medRxiv preprint doi: https://doi.org/10.1101/2024.09.10.24313425; this version posted September 11, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

[Title Page]

Title: Performance Evaluation of the Verily Numetric Watch sleep suite for digital sleep assessment against in-lab polysomnography

Authors: Benjamin W. Nelson, PhD[1, 2*]; Sohrab Saeb, PhD[1*]; Poulami Barman, MS[1]; Nishant Verma, PhD[1]; Hannah Allen, BS[1]; Massimiliano de Zambotti, PhD[3]; Fiona C. Baker, PhD[3]; Nicole Arra, BA[3]; Niranjan Sridhar, PhD[1]; Shannon S. Sullivan, MD, MSc[1, 4]; Scooter Plowman, MD, MBA, MHSA, MSc[1]; Erin Rainaldi, MS[1]; Ritu Kapur, PhD[1, 5]; Sooyoon Shin, PhD[1]

Affiliations:

[1]Verily Life Sciences, South San Francisco, CA, United States

[2]Division of Digital Psychiatry, Department of Psychiatry, Harvard Medical School and Beth Israel Deaconess Medical Center, Boston, MA

[3]Center for Health Sciences, SRI International, Menlo Park, California, USA

[4]Division of Pulmonary, Asthma, and Sleep Medicine, Department of Pediatrics, Stanford University School of Medicine, Palo Alto, California

[5]Department of Neurology, Radboud UMC

*These authors contributed equally to this work

Corresponding Author:

Benjamin W. Nelson, PhD

Verily Life Sciences

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.09.10.24313425; this version posted September 11, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

269 E Grand Ave, South San Francisco, CA 94080

Email: bwn@verily.com

Phone: 650-495-7100

Declaration of conflicts of interest: BWN, SSaeb, PB, NV, HA, NS, SSSullivan, SP, ER, RK, SShin report employment and equity ownership in Verily Life Sciences. In addition, SSaeb and NV are listed inventors in a pending patent broadly relevant to the work. MDZ, FB and NA received institutional research funding from Verily Life Sciences for study execution.

Funding: This study was funded by Verily Life Sciences

Data sharing statement:

Data from this study are not available due to the nature of this program. Participants did not consent for their data to be shared publicly.

Counts:

Word count: Abstract, 249. Overall, 3320 (limit, 5K)

Reference count: 35

Table count: 5

Figure count: 2

Supplemental file count: 16 pages

Keywords: digital measures; mHealth; polysomnography; PSG; sleep; wearable technology; Verily Numetric Watch

medRxiv preprint doi: https://doi.org/10.1101/2024.09.10.24313425; this version posted September 11, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

Abstract

The goal was to evaluate the performance of a multi-sensor wrist-worn wearable device for generating 12 sleep measures in a diverse cohort. Our study technology was the sleep suite of the Verily Numetric Watch (VNW), using polysomnography (PSG) as reference during 1-night simultaneous recording in a sample of N=41 (18 male, age range: 18-78 years). We performed epoch-by-epoch comparisons for all measures. Key specific analyses were: core accuracy metrics for sleep vs wake classification; bias for continuous measures (Bland-Altman); Cohen's kappa and accuracy for sleep stage classifications; and mean count difference and linearly weighted Cohen's kappa for count metric. In addition, we performed subgroup analyses by sex, age, skin tone, body mass index, and arm hair density. Sensitivity and specificity (95% CI) of sleep versus wake classification were 0.97 (0.96, 0.98) and 0.66 (0.61, 0.71), respectively. Mean total sleep time bias was 14.55 minutes (1.61, 27.16); wake after sleep onset, -11.77 minutes (-23.89, 1.09); sleep efficiency, 3.15% (0.68, 5.57); sleep onset latency, -3.24 minutes (-9.38, 3.57); light-sleep duration, 3.78 minutes (-7.04, 15.06); deep-sleep duration, 3.91 minutes (-4.59, 12.60); rapid eye movement-sleep duration, 6.94 minutes (0.57, 13.04). Median difference for number of awakenings, 0.00 (0.00, 1.00); and overall accuracy of sleep stage classification, 0.78 (0.51, 0.88). Most measures showed statistically significant proportional biases and/or heteroscedasticity. Subgroup results appeared largely consistent with the overall group, although small samples preclude strong conclusions. These results support the use of VNW's in classifying sleep versus wake, sleep stages, and for related overnight sleep measures.

medRxiv preprint doi: https://doi.org/10.1101/2024.09.10.24313425; this version posted September 11, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

Introduction

Sleep behaviors may provide important insights into both mental and physical health status, as well as therapeutic and disease outcomes. Sleep disturbance constitutes a meaningful aspect of health, and inadequate sleep is associated with increased risk for major depressive disorder, cardiovascular disease, cancer, and diabetes.[1,2]

However, the traditional reference standard for measuring sleep staging, polysomnography (PSG), is resource-intensive and in limited supply. Measuring sleep via PSG imparts high burden and cost, and for many purposes, may diminish ecological validity.[3] The emergence of user-friendly, low-burden, yet highly accurate and reliable digital technologies is a promising development. These technologies may open the door to long-term sleep monitoring, allowing for evaluation of changes in sleep over time (e.g., regularity, variance patterns, and trends over days, weeks, or months). Such methods may also facilitate the wider evaluation of sleep-related outcomes across multiple diseases where they remain under-studied despite their importance.

Wearable sensor-based sleep-tracking technology may lead to important advances towards these goals by providing accessible, convenient, long-term, free-living sleep behavior monitoring.[4] However, the responsible deployment of these devices demands proper performance evaluation studies to ensure their accuracy and generalizability across demographic groups, [5-7] and to mitigate any potential biases that could lead to health disparities.[8-15]

Following the widespread adoption of sleep-tracking technology by the general public and the subsequent emergence of reports in the clinical literature,[16–18] device performance evaluation studies have proliferated in recent years,[5-7] although concerns remain about adoption for diagnosis and treatment.[19-21] Those studies have shown that, in general, latest-generation wearable devices produce suitable estimates of sleep measures, although these vary by device

medRxiv preprint doi: https://doi.org/10.1101/2024.09.10.24313425; this version posted September 11, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

and firmware version and their capabilities to classify sleep stages are less sufficient.[22,23] Despite growing research to understand overall device performance, investigations about generalizability across diverse populations, including sex, age, skin tone, body mass index (BMI), and arm hair density are still lacking.[4]

We undertook this study in order to evaluate the analytic performance of sleep tracking by the Verily Numetric Watch (VNW), a wrist-worn device that classifies every 30-second epoch into one of the following 4 sleep stages: wake, light sleep, deep sleep, and rapid eye movement (REM) sleep, allowing for the derivation of a number of clinically meaningful overnight sleep measures including total sleep time (TST); wake after sleep onset (WASO); sleep efficiency (SE); sleep onset latency (SOL); number of awakenings (NAWK); and light, deep, and REM sleep duration.

Our objectives were to evaluate the VNW's performance for epoch-by-epoch sleep versus wake classification; 4-stage sleep classification (wake, light, deep, REM); and TST, WASO, SE, SOL, NAWK, and sleep stage duration in a diverse sample of sleepers, compared to PSG-derived labels. In addition, the exploratory objectives included performing subgroup analyses based on sex, age, skin tone, BMI, and arm hair density.

medRxiv preprint doi: https://doi.org/10.1101/2024.09.10.24313425; this version posted September 11, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

Methods

Sample

This study included 41 participants. Eligible participants were between 18-80 years of age and agreed to abstain from caffeine, nicotine, alcohol, and cannabis products for eight hours prior to the overnight lab visit and until the visit was complete. They also had to agree to abstain from medications that may affect sleep/wakefulness for 24 hours prior to the lab visit and during the study, unless the medications were taken on a routine basis and had approval from the study team. In order to be eligible, participants had to be considered "typical sleepers," based on the following: obstructive sleep apnea 50 (OSA-50)[24] scores < 5; insomnia severity index (ISI)[25] < 8; Epworth Sleepiness Scale (ESS) scores < 10; no evidence of sleep-disordered breathing at the PSG evaluation; and apnea-hypopnea index (AHI)[26] threshold of 5 defining hypopnea as ≥ 30% reduction in airflow for ≥ 10 seconds associated with a ≥ 3% decrease in oxygen saturation or an arousal).[27] Individuals were ineligible if they had a major medical or psychiatric condition, if they used supplemental oxygen, or were unwilling to cease use of therapy, such as continuous positive airway pressure or oral appliance for sleep-disordered breathing during the visit. Additional exclusion criteria included use of medications that affect sleep (e.g., hypnotics or antidepressants) or any sleep medications in the previous 24 hours; pregnancy, lactation, or breastfeeding; having an implantable medical device; night-shift work; or travel over 3 time zones within two weeks prior to the study. The study was approved by the WCG Institutional Review Board (20215892) and was conducted in accordance with the Declaration of Helsinki. All participants provided informed consent, and the study was registered at clinicaltrials.gov (NCT05276362).

medRxiv preprint doi: https://doi.org/10.1101/2024.09.10.24313425; this version posted September 11, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

Focus method/technology: Verily Numetric Watch

Participants wore the VNW wrist-worn device on the non-dominant wrist and this device was equipped with two sensors: a photoplethysmography (PPG) sensor with a sampling rate of 60 Hz and a 3-axis accelerometer with a sampling rate of 104 Hz.

