1 RS	72.10 ± 9.00%	57.90 ± 14.10%	85.90 ± 4.60%

Accuracy is represented by AUC; sensitivity is the ability to detect the sleep stage of interest compared with PSG; and specificity is the ability to detect everything but the sleep stage of interest compared with PSG. AUC = area under the curve, DS = deep sleep, LS = light sleep, PSG = polysomnography, RS = rapid eye movement sleep.

Table 3—Correlation matrix demonstrating correspondence of Basis B1, ACT-auto, and ACT-human assessment of sleep variables with PSG assessment of the same sleep variables.

PSG Sleep Variables	Basis B1	ACT-auto	ACT-human
TST	*r* = .82, 95% CI [0.69 to 0.91]	*r* = .83, 95% CI [0.68 to 0.91]	*r* = .87, 95% CI [0.75 to 0.93]
SOL	*r* = .71, 95% CI [0.50 to 0.84]	*r* = .72, 95% CI [0.50 to 0.85]	*r* = .67, 95% CI [0.42 to 0.82]
WASO	*r* = .01, 95% CI [−0.31 to 0.33]	*r* = .50, 95% CI [0.19 to 0.72]	*r* = .51, 95% CI [0.21 to 0.73]
SE	*r* = .27, 95% CI [−0.06 to 0.55]	*r* = .61, 95% CI [0.33 to 0.79]	*r* = .59, 95% CI [0.31 to 0.78]
TST-LS	*r* = .29, 95% CI [−0.04 to 0.56]		
TST-DS	*r* = .15, 95% CI [−0.18 to 0.45]		
TST-RS	*r* = .19, 95% CI [−0.14 to 0.48]		

CI = confidence interval, ACT-auto = actigraphy automatically scored by the software program, ACT-human = human-adjusted actigraphy scoring, SE = sleep efficiency, TST = total sleep time, TST-DS = duration of deep sleep, TST-LS = duration of light sleep, TST-RS = duration of REM sleep, WASO = wake after sleep onset.

Table 4—Sleep variable means across devices; sleep variable mean differences for Basis B1 vs PSG, ACT-auto vs PSG, and ACT-human vs PSG derived from Bland Altman; and a comparison of mean differences across devices.

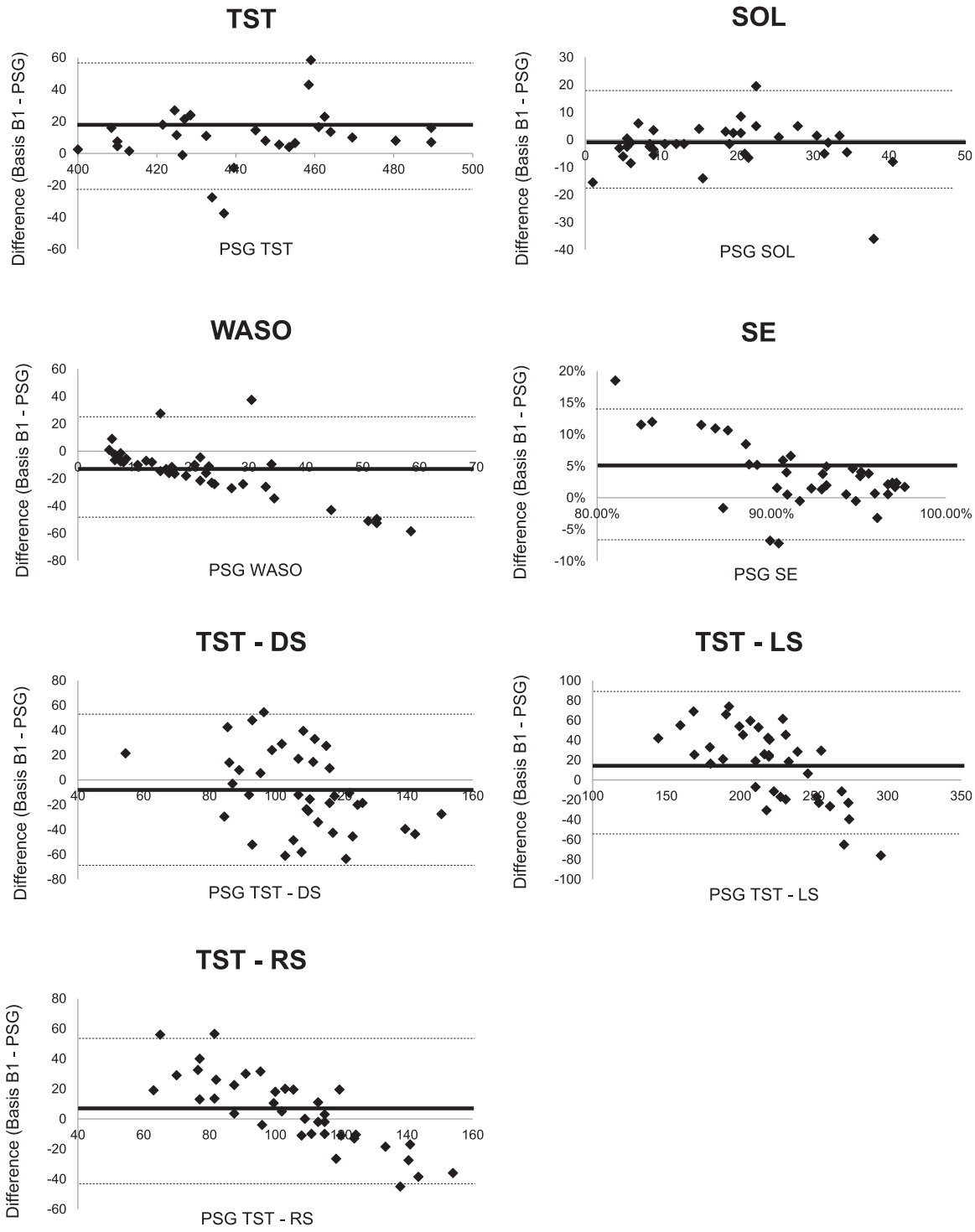
Gross Sleep Parameters	PSG (Mean ± SD)	Basis B1 (Mean ± SD)	ACT-auto (Mean ± SD)	ACT-human (Mean ± SD)	Mean Differences		
					Basis B1 vs PSG (Bias ± SD, [95% CI])	ACT-auto vs PSG (Bias ± SD, [95% CI])	ACT-human vs PSG (Bias ± SD, [95% CI])
TST	433.84 ± 31.80	448.95 ± 33.52	450.24 ± 35.50	448.10 ± 37.80	15.11 ± 19.42, [8.93, 21.28]	14.17 ± 16.22, [8.63, 19.70]	12.02 ± 15.47, [6.74, 17.29]
SOL	17.65 ± 10.70	15.36 ± 11.80	16.92 ± 9.14	17.76 ± 10.44	−2.28 ± 8.60, [−21.28, −8.93]	−0.37 ± 7.69, [−3.04, 2.29]	0.47 ± 8.83, [−2.56, 3.49]
WASO	21.88 ± 14.47	6.77 ± 13.10	7.18 ± 8.00	6.97 ± 7.89	−15.11 ± 19.42, [−21.28, −8.93]	−15.11 ± 13.06, [−19.65, −10.65]	−15.32 ± 12.92, [−19.73, −10.91]
SE	91.74 ± 4.34	95.37 ± 4.22	98.38 ± 1.86	98.42 ± 1.84	4.00 ± 5.00, [2.00, 5.00]	3.00 ± 4.00, [2.00, 5.00]	3.00 ± 4.00, {2.00, 4.00]
TST-LS	220.80 ± 34.75	236.95 ± 27.56			16.14 ± 37.62, [4.18, 28.11]		
TST-DS	107.86 ± 18.05	99.21 ± 30.17			−8.64 ± 32.78, [−19.07, 1.78]		
TST-RS	105.18 ± 23.15	110.37 ± 13.13			5.18 ± 24.38, [−2.57, 12.94]		

Lower limit and upper limit, lower and upper agreement (mean difference ± 1.96 SD). ACT-auto = actigraphy automatically scored by the software program, ACT-human = human-adjusted actigraphy scoring, CI = confidence interval, ACT-auto = actigraphy automatically scored by the software program, ACT-human = human-adjusted actigraphy scoring, SE = sleep efficiency, TST = total sleep time, TST-DS = duration of deep sleep, TST-LS = duration of light sleep, TST-RS = duration of REM sleep, WASO = wake after sleep onset.

DISCUSSION

The goal of the present study was to determine the ability of a multisensor consumer wearable device (Basis B1) with accelerometer and PPG technology to accurately assess sleep patterns in a sample of healthy young adults. Results from this study demonstrated that Basis B1 showed some utility for estimates of gross sleep parameters and performed similarly to actigraphy for estimates of total sleep time. However, the specificity of Basis B1 was extremely poor, suggesting that

Figure 1—Bland-Altman plots for TST, SOL, WASO, SE, TST-LS, TST-DS, and TST-S recorded by Basis B1 and PSG.