The PPG sensor consists of a green light emitter diode and two PPG signal channels.

Using the PPG and accelerometer signals, the VNW classifies every 30-second epoch into one of the following 4 classes: wake, light sleep, deep sleep, and REM sleep.

The sleep staging algorithm consists of a deep convolutional neural network that was initially trained using 10,000 nights of data from the Sleep Heart Health Study (SHHS) and Multi-Ethnic Study of Atherosclerosis (MESA) public datasets,[28] and tuned on a previous generation of the VNW with a dataset collected at SRI.

Reference Labels

Standard laboratory PSG sleep assessment including electroencephalography (EEG; F3/4, C3/4, O1/2 referred to the contralateral mastoid; 256 Hz sampled), submental electromyography and bilateral electrooculography was performed according to the American Academy of Sleep Medicine (AASM) guidelines.[29] Leg movement (bilateral anterior tibialis), electrocardiography (ECG), respiratory (thoracic and abdominal piezoelectric bands, nasal cannula and thermistor), and oxygen saturation (pulse oximeter) signals were also collected and used to confirm the absence of sleep disordered breathing. All recordings were performed using the Compumedics Grael® HD-PSG system (Compumedics, Abbotsford, Victoria, Australia). Sleep scoring (wake, N1, N2, N3, REM) was performed by two independent scorers according to the AASM rules. Inter-rater reliability (Kappa) was 91%. Discrepancies were resolved by a third scorer.

medRxiv preprint doi: https://doi.org/10.1101/2024.09.10.24313425; this version posted September 11, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

To make the labels consistent with the VNW sleep suite, categories N1 and N2 light sleep were combined into a single "light sleep" category.[19]

Design, study setting, and procedures

This was a single-arm, observational study to evaluate the performance of the sleep measures from the VNW against PSG-derived labels (as reference). Data were collected during a single overnight stay in a sleep laboratory from a diverse sample of sleepers without elevated insomnia symptoms or OSA. All study protocols and procedures were conducted at a single site (SRI; Menlo Park, California).

Recruitment and Phone Screen

Potential participants were recruited through fliers, an existing site participant database, and postings on public websites. Study personnel pre-screened potential participants for eligibility, via phone and online screen questionnaires.

In-Lab/Remote Screening and Enrollment Visit

After checking verbal interest and eligibility during a phone pre-screening, participants were invited to an in-lab or remote screening visit to sign an informed consent. Study personnel collected demographic, clinical, and other relevant information including skin tone and arm hair density. Candidates whose eligibility was confirmed in the in-lab/remote screening visit were scheduled for the in-lab overnight visit. Screen failures (defined as those who consented to participate but did not meet one or more eligibility criteria) were not entered in the study.
The questionnaires completed by participants or study personnel and used for screening were OSA50,[24] AHI,[26] ISI,[25] ESS,[30] Fitzpatrick Skin Scale,[31] and Arm Hair Index (see Supplement).

medRxiv preprint doi: https://doi.org/10.1101/2024.09.10.24313425; this version posted September 11, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

In-Lab Overnight Visit

Participants slept in comfortable sound-proof and temperature-controlled bedrooms where they were able to go to bed and wake up at their preferred times. During this visit, standard PSG protocols were used for preparation, recording procedures, and instrument calibration; and participants were outfitted with the VNW on their non-dominant wrist (see Supplement for additional information).

Statistical Analysis

All analyses were conducted between lights off and lights on and structured to evaluate the performance of the VNW against the PSG reference, following published scientific recommendations for performance evaluation.[7] The unit of analysis was the 30-second epoch. To evaluate the endpoints, all epochs with data from PSG and VNW from lights-out to lights-on were included in analyses.

Our core analysis included evaluation of sensitivity, specificity, negative predictive value (NPV) and positive predictive value (PPV) of sleep versus wake classification. Estimates of sensitivity, specificity, PPV, and NPV were obtained using all epochs that were non-missing in both devices. We accounted for clustering of epochs within an individual using logistic mixed-effect regression models, with subject added as random effect. 95% CIs were calculated using clustered bootstrap method.[32] Additionally, we evaluated classification of different sleep stages (light, deep, REM) and derived sleep measures TST, WASO, SE, SOL, NAWK, and duration of different sleep stages (light, deep, REM) as part of the core analysis.

VNW's performance for derived sleep measurements, at each participant level, were calculated using all epochs from lights-off to lights-on, based on existing performance testing standardization frameworks.[19,33] For these measurements, we performed Bland Altman analyses, estimating the mean bias and lower and upper limits of agreement (with their 95%

medRxiv preprint doi: https://doi.org/10.1101/2024.09.10.24313425; this version posted September 11, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

CIs), testing for assumptions of proportional bias, heteroscedasticity, and normality. 95% percentile bootstrap CIs are reported for TST, WASO, SE, SOL, NAWK, sleep stage duration.

We evaluated the VNW's measurements of NAWKs, using mean and median difference in counts, and linearly weighted Cohen's kappa (with their 95% CIs).

To assess the accuracy of 4-stage classification of sleep stages (light, deep, REM and wake) we report confusion matrix, overall Cohen's kappa and accuracy, with associated 95% CIs.

Additionally for each sleep stage we report stage-specific Cohen's Kappa, stage-specific accuracy, PPV and sensitivity. To obtain accuracy measures on each sleep stage, the outcomes were dichotomized to the sleep stage of interest against all others. The average method calculates the measure for each individual participant and then averages out the measure across all participants [19]. 95% percentile CIs were obtained using bootstrap method.

All analyses were performed with R version 4.3.1 (2023-06-16).

As part of exploratory analysis, we further evaluated all core analytics endpoints across relevant participant subgroups (i.e., age, sex, BMI, skin tone, arm hair density). Analyses were performed among subgroups with a sample size ≥ 10. In some cases, groups with fewer than 10 participants were combined with other groups that shared similar characteristics, when possible, to obtain a minimum sample size of at least 10 participants.

medRxiv preprint doi: https://doi.org/10.1101/2024.09.10.24313425; this version posted September 11, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

Results

Participants included 41 adults (18 male) with a mean age of 40.5 years (SD, 16.5) and ranging from 18-78 years. The majority of participants were White (22 [53.7%]) and not hispanic or latino (37 [90.2%]), but there was diversity in subgroups, including different skin tones (light, n=21; medium, n=15; dark, n=5), body mass index (BMI; range: 17.8 - 36.0), and arm hair density (little to no visible hair, n=17; visible fine hair, n=16; coarse and very coarse hair, n=8) (Table 1, Supplemental Table 1).

Core analytics and main outcome variables

We collected data for a total of 38,796 epochs and all epochs were used for analyses.

The sensitivity of the VNW classifying sleep vs wake compared to the PSG was 0.96 (95% CI: 0.95, 0.98), specificity was 0.65 (95% CI: 0.60, 0.70), PPV was 0.92 (95% CI: 0.90, 0.94) and NPV was 0.79 (95% CI: 0.72, 0.88) (Table 2).