Mean bias of the differences between Basis B1 and PSG outcomes are demonstrated by a solid black line, and lower and upper agreement limits (mean difference ± 1.96 SD) are demonstrated by dotted lines for each Bland-Altman plot. PSG = polysomnography, SE = sleep efficiency, SOL = sleep onset latency, TST = total sleep time, TST-DS = duration of deep sleep, TST-LS = duration of light sleep, TST-S = duration of REM sleep, WASO = wake after sleep onset.

Basis B1 is not able to provide accurate measurements of wake during sleep periods. Given that Basis B1 and actigraphy performed similarly, the addition of PPG technology does not seem to improve the accuracy of sleep and wake detection, at least in this particular consumer wearable device. In terms of distinguishing between the different epoch-by-epoch sleep stages, the Basis B1 performed similarly to other wearable devices using PPG technology[8] and performed better than chance. However,

Basis B1, a consumer wearable device with accelerometry and PPG technology, is not a suitable replacement for gold standard PSG assessment in clinical and research settings.

Epoch-by-epoch analyses revealed that Basis B1 has excellent sensitivity (98.9%), only slightly better than chance accuracy (54.2%), and extremely poor specificity (8.1%). These results demonstrate that Basis B1 can accurately identify sleep epochs, but its ability to correctly identify wake epochs during the sleep period in our sample of healthy sleepers was extremely poor. ACT-auto and ACT-human specificity values were also poor at 30.1% and 29.8%, respectively. However, these values are substantially higher than specificity of Basis B1. Therefore, research-grade actigraphy is the better choice when trying to capture wake during sleep periods. Notably, accuracy, sensitivity, and specificity values are likely influenced by the fact that we conducted epoch-by-epoch analyses only for defined sleep periods. These values would likely change when examining epoch-by-epoch comparisons across a 24-hour day when an individual is not asleep for most of the time. To our knowledge, no studies have conducted these types of analyses using PSG as the gold standard comparison.

In terms of the accuracy of Basis B1 to correctly identify the different sleep stages, Basis B1 performed better than chance. Sensitivity was highest for light sleep at 62.1% and is comparable to previous studies using heart rate variability technology.[8,14] This result may be partially influenced by the fact that most of the sleep period consists of LS and thus has more opportunity to be correctly identified. Sensitivity values for DS and RS fell within the poor range at 48.8% and 57.9%, respectively. Results from previous studies examining sensitivity of devices with heart rate variability sensors for the detection of DS and RS vary. For example, Fonseca et al[14] found a deep sleep sensitivity value of 63.9% and REM sensitivity value of 70.7% in a sample of healthy middle-aged adults, which are higher than the sensitivity values reported here. Cook et al[8] found poorer sensitivity values than reported here (deep sleep sensitivity of 49% and REM sensitivity of 30%) in individuals with excessive daytime sleepiness. These mixed findings demonstrate that performance of devices that measure heart rate variability can vary depending on device technology, the algorithms used, and the populations studied.

Basis B1 performed better when estimating gross sleep parameters. Associations between Basis B1 and PSG estimates of TST (r = .82) and SOL (r = .71) were strong and were comparable to associations between PSG and actigraphy for the same sleep variables. The association between Basis B1 and PSG estimates of SE was weak to moderate (r = .27) and was weaker than the association between PSG and actigraphy for estimates of SE. This is likely driven by Basis B1 measurement of WASO. The association between PSG and Basis B1 WASO was extremely weak (r = .01), which undoubtedly influenced estimates of SE. Bland-Altman analyses demonstrated that Basis B1 significantly overestimated TST by an average of 15.11 minutes and SE by an average of 4% and significantly underestimated WASO by an average of 15.11 minutes. Parallel findings were found for ACT-auto and ACT-human. Taken together, these results suggest that Basis B1 may have some usefulness for the assessment of gross sleep parameters, with the exception of WASO.

Results from analyses examining Basis B1 measurement of WASO are interesting. Mean WASO derived from Basis B1 (6.77 minutes) was not significantly different than mean WASO from ACT-auto (7.18 minutes) and ACT-human (6.97 minutes). However, specificity (eg, the ability to identify wake epochs during the sleep period) of Basis B1 was extremely low (8.1%), and the association between Basis B1 WASO and PSG was extremely weak (r = .01) and remarkably lower than associations between PSG and actigraphy assessments of WASO. These results demonstrate that Basis B1 is estimating a similar amount of WASO as actigraphy but is not correctly identifying the specific wake epochs (eg, when the WASO occurs). One possibility is that the Basis B1 is scoring sleep stages with more movement (eg, lighter sleep stages) as wake and is scoring motionless periods of wake as sleep. If the latter possibility is true, then the addition of PPG to a consumer wearable device is not improving measurement of ambivalent epochs that are not accurately captured by accelerometry (eg, sleep vs quiet wakefulness). Furthermore, the inability of the Basis B1 to capture WASO indicates that multisensor consumer wearable devices are not yet appropriate for sleep assessment in clinical populations (eg, insomnia, sleep apnea).