The mean bias for TST was 14.55 minutes (95% CI: 1.61, 27.16), and for WASO was -11.77 minutes (-23.89, 1.09). For SE, the mean bias was 3.15% (95% CI: 0.68, 5.67) and for SOL, the mean bias was -3.24 minutes (95% CI: -9.38, 3.57). The mean bias for the duration of different sleep stages was 3.78 minutes (95% CI: -7.04, 15.06) for light sleep, 3.91 minutes (95% CI: -4.59, 12.60) for deep sleep, and 6.94 minutes (0.57, 13.04) for REM sleep. The median NAWK counts for PSG is 1 and VNW is 1 with the difference in median counts of NAWKs was 0.00 (95% CI: 0.00, 1.00), the difference in the mean counts of NAWKs was -0.02: (95% CI: -0.56, -0.54), and the linear weighted Cohen's kappa coefficient was 0.41 (95% CI: 0.26, 0.57). Bland-Altman analyses showed that all measures had significant proportional bias (Table 3), with the VNW slightly overestimating values at the low end of the distribution, and underestimating them at the high end (Figure 1). For all measures, proportional bias was true; the assumption of

medRxiv preprint doi: https://doi.org/10.1101/2024.09.10.24313425; this version posted September 11, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

normality was false for all measures, except for deep and REM sleep duration; and heteroscedasticity was false for all measures, except for SOL (Table 3).

The overall accuracy of the VNW algorithm in classifying sleep stages was 0.78 (95% CI: 0.51, 0.88), and the overall kappa was 0.64 (95% CI: 0.08, 0.82) (Table 4). There was variability in the performance across different sleep stages, with light sleep stage prediction having the lowest accuracy (Table 4), as there were instances of confusion between light sleep stage and all other stages (Table 5).

Additional analytics and exploratory analyses

Subgroup analyses of the performance for the sleep vs wake classification, as well as the derived sleep measures, revealed results largely consistent with the overall group (Figure 2; Supplement Tables 2-12).

medRxiv preprint doi: https://doi.org/10.1101/2024.09.10.24313425; this version posted September 11, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

Discussion

This evaluation of the performance of the VNW's algorithm-derived sleep measures compared to PSG epoch-by-epoch in sleepers without elevated insomnia symptoms or OSA showed that sleep versus wake classification performance estimates were largely comparable or numerically higher than previously published results for other commercial wearable devices.[34,35]

Similar to other novel wearable devices, the ability to provide a 4-stage sleep classification was another strength of the VNW as some actigraphy-based devices only allow for 2-stage sleep classification, which precludes the detection of specific sleep stages that are indicated in various cognitive functioning and disease processes, such as depression and Parkinson's Disease.[36–40]

Overall accuracy of the VNW for sleep stage classification was similar to other studies for the 4-stage model on a sample without OSA and heightened insomnia symptoms (see Schyvens et al.[22] and Chinoy et al.[23] for example). There were a few participants that had a low number of epochs in a particular sleep stage, including one participant with no deep sleep epochs. The sparsity of epochs in a particular sleep stage caused highly variable Kappa estimates, which likely led to a wide 95% CI. We found VNW to show significant heteroskedasticity for SOL and significant proportional bias for TST, WASO, SE, SOL, and light, deep, and REM sleep duration, overestimating and underestimating shorter and longer values, respectively, both of which are common in other wearable devices.[41] Based on a mean of the bias estimates (Table 3), the 95% CI for TST, WASO, and SOL biases fell within the range of allowable differences for actigraphy in clinical populations recommended by the AASM clinical practice guideline. However, when using the proportional mean bias estimate, which accounts for variation in bias over the range of measurement, the 95% CI exceeds the allowable difference at lower and higher ends of the distributions (Figure 1, Table 3). These allowable differences are also based on adults with specific sleep disorders and the current sample did not include patients with these characteristics.

medRxiv preprint doi: https://doi.org/10.1101/2024.09.10.24313425; this version posted September 11, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

Investigating the generalizability of the performance of the VNW among different demographic subgroups was an aspect of special interest in our study. VNW performed consistently across subgroups for the classification of sleep vs wake and sleep stages, albeit with larger variability in estimates for those with darker skin tone. In addition, for the calculated sleep measures, all mean bias 95% CIs overlapped across demographic subgroups (i.e., sex, age, BMI, skin tone, arm hair density), indicating that there were likely no statistically significant differences between groups; but as stated previously, limited sample sizes may preclude the ability to detect differences between groups. Lastly, a slight difference in performance was observed between age groups (younger vs older) and between sexes.

Limitations and future perspectives

Strengths of this study include thoroughness of statistical methods,[19] the compliance with state-of-the art recommendation for performance evaluation studies,[7] and the diverse sample of participants, with a range of ages, BMI, skin tones, and arm hair density, which are known factors to influence PPG signal quality. Nonetheless, there are important limitations to consider in our study.

First, we focused this investigation on data collected from a sample of adult sleepers without elevated insomnia symptoms or OSA. In order to fully characterize performance in clinically relevant scenarios, future studies should evaluate the performance of this device in populations with disturbances in sleep, such as people with sleep disorders. Similarly, although our study group had about 30% of participants with overweight/obese BMI, it is unknown whether performance would generalize to morbidly obese populations, whose watch fit and tissue characteristics may vary and who often have sleep apnea as a comorbidity.

Second, as a typical procedure for the performance evaluation of sleep measures with in-lab PSG as the reference standard, epoch by epoch accuracy evaluation was conducted between lights-off and lights-on. This may limit the interpretation or generalizability of SOL results, as

medRxiv preprint doi: https://doi.org/10.1101/2024.09.10.24313425; this version posted September 11, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

there is no way to easily passively identify the moment a participant begins to attempt to initiate sleep in real-world settings. In addition, as is common practice, participants started wearing the VNW before PSG was on, the start of the PSG recording was labeled as “Time 0”, and time was rounded (e.g., 22:32:00).[42] Yet, the mean overall difference between the onset times of the devices was below the range shown to introduce significant bias in study outcomes (see Supplementary Materials). While there is always some degree of error introduced by alignment methods, we cannot rule out that this may have introduced some error that would have resulted in a small degree of underestimating performance of staging.

Third, our subgroup analyses were underpowered, limiting our ability to extract robust conclusions from them. Future studies should aim to recruit larger diverse cohorts, particularly sampling for specific key features that may affect device performance, including but not limited to skin tone, arm hair density, and clinical status.

Conclusion

Results demonstrate the potential of the VNW to effectively measure 12 standard sleep metrics, as compared to gold-standard PSG-based labels, in a demographically diverse sample of adults. Results from the epoch by epoch sleep versus wake classification, sleep stage classification, and from the derived overnight sleep measures showed comparable performance across demographic subgroups.

The results support the application of this device to monitor and understand sleep behaviors in sleepers without elevated insomnia symptoms or OSA in free-living settings for long durations, when PSG collection is not an optimal method.

medRxiv preprint doi: https://doi.org/10.1101/2024.09.10.24313425; this version posted September 11, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

Tables

Table 1. Summary of key participant characteristics

		N=41
Age (years)	Median (range)	34.0 (18.0 - 78.0)
	Mean (SD)	40.5 (16.5)
Age categories, n (%)	18-40	25 (61.0)
	41-80	16 (39.02)
Sex, n (%)	Female	23 (56.1)
	Male	18 (43.9)
BMI categories, n (%)	< 25	30 (73.17)
	≥ 25	11 (26.82)
Skin tone, n (%)	Light Skin Tone	21 (51.21)
	Medium Skin Tone	15 (36.59)
	Dark Skin Tone	5 (12.20)
Arm hair index, n (%)	1: Little to no visible arm hair, light in color	17 (41.5)
	2: Visible, fine, arm hair, light to medium color	16 (39.0)
	3 and 4: Coarse and very coarse arm hair, medium to dark color	8 (19.51)
Race, n (%)	American Indian or Alaska Native	1 (2.4)
	Asian	8 (19.5)
	Black of African American	4 (9.8)

medRxiv preprint doi: https://doi.org/10.1101/2024.09.10.24313425; this version posted September 11, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

	Mixed race	4 (9.8)
	Native Hawaiian or Other Pacific Islander	1 (2.4)
	Other	1 (2.4)
	White	22 (53.7)
Ethnicity, n(%)	Hispanic or Latino	4 (9.76)
	Not Hispanic or Latino	37 (90.20)
Dominant hand, n (%)	Ambidextrous	1 (2.4)
	Left	5 (12.2)
	Right	35 (85.4)
OSA score	Median (range)	0.0 (0.0 - 7.0)
	Mean (SD)	1.3 (1.6)
ISI score	Median (range)	3.0 (0.0 - 7.0)
	Mean (SD)	3.0 (1.9)
ESS score	Median (range)	5.0 (0.0 - 9.0)
	Mean (SD)	5.0 (2.7)
AHI Index	Median (range)	1.3 (0.1 - 4.7)
	Mean (SD)	1.7 (1.2)

Note: AHI=Apnea hypopnea index ; BMI=body mass index; ESS=Epworth sleepiness scale ; ISI=insomnia severity index ; OSA=Obstructive Sleep Apnea; SD=standard deviation

medRxiv preprint doi: https://doi.org/10.1101/2024.09.10.24313425; this version posted September 11, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

Table 2. Summary of performance metrics for 'sleep vs wake' classification of 30-second epochs

	Sensitivity [95% CI]	**Specificity [95% CI]**	**NPV [95% CI]**	**PPV [95% CI]**
Sleep vs Wake	0.96 (0.95, 0.98)	0.65 (0.60, 0.70)	0.79 (0.72, 0.88)	0.92 (0.90, 0.94)

CI = confidence interval; NPV = negative predictive value (prediction of wake); PPV = positive predictive value (prediction of sleep)

Table 3. Summary of performance metrics for derived sleep measures

Measure	PSG (SD)	Study Watch (SD)	Bias Estim. [95% CI]	Assumptions	Proportional Mean Bias Estim.	[95% CI]	LOA			
							Lower LOA	[95% CI]	Upper LOA	[95% CI]
TST (min)	384.98 (60.85)	399 (46.33)	14.55 [1.61, 27.16]	Prop Bias = T Normality = F Heteroscedasticity = F	184 + (-0.44) * Ref	intercept = [76.3, 275.11], slope = [-0.69,-0.16]	-68.0	[-136.02, 18.23]	94.0	[25.09, 179.96]
WASO (min)	62.72 (49.97)	50.95 (42.99)	-11.77 [-23.89, 1.09]	Prop Bias = T Normality = F Heteroscedasticity = F	17.55 - 0.47 x Ref	intercept = [1.97, 35.03], slope = [-0.68, -0.19]	-86.4	[-163.51, -20.24]	74.48	[-3.34, 141.15]
SE (%)	81.69 (11.71)	84.84 (8.99)	3.15 [0.68, 5.67]	Prop Bias = T Normality = F Heteroscedasticity = F	39.55 - (0.45 * Ref)	Intercept = [19.51, 54.31]; slope = [-0.63, -0.21]	-12.99	[-25.85, 4.47]	18.80	[5.79, 36.21]
SOL (min)	25.43 (20.37)	22.18 (22.79)	-3.24 [-9.38, 3.57]	Prop Bias = T Normality = F Heteroscedasticity = T	7.55 - (0.42 * Ref)	intercept = [-1.1, 19.72], slope = [-0.83, -0.09]	bias - 2.46 (5.2 + 0.23 x Ref)	intercept = [-1.26, 15.02], slope = [0.01, 0.41]	bias + 2.46(5.2 + 0.23 x Ref)	intercept = [-1.26, 15.02], slope = [0.01, 0.41]
Duration of sleep stages										
Light (min)	240.65 (49.27)	244.43 (44.76)	3.78 [-7.04,15.06]	Prop Bias = T Normality = F Heteroscedasticity = F	91.75 - (0.37 * PSG)	intercept = [28.55, 162.77] slope = [-0.67, -0.12]	-65.05	[-130.33, -1.63]	79.04	[13.73,142.15]
Deep (min)	63.39 (27.19)	67.30 (20.75)	3.91 [-4.59, 12.60]	Prop Bias = T Normality = T Heteroscedasticity = F	51.27 - (0.75 x PSG)	intercept = [35.28, 69.02] slope = [-1, -0.52]	-45.30	[-101.36, 2.24]	65.30	[9.30,112.81]

REM (min)	82.49 (25.46)	89.43 (22.26)	6.94 [0.57, 13.04]	Prop Bias = T Normality = T Heteroscedasticity = F	44.08 PSG	-0.45 x	intercept = [26.57, 64.85] slope = [-0.71, -0.24]	-44.69	[-74.91, 2.55]	36.69	[6.57,84.11]

CI= confidence interval; LOA = limits of agreement; REM = rapid eye movement; SD = standard deviation; SE = sleep efficiency; SOL = sleep onset latency; TST = total sleep time; WASO = wake after sleep onset

medRxiv preprint doi: https://doi.org/10.1101/2024.09.10.24313425; this version posted September 11, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

Table 4. Summary of performance metrics for the classification of sleep stages

Sleep Stage	Kappa [95% CI]	Accuracy [95% CI]	PPV [95% CI]	Sensitivity [95% CI]
Overall	0.64 [0.08, 0.82]	0.78 [0.51, 0.88]	NA	NA
Wake	0.67 [0.26, 0.89]	0.91 [0.69, 0.98]	0.82 [0.39, 0.98]	0.69 [0.38, 0.91]
Light	0.60 [0.21, 0.78]	0.80 [0.60, 0.89]	0.80 [0.51, 0.92]	0.81 [0.58, 0.93]
Deep	0.66 [0.08, 0.88]	0.93 [0.82, 0.98]	0.68 [0.05, 0.97]	0.77 [0.21, 0.98]
REM	0.73 [0.36, 0.91]	0.92 [0.83, 0.98]	0.75 [0.41, 0.93]	0.83 [0.43, 0.98]

CI= confidence interval; REM = rapid eye movement

medRxiv preprint doi: https://doi.org/10.1101/2024.09.10.24313425; this version posted September 11, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

Table 5. Confusion matrix for the classification of sleep stages.

		Device (Verily Numetric Watch)				
		Wake	**Light**	**Deep**	**REM**	**Total Reference**
Reference	**Wake**	4,773	1,903	44	508	7,228
	Light	886	16,011	1,516	1,320	19,733
	Deep	190	1,704	3,799	8	5,071
	REM	185	1,056	26	5,497	6,764
Total Device		6,034	2,0044	5,385	7,333	38,796

REM = rapid eye movement

medRxiv preprint doi: https://doi.org/10.1101/2024.09.10.24313425; this version posted September 11, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

Figure Legends

Figure 1. Bland-Altman plots of derived sleep measures.

Note: Solid red lines indicate mean bias. Dotted red lines indicate 95% CI of mean bias. Solid gray lines indicate the 95% LOAs. Dotted gray lines indicate 95% CI of LOAs. Black dots are observations.

Abbreviations: CI= confidence interval; LOA = limits of agreement; REM = rapid eye movement; SD = standard deviation; SE = sleep efficiency; SOL = sleep onset latency; TST = total sleep time; WASO = wake after sleep onset

Figure 2. Performance of sleep vs wake classification across various subgroups based on sex, age, BMI, skin tone, and arm hair index.

Notes: Bars show 95% CI ranges; dots, point estimates. For reference: red dotted line, overall point estimate; blue shade, overall 95% CI. For the skin tone classification, larger values indicate darker skin tone; for Arm Hair Index, larger values indicate coarser arm hair. Gray bars indicate subgroup with sample size < 10 participants.