There are several limitations to this study. First, the sample consisted of 18 healthy adults. Previous validation studies using actigraphy and consumer wearable devices have demonstrated that these devices perform best in healthy, young sleepers.[9,27] Thus, results from this study are likely not generalizable to other populations (eg, older adults, individuals with insomnia). Second, because Basis Science, Inc. provided us with a random selection of 41 nights, there was unbalanced data distribution across participants. Third, our results cannot extend to other consumer wearable devices because technology and algorithms likely differ across different manufacturers and models. This may partially explain the mixed findings regarding differences in accuracy, sensitivity, and specificity values across studies. Fourth, it is possible that the devices desynchronized across the night. However, we believe that this is highly unlikely, and our synchronization methods are similar to other validation studies.[3,5,7] Fifth, epoch-by-epoch comparisons were calculated for a defined sleep period, which contains mostly sleep and very little wake. Measurements of accuracy, sensitivity, and specificity would likely change if assessed across a 24-hour day (eg, the ability of a device to capture short daytime naps[28]).

The present study adds to the literature examining the ability of multisensory consumer wearable devices to accurately measure sleep and wake. Results from this study suggest that wearable devices with accelerometer and heart rate technology are not yet a suitable replacement for traditional actigraphy and PSG sleep assessment in clinical or research settings. Although the Basis B1 demonstrated utility for the assessment of gross sleep parameters, the inability of the device to capture wake after sleep onset is a huge limitation of this methodology. Despite low cost, ease of use, and attractiveness for patients, consumer devices such as the Basis B1 are not yet accurate or reliable enough to guide treatment decision making in clinical settings. Further studies are needed to validate the usefulness of consumer wearable devices for sleep assessment.

ABBREVIATIONS

ACT-auto, automated scoring algorithm for actigraphy
ACT-human, human-adjusted sleep period for actigraphy
AUC, area under the curve
DS, deep sleep
EKG, electrocardiogram
EMG, electromyogram
LS, light sleep
PPG, photoplethysmography
PSG, polysomnography
REM, rapid eye movement
RS, REM sleep
SE, sleep efficiency
SOL, sleep onset latency
TST, total sleep time
TST-DS, deep sleep total sleep time
TST-LS, light sleep total sleep time
TST-RS, REM sleep total sleep time
WASO, wake after sleep onset