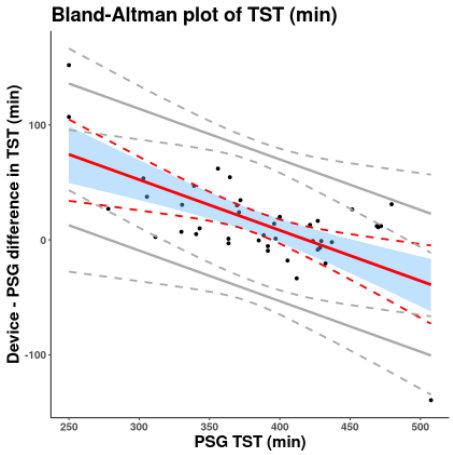
Abbreviations: BMI = body mass index; PPV = positive predictive value; NPV=negative predictive value

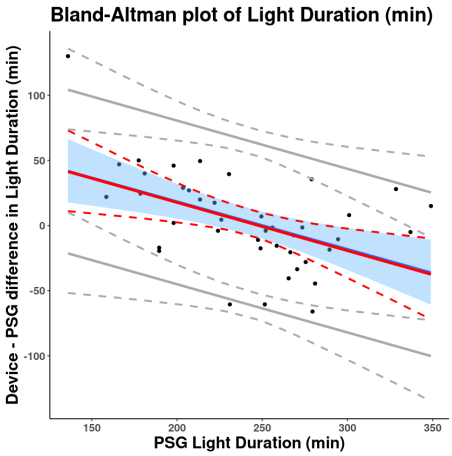
medRxiv preprint doi: https://doi.org/10.1101/2024.09.10.24313425; this version posted September 11, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

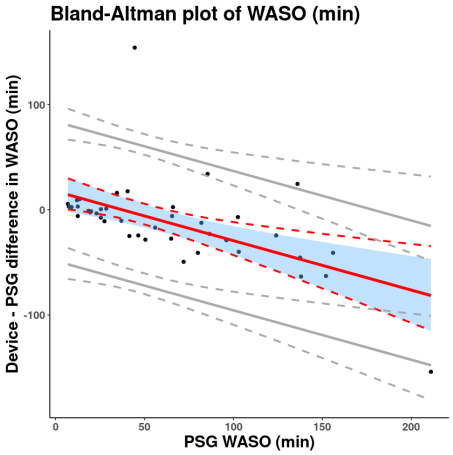
It is made available under a CC-BY-NC-ND 4.0 International license .

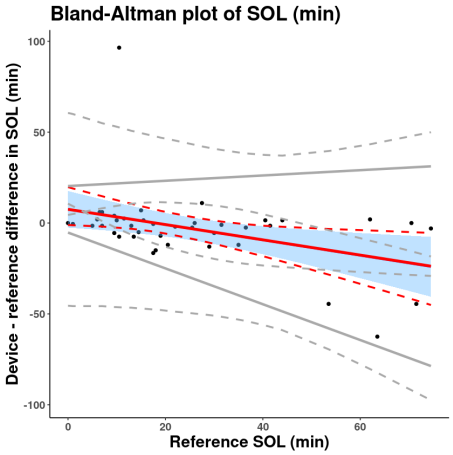
Figures

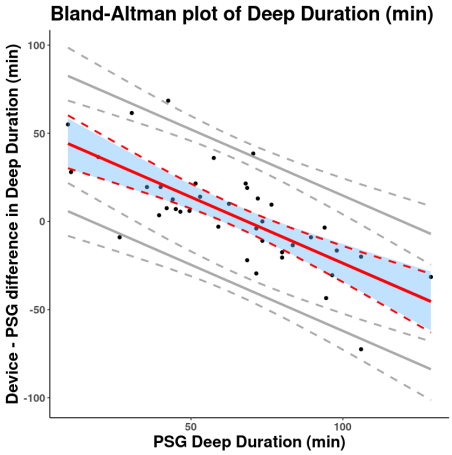
Figure 1

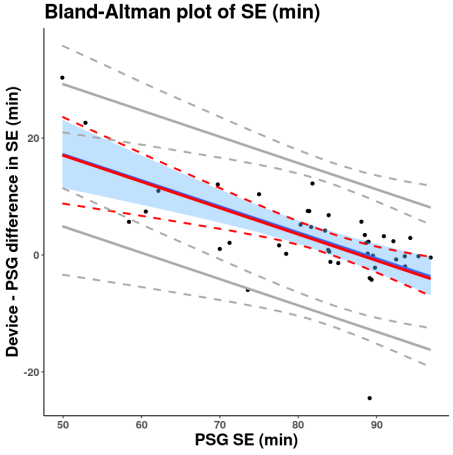


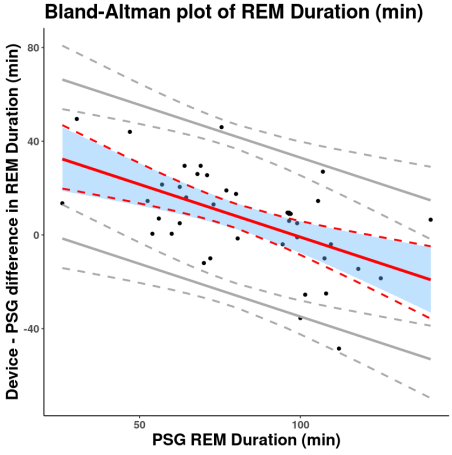








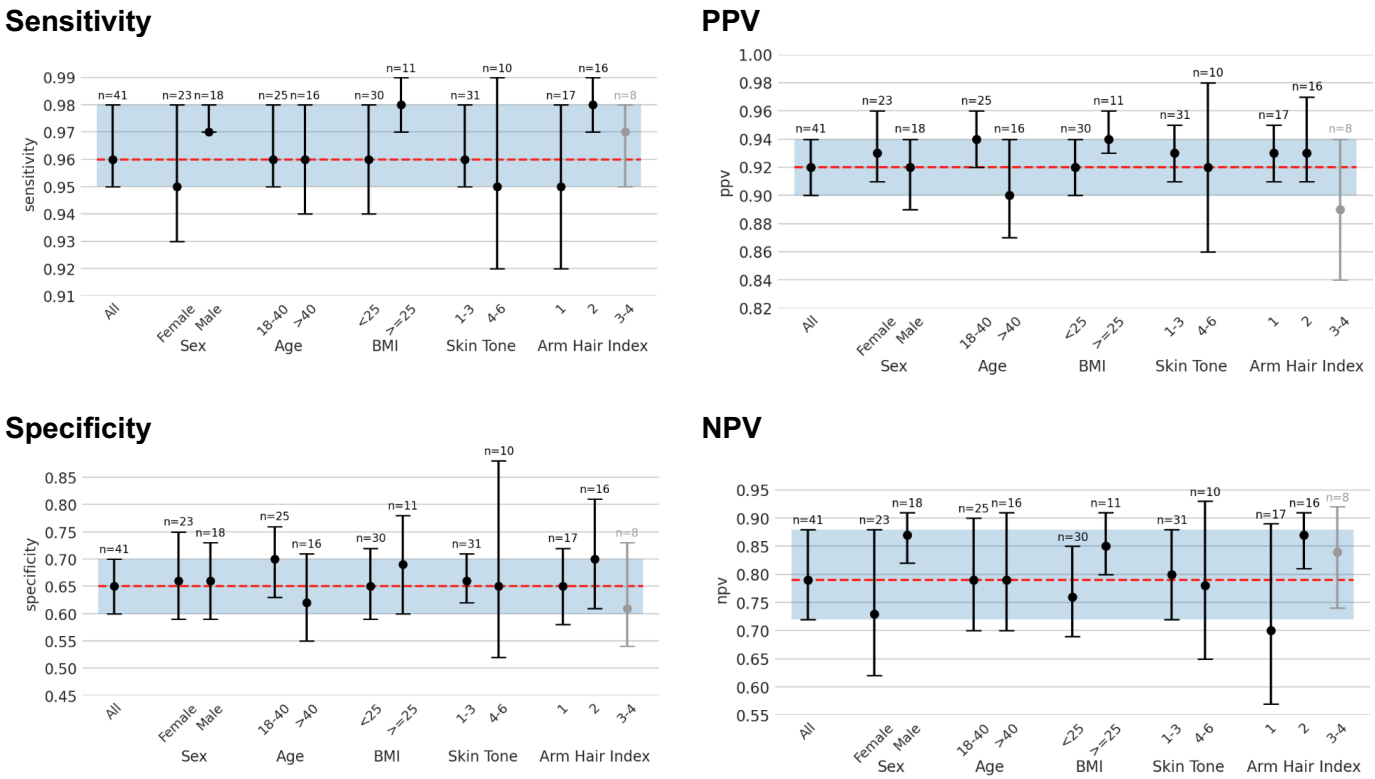




medRxiv preprint doi: https://doi.org/10.1101/2024.09.10.24313425; this version posted September 11, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 2.



medRxiv preprint doi: https://doi.org/10.1101/2024.09.10.24313425; this version posted September 11, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

References

1. Freeman D, Sheaves B, Waite F, Harvey AG, Harrison PJ. Sleep disturbance and psychiatric disorders. *Lancet Psychiatry*. 2020;7(7):628-637.

2. Tobaldini E, Fiorelli EM, Solbiati M, Costantino G, Nobili L, Montano N. Short sleep duration and cardiometabolic risk: from pathophysiology to clinical evidence. *Nat Rev Cardiol*. 2019;16(4):213-224.

3. Toussaint M, Luthringer R, Schaltenbrand N, et al. First-night effect in normal subjects and psychiatric inpatients. *Sleep*. 1995;18(6):463-469.

4. de Zambotti M, Goldstein C, Cook J, et al. State of the science and recommendations for using wearable technology in sleep and circadian research. *Sleep*. 2024;47(4). doi:10.1093/sleep/zsad325

5. Goldsack JC, Coravos A, Bakker JP, et al. Verification, analytical validation, and clinical validation (V3): the foundation of determining fit-for-purpose for Biometric Monitoring Technologies (BioMeTs). *NPJ Digit Med*. 2020;3:55.

6. Benedetti D, Menghini L, Vallat R, et al. Call to action: an open-source pipeline for standardized performance evaluation of sleep-tracking technology. *Sleep*. 2023;46(2). doi:10.1093/sleep/zsac304

7. 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(3):263-269.

8. Zinzuwadia A, Singh JP. Wearable devices-addressing bias and inequity. *Lancet Digit Health*. 2022;4(12):e856-e857.

9. Shachar C, Gerke S. Prevention of Bias and Discrimination in Clinical Practice Algorithms. *JAMA*. 2023;329(4):283-284.

10. Goodman KE, Morgan DJ, Hoffmann DE. Clinical Algorithms, Antidiscrimination Laws, and Medical Device Regulation. *JAMA*. 2023;329(4):285-286.

11. Colvonen PJ, DeYoung PN, Bosompra NOA, Owens RL. Limiting racial disparities and bias for wearable devices in health science research. *Sleep*. 2020;43(10). doi:10.1093/sleep/zsaa159

12. Valbuena VSM, Seelye S, Sjoding MW, et al. Racial bias and reproducibility in pulse oximetry among medical and surgical inpatients in general care in the Veterans Health Administration 2013-19: multicenter, retrospective cohort study. *BMJ*. 2022;378:e069775.

13. Fawzy A, Wu TD, Wang K, et al. Racial and Ethnic Discrepancy in Pulse Oximetry and Delayed Identification of Treatment Eligibility Among Patients With COVID-19. *JAMA Intern Med*. 2022;182(7):730-738.

medRxiv preprint doi: https://doi.org/10.1101/2024.09.10.24313425; this version posted September 11, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

14. Sjoding MW, Dickson RP, Iwashyna TJ, Gay SE, Valley TS. Racial Bias in Pulse Oximetry Measurement. *N Engl J Med*. 2020;383(25):2477-2478.

15. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447-453.

16. Weatherall J, Paprocki Y, Meyer TM, Kudel I, Witt EA. Sleep Tracking and Exercise in Patients With Type 2 Diabetes Mellitus (Step-D): Pilot Study to Determine Correlations Between Fitbit Data and Patient-Reported Outcomes. *JMIR Mhealth Uhealth*. 2018;6(6):e131.

17. Imhoff-Smith TP, Grupe DW. The impact of mindfulness training on posttraumatic stress disorder symptoms, subjective sleep quality, and objective sleep outcomes in police officers. *Psychol Trauma*. Published online August 31, 2023. doi:10.1037/tra0001566

18. Kasparian AM, Badawy SM. Utility of Fitbit devices among children and adolescents with chronic health conditions: a scoping review. *Mhealth*. 2022;8:26.

19. Menghini L, Cellini N, Goldstone A, Baker FC, de Zambotti M. A standardized framework for testing the performance of sleep-tracking technology: step-by-step guidelines and open-source code. *Sleep*. 2021;44(2). doi:10.1093/sleep/zsaa170

20. Khosla S, Wickwire EM. Consumer sleep technology: accuracy and impact on behavior among healthy individuals. *J Clin Sleep Med*. 2020;16(5):665-666.

21. Khosla S, Deak MC, Gault D, et al. Consumer Sleep Technology: An American Academy of Sleep Medicine Position Statement. *J Clin Sleep Med*. 2018;14(5):877-880.

22. Schyvens AM, Van Oost NC, Aerts JM, et al. Accuracy of Fitbit Charge 4, Garmin Vivosmart 4, and WHOOP Versus Polysomnography: Systematic Review. *JMIR Mhealth Uhealth*. 2024;12:e52192.

23. Chinoy ED, Cuellar JA, Huwa KE, et al. Performance of seven consumer sleep-tracking devices compared with polysomnography. *Sleep*. 2021;44(5). doi:10.1093/sleep/zsaa291

24. Chai-Coetzer CL, Antic NA, Rowland LS, et al. A simplified model of screening questionnaire and home monitoring for obstructive sleep apnoea in primary care. *Thorax*. 2011;66(3):213-219.

25. Bastien CH, Vallières A, Morin CM. Validation of the Insomnia Severity Index as an outcome measure for insomnia research. *Sleep Med*. 2001;2(4):297-307.

26. AASM. American Academy of Sleep Medicine. Obstructive sleep Apnea. Published online 2008.

27. Berry RB, Abreu AR, Krishnan V, Quan SF, Strollo PJ, Malhotra RK. A transition to the American Academy of Sleep Medicine-recommended hypopnea definition in adults: initiatives of the Hypopnea Scoring Rule Task Force. *J Clin Sleep Med*. 2022;18(5):1419-1425.

28. Sridhar N, Shoeb A, Stephens P, et al. Deep learning for automated sleep staging using

medRxiv preprint doi: https://doi.org/10.1101/2024.09.10.24313425; this version posted September 11, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

instantaneous heart rate. *NPJ Digit Med*. 2020;3:106.

29. American Academy of Sleep Medicine. *The AASM Manual for the Scoring of Sleep and Associated Events, Version 3*.; 2023.

30. Johns MW. A new method for measuring daytime sleepiness: the Epworth sleepiness scale. *Sleep*. 1991;14(6):540-545.

31. Fitzpatrick TB. The Validity and Practicality of Sun-Reactive Skin Types I Through VI. *Arch Dermatol*. 1988;124(6):869-871.

32. Cameron AC, Gelbach JB, Miller DL. Bootstrap-based improvements for inference with clustered errors. *Rev Econ Stat*. 2008;90(3):414-427.

33. Ohayon M, Wickwire EM, Hirshkowitz M, et al. National Sleep Foundation's sleep quality recommendations: first report. *Sleep Health*. 2017;3(1):6-19.

34. de Zambotti M, Goldstone A, Claudatos S, Colrain IM, Baker FC. A validation study of Fitbit Charge 2™ compared with polysomnography in adults. *Chronobiol Int*. 2018;35(4):465-476.

35. Miller DJ, Sargent C, Roach GD. A Validation of Six Wearable Devices for Estimating Sleep, Heart Rate and Heart Rate Variability in Healthy Adults. *Sensors* . 2022;22(16). doi:10.3390/s22166317

36. Postuma RB, Gagnon JF, Vendette M, Fantini ML, Massicotte-Marquez J, Montplaisir J. Quantifying the risk of neurodegenerative disease in idiopathic REM sleep behavior disorder. *Neurology*. 2009;72(15):1296-1300.

37. MacDonald KJ, Cote KA. Contributions of post-learning REM and NREM sleep to memory retrieval. *Sleep Med Rev*. 2021;59:101453.

38. Postuma RB, Lang AE, Massicotte-Marquez J, Montplaisir J. Potential early markers of Parkinson disease in idiopathic REM sleep behavior disorder. *Neurology*. 2006;66(6):845-851.

39. Sixel-Döring F, Trautmann E, Mollenhauer B, Trenkwalder C. Associated factors for REM sleep behavior disorder in Parkinson disease. *Neurology*. 2011;77(11):1048-1054.

40. Palagini L, Baglioni C, Ciapparelli A, Gemignani A, Riemann D. REM sleep dysregulation in depression: state of the art. *Sleep Med Rev*. 2013;17(5):377-390.