REFERENCES

1. Ko PRT, 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.
2. Baron KG, Duffecy J, Berendsen MA, Cheung Mason I, Lattie EG, Manalo NC. Feeling validated yet? A scoping review of the use of consumer-targeted wearable and mobile technology to measure and improve sleep. *Sleep Med Rev*. 2018;40: 151–159.
3. 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.
4. de Zambotti M, Baker FC, Colrain IM. Validation of sleep-tracking technology compared with polysomnography in adolescents. *Sleep*. 2015;38(9):1461–1468.
5. Meltzer LJ, Hiruma LS, Avis K, Montgomery-Downs H, Valentin J. Comparison of a commercial accelerometer with polysomnography and actigraphy in children and adolescents. *Sleep*. 2015;38(8):1323–1330.
6. Montgomery-Downs HE, Insana SP, Bond JA. Movement toward a novel activity monitoring device. *Sleep Breath*. 2012;16(3):913–917.
7. Toon E, Davey MJ, Hollis SL, Nixon GM, Horne RSC, Biggs SN. Comparison of commercial wrist-based and smartphone accelerometers, actigraphy, and PSG in a clinical cohort of children and adolescents. *J Clin Sleep Med*. 2016;12(3):343–350.
8. Cook JD, Prairie ML, Plante DT. Ability of the multisensory jawbone UP3 to quantify and classify sleep in patients with suspected central disorders of hypersomnolence: a comparison against polysomnography and actigraphy. *J Clin Sleep Med*. 2018;14(5):841–848.
9. Kang SG, Kang JM, Ko KP, Park SC, Mariani S, Weng J. Validity of a commercial wearable sleep tracker in adult insomnia disorder patients and good sleepers. *J Psychosom Res*. 2017;97:38–44.
10. Ancoli-Israel S, Martin JL, Blackwell T, et al. The SBSM guide to actigraphy monitoring: clinical and research applications. *Behav Sleep Med*. 2015;13(suppl 1):S4–S38.
11. Manber R, Edinger JD, Gress JL, San Pedro-Salcedo MG, Kuo TF, Kalista T. Cognitive behavioral therapy for insomnia enhances depression outcome in patients with comorbid major depressive disorder and insomnia. *Sleep*. 2008;31(4):489–495.
12. Yuda E, Yoshida Y, Sasanabe R, Tanaka H, Shiomi T, Hayano J. Sleep stage classification by combination of actigraphic and heart rate signals. *J Low Power Electron Appl*. 2017;7(4):28–33.
13. Muzet A, Werner S, Fuchs G, et al. Assessing sleep architecture and continuity measures through the analysis of heart rate and wrist movement recordings in healthy subjects: comparison with results based on polysomnography. *Sleep Med*. 2016;21:47–56.
14. Fonseca P, Weysen T, Goelema MS, et al. Validation of photoplethysmography-based sleep staging compared with polysomnography in healthy middle-aged adults. *Sleep*. 2017;40(7):zsx097.
15. de Zambotti M, Baker FC, Willoughby AR, et al. Measures of sleep and cardiac functioning during sleep using a multi-sensory commercially-available wristband in adolescents. *Physiol Behav*. 2016;158:143–149.
16. Lee XK, Chee NIYN, Ong JL, et al. Validation of a consumer sleep wearable device with actigraphy and polysomnography in adolescents across sleep opportunity manipulations. *J Clin Sleep Med*. 2019;15(9):1337–1346.
17. Edinger JD, Bonnet MH, Bootzin RR, et al. Derivation of research diagnostic criteria for insomnia: report of an American Academy of Sleep Medicine Work Group. *Sleep*. 2004;27(8):1567–1596.
18. First MB, Spitzer RL, Gibbon M, Williams JBWS. *Structured Clinical Interview for DSM-IV-TR Axis I Disorders, Research Version*. New York, NY: Biometrics Research, New York State Psychiatric Institute; 2002.
19. Heinzer R, Andries D, Bastardot F, et al. Berlin questionnaire performance for detecting sleep apnea in the general population. *Eur Respir J*. 2011;38(Suppl 55):4957.
20. Senaratna CV, Perret JL, Matheson MC, et al. Validity of the Berlin questionnaire in detecting obstructive sleep apnea: a systematic review and meta-analysis. *Sleep Med Rev*. 2017;36:116–124.
21. Varghese B. Identification of risk for obstructive sleep apnea by Berlin questionnaire. *Res J Pharm Biol. Chem. Sci*. 2011;2(4):1035–1040.
22. Littner M, Kushida CA, Anderson WMD, et al. Practice parameters for the role of actigraphy in the study of sleep and circadian rhythms: an update for 2002. *Sleep*. 2003;26(3):337–341.
23. Iber C, Ancoli-Israel S, Chesson AL Jr, Quan SF; for the American Academy of Sleep Medicine. *The AASM Manual for the Scoring of Sleep and Associates Events: Rules, Terminology and Technical Specifications* 1st ed. Westchester, IL: American Academy of Sleep Medicine; 2007.
24. Genders TSS, Spronk S, Stijnen T, Steyerberg EW, Lesaffre E, Hunink MGM. Methods for calculating sensitivity and specificity of clustered data: a tutorial. *Radiology*. 2012;265(3):910–916.
25. Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. *Lancet*. 1995;346(8982):1085–1087.
26. Bland JM, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;1(8476):307–310.
27. Lichstein K, Stone K, Donaldson J, et al. Actigraphy validation with insomnia. *Sleep*. 2006;29(2):232–239.
28. Kanady JC, Drummond SPA, Mednick SC. Actigraphic assessment of a polysomnographic-recorded nap: A validation study. *J Sleep Res*. 2011; 20(1 Pt 2):214–222.

ACKNOWLEDGMENTS

Author contributions: JH coordinated data collection. LR and TM collected and analyzed the data. JK interpreted the data and wrote the manuscript. All remaining authors provided edits to the manuscript and approved the final manuscript.

SUBMISSION & CORRESPONDENCE INFORMATION

Submitted for publication December 5, 2019
Submitted in final revised form January 31, 2020
Accepted for publication January 31, 2020
Address correspondence to: Jennifer C. Kanady, PhD, 4150 Clement Street, San Francisco, CA 94121; Tel: (510) 393-5935; Email: jen.kanady@gmail.com

DISCLOSURE STATEMENT

All authors have seen and approved the manuscript. The authors report Basis B1 watches provided by Basis Science, Inc. This study was funded by USAMRMC Grant W81XWH-09-2-0080. The authors report no conflicts of interest.