41. Menghini L, Yuksel D, Goldstone A, Baker FC, de Zambotti M. Performance of Fitbit Charge 3 against polysomnography in measuring sleep in adolescent boys and girls. *Chronobiol Int*. 2021;38(7):1010-1022.

42. de Zambotti M, Cellini N, Goldstone A, Colrain IM, Baker FC. Wearable Sleep Technology in Clinical and Research Settings. *Med Sci Sports Exerc*. 2019;51(7):1538-1557.

43. Johns MW. Reliability and factor analysis of the Epworth Sleepiness Scale. *Sleep*. 1992;15(4):376-381.

medRxiv preprint doi: https://doi.org/10.1101/2024.09.10.24313425; this version posted September 11, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

44. Shi C, Goodall M, Dumville J, et al. The accuracy of pulse oximetry in measuring oxygen saturation by levels of skin pigmentation: a systematic review and meta-analysis. *BMC Med*. 2022;20(1):267.

medRxiv preprint doi: https://doi.org/10.1101/2024.09.10.24313425; this version posted September 11, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

[Supplement]

Title: Performance Evaluation of the Verily Numetric Watch sleep suite for digital sleep assessment against in-lab polysomnography

Authors: Benjamin W. Nelson, PhD; Sohrab Saeb, PhD; Poulami Barman, MS; Nishant Verma, PhD; Hannah Allen, BS; Massimiliano de Zambotti, PhD; Fiona C. Baker, PhD; Nicole Arra, BA; Niranjan Sridhar, PhD; Shannon S. Sullivan, MD, MSc; Scooter Plowman, MD, MBA, MHSA, MS[1]; Erin Rainaldi, MS; Ritu Kapur, PhD; Sooyoon Shin, PhD

medRxiv preprint doi: https://doi.org/10.1101/2024.09.10.24313425; this version posted September 11, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

Supplement Table 1.

Full summary of participant characteristics

		N=41
Age (yrs)	Median (range)	34.0 (18.0 - 78.0)
	Mean (SD)	40.5 (16.5)
Age categories, n (%)	18-40	25 (61.0)
	40-60	8 (19.5)
	60-80	8 (19.5)
Sex, n (%)	Female	23 (56.1)
	Male	18 (43.9)
BMI	Median (range)	23.3 (17.8 - 36.0)
	Mean (SD)	24.2 (4.2)
BMI categories, n (%)	<18.5	1 (2.4)
	18.5-25	29 (70.7)
	25-30	6 (14.6)
	≥30	5 (12.2)
Ethnicity, n (%)	Hispanic or Latino	4 (9.8)
	Not Hispanic or Latino	37 (90.2)
Skin tone, n (%)	Type I - Always burns, never tans	1 (2.44)
	Type II - Usually burns, then tans	20 (48.8)
	Type III - May burn, tans well	10 (24.4)

medRxiv preprint doi: https://doi.org/10.1101/2024.09.10.24313425; this version posted September 11, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

	Type IV - Rarely burns, tans well	5 (12.2)
	Type V - Very rarely burns, tans well, brown skin	3 (7.3)
	Type VI - Very rarely burns, tans well, very dark skin	2 (4.9)
Arm hair index, n (%)	1: Little to no visible arm hair, light in color	17 (41.5)
	2: Visible, fine, arm hair, light to medium color	16 (39.0)
	3: Coarse arm hair, medium to dark color	6 (14.6)
	4: Very coarse arm hair, dark in color	2 (4.9)
Right wrist circumference, cm	Median (range)	6.2 (5.5 - 7.4)
	Mean (SD)	6.29 (0.49)
Left wrist circumference, cm	Median (range)	6.2 (5.3 - 7.5)
	Mean (SD)	6.28 (0.48)
Right wrist circumference categories, n (%)	Large	13 (31.7)
	Medium	14 (34.1)
	Small	14 (34.1)
Left wrist circumference categories, n (%)	Large	13 (31.7)
	Medium	14 (34.1)
	Small	14 (34.1)
Race, n (%)	American Indian or Alaska Native	1 (2.4)
	Asian	8 (19.5)
	Black of African American	4 (9.8)
	Mixed race	4 (9.8)

medRxiv preprint doi: https://doi.org/10.1101/2024.09.10.24313425; this version posted September 11, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

	Native Hawaiian or Other Pacific Islander	1 (2.4)
	Other	1 (2.4)
	White	22 (53.7)
Dominant hand, n (%)	Ambidextrous	1 (2.4)
	Left	5 (12.2)
	Right	35 (85.4)
OSA score	Median (range)	0.0 (0.0 - 7.0)
	Mean (SD)	1.34 (1.64)
ISI score	Median (range)	3.0 (0.0 - 7.0)
	Mean (SD)	2.95 (1.92)
ESS score	Median (range)	5.0 (0.0 - 9.0)
	Mean (SD)	5.0 (2.65)
AHI Index	Median (range)	1.3 (0.1 - 4.7)
	Mean (SD)	1.69 (1.2)

AHI=Apnea hypopnea index ; BMI=body mass index; ESS=Epworth sleepiness scale ; ISI=insomnia severity index ; OSA=Obstructive Sleep Apnea; SD=standard deviation; Skin Type I and II were categorized as light skin tone; Skin Type III and IV were categorized as medium skin tone; Skin Type V and VI were categorized as dark skin tone

medRxiv preprint doi: https://doi.org/10.1101/2024.09.10.24313425; this version posted September 11, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

Detailed description of enrollment screening

Questionnaires completed by participants or study personnel and used for screening were:

- Obstructive Sleep Apnea 50 (OSA50): a brief 4-item (obesity, snoring, apneas, and age) diagnostic screening questionnaire for OSA in primary care that has been shown to identify patients with moderate to severe OSA with 84% accuracy [24]. This questionnaire has a cut-off score of ≥ 5, with typical sleepers having an OSA50 score of < 5 (sensitivity 100%, specificity 29%).
- AHI: calculated from PSG signals, it is the total number of apnea or hypopnea events in a night divided by the hours of sleep. The AHI assists in the diagnosis of OSA, classifying participants as having Normal sleep (< 5 events per hour), Mild (5-14 events per hour), Moderate (15-29 events per hour), and Severe (30 or more events per hour).
- Insomnia Severity Index (ISI): a 7-item questionnaire designed to screen for insomnia [25]. A 5-point Likert scale ranging from 0 = no problem to 4 = very severe problem is used for each question to calculate a total score ranging from 0 to 28. Total scores are categorized as absence of insomnia (0–7); sub-threshold insomnia (8–14); moderate insomnia (15–21); and severe insomnia (22–28). An ISI < 8 is used to identify no clinically significant insomnia symptoms.
- ESS: an 8-item self-reported questionnaire used to assess daytime sleepiness [30]. A 4-point Likert scale ranging from 0 = would never doze to 3 = high chance of dozing is used for each question to calculate a total score ranging from 0 to 24. Total scores are categorized as normal (0-10) and then increasing degree of daytime sleepiness (11-24). This measure has good internal consistency[43].
- Fitzpatrick Skin Scale (completed by study personnel) [31]: a 10-item questionnaire that assesses Genetic (physical traits), Sensitivity (reaction to sun exposure), and Intentional

medRxiv preprint doi: https://doi.org/10.1101/2024.09.10.24313425; this version posted September 11, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

Exposure (tanning habits) to categorize participants on a scale from 1-6 as research has shown that skin tone can influence the accuracy of PPG signals. Categories include, Type I (scores 0–6) always burns, never tans (palest; freckles), Type II (scores 7–13) usually burns, tans minimally (light colored but darker than fair), Type III (scores 14–20) sometimes mild burn, tans uniformly (golden honey or olive), Type IV (scores 21–27) burns minimally, always tans well (moderate brown), Type V (scores 28–34) very rarely burns, tans very easily (dark brown), Type VI (scores 35–36) never burns (deeply pigmented dark brown to darkest brown). These can be categorized into light (Type I and II), medium (Type III and IV), and dark (Type V and VI)[44].

- Arm Hair Index (scored by study personnel): assesses the density of participants' arm hair on a scale of one to four, including 1 (Little to no visible hair), 2 (Visible fine arm hair), 3 (Coarse arm hair), and 4 (Very coarse arm hair).

Device synchronization during overnight study visit

Time synchronization between the VNW and PSG devices was performed using the following steps. [42]

First, the VNW was placed on the non-dominant wrist and the time (HH:MM) was recorded in the Device Accountability Log and electronic data capture (EDC) system.

Next, PSG hook up and calibration were performed; PSG was then turned on at the top of the minute on the VNW device (i.e., when the minute on the watch face changed) and the time (HH:MM) was logged in the EDC system.

Participants started wearing VNW before PSG was on and the overall difference between the onset times (mean = 3.51 seconds, SD = 4.06 seconds) was well below the range that has been shown to introduce significant bias in study outcomes. As is common practice, the start of the PSG recording was labeled as “Time 0” and time was rounded (e.g., 22:32:00).[42]

medRxiv preprint doi: https://doi.org/10.1101/2024.09.10.24313425; this version posted September 11, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

Success Criteria for Sleep Versus Wake Classification as Product Requirement Specification

We pre-specified thresholds as “success criteria” for sleep versus wake classification with success being defined as sensitivity ≥ 0.90 and specificity ≥ 0.50 (the lower 95% CI bound over these values) as internal criteria for a successful sleep versus wake classification to proceed with subsequent development of the algorithm.

Supplement Table 2.

Summary of performance metrics for ‘sleep vs wake’ classification, according to participant subgroups

Demographic Variable	Subgroup (n)	Sensitivity [95% CI]	Specificity [95% CI]	PPV [95% CI]	NPV [95% CI]
Sex	Female (23)	0.95 [0.93, 0.98]	0.66 [0.59, 0.75]	0.93 [0.91, 0.96]	0.73 [0.62, 0.88]
	Male (18)	0.97 [0.97, 0.98]	0.66 [0.59, 0.73]	0.92 [0.89, 0.94]	0.87 [0.82, 0.91]
Age	18-40 (25)	0.96 [0.95, 0.98]	0.70 [0.63, 0.76]	0.94 [0.92, 0.96]	0.79 [0.70, 0.90]
	> 40 (16)	0.96 [0.94, 0.98]	0.62 [0.55, 0.71]	0.90 [0.87, 0.94]	0.79 [0.70, 0.91]
BMI	< 25 (30)	0.96 [0.94, 0.98]	0.65 [0.59, 0.72]	0.92 [0.90, 0.94]	0.76 [0.69, 0.85]
	25 and greater (11)	0.98 [0.97, 0.99]	0.69 [0.60, 0.78]	0.94 [0.93, 0.96]	0.85 [0.80, 0.91]
Skin Tone	I - III (31)	0.96 [0.95, 0.98]	0.66 [0.62, 0.71]	0.93 [0.91, 0.95]	0.80 [0.72, 0.88]
	IV - VI (10)	0.95 [0.92, 0.99]	0.65 [0.52, 0.88]	0.92 [0.86, 0.98]	0.78 [0.65, 0.93]
Arm Hair Index	1 (17)	0.95 [0.92, 0.98]	0.65 [0.58, 0.72]	0.93 [0.91, 0.95]	0.70 [0.57, 0.89]
	2 (16)	0.98 [0.97, 0.99]	0.70 [0.61, 0.81]	0.93 [0.91, 0.97]	0.87 [0.81, 0.91]
	3-4 (8)*	–	–	–	–

BMI = body mass index; CI= confidence interval; NPV = negative predictive value; PPV = positive predictive value; * subgroups < 10

Supplement Table 3.

Summary of performance metrics for ‘sleep stage’’ classification, according to participant subgroups

Variable	Subgroup (n)	Overall Cohen’s Kappa [95% CI]	Light Sleep Kappa [95% CI]	Deep Sleep Kappa [95% CI]	REM Sleep Kappa [95% CI]	Wake Kappa [95% CI]
Sex	Female (23)	0.64[0.63,0.65]	0.58[0.17,0.76]	0.65[0.07,0.88]	0.71[0.32,0.87]	0.67[0.24,0.89]
	Male (18)	0.69[0.68,0.70]	0.62[0.46,0.77]	0.67[0.31,0.87]	0.75[0.56,0.91]	0.67[0.45,0.87]
Age	18-40 (25)	0.68[0.67,0.69]	0.63[0.41,0.78]	0.7[0.24,0.89]	0.72[0.48,0.87]	0.69[0.34,0.89]
	> 40 (16)	0.62[0.61,0.63]	0.55[0.16,0.74]	0.58[0.13,0.85]	0.74[0.35,0.91]	0.65[0.29,0.86]
BMI	< 25 (30)	0.65[0.64,0.65]	0.59[0.19,0.77]	0.66[0.09,0.87]	0.72[0.33,0.91]	0.66[0.25,0.89]
	25 and greater (11)	0.69[0.68,0.7]	0.62[0.47,0.76]	0.66[0.31,0.87]	0.76[0.65,0.87]	0.71[0.47,0.87]
Skin Tone	I - III (31)	0.67[0.66,0.68]	0.61[0.4,0.78]	0.63[0.12,0.87]	0.74[0.48,0.91]	0.67[0.34,0.84]
	IV - VI (10)	0.65[0.64,0.66]	0.59[0.18,0.76]	0.69[0.24,0.87]	0.72[0.37,0.89]	0.67[0.3,0.9]
Arm Hair Index	1 (17)	0.62[0.61,0.63]	0.57[0.19,0.77]	0.59[0.04,0.87]	0.69[0.3,0.85]	0.64[0.25,0.81]
	2 (16)	0.7[0.69,0.71]	0.63[0.38,0.76]	0.73[0.39,0.88]	0.76[0.6,0.91]	0.73[0.45,0.91]
	3-4 (8)*	–	–	–	–	–

AHI = arm hair index; BMI = body mass index; CI= confidence interval; REM = rapid eye movement; * subgroups < 10

Supplement Table 4.

Summary of performance metrics for TST according to subgroups.

Variable	Subgroup	n (%)	Device Mean (SD)	PSG Mean (SD)	Assumptions	Bias [95% CI]	Proportional Bias	Proportional Bias 95% CI	LOA lower [95% CI]	LOA upper [95% CI]	ICC [95% CI]
Age	**18-40 yrs**	25 (61)	397.58 (54.19)	388.3 (66.32)	Prop Bias = T Normality = F Heterosceda sticity = F	9.28 [-7.56,23.82]	146.6 - (0.35 x PSG)	Intercept = [23.02,253.59] Slope = [-0.65, -0.03]	-66.59 [-144.23,-6.18]	92.59 [15.11,153.35]	0.76 [0.49,0.94]
	>40 yrs	16 (39)	402.56(31.77)	379.78 (52.81)	Prop Bias = T Normality = F Heterosceda sticity = F	22.78 [5.062,45.16]	266.64 - (0.64 x PSG)	Intercept = [94.25, 375.24] Slope = [-0.93, -0.2]	-92.71 [-107.27,19.59]	73.71 [59.15,185.76]	0.42 [0,0.82]
Sex	**Female**	23 (56.1)	407.52(51.21)	400.11 (62.33)	Prop Bias = F Normality = F Heterosceda sticity = F	7.41 [-11.305,26.17]	7.41	[-11.31,26.17]	-78.50 [-165.53,-8.4]	104.50 [17.31,175.73]	0.66 [0.26,0.95]
	Male	18 (43.9)	389.31(38.2)	365.64 (54.59)	Prop Bias = T Normality = T Heterosceda sticity = F	23.67 [9.778,38.972]	181.11 - (0.43 x PSG)	Intercept = [93.36, 253.85] Slope = [-0.64, -0.2]	-58.37 [-75.85,18.20]	68.37 [50.87,144.60]	0.65 [0.31,0.86]
BMI	**<25**	30 (73.2)	398.68(46.77)	383.85 (66.08)	Prop Bias = T Normality = F Heterosceda sticity = F	14.83 [-1.867,31.38]	207.78 - (0.5 x PSG)	Intercept = [82.6, 303.29] Slope = [-0.76, -0.18]	-79.16 [-157.73,14.65]	105.16 [25.11,199.68]	0.63 [0.34,0.88]
	>=25	11 (26.8)	401.82(47.25)	388.05 (46.21)	Prop Bias = F Normality = T Heterosceda sticity = F	13.77 [2.409,25.46]	13.77	[2.41,25.46]	-20.44 [-51.91,2.98]	60.44 [29.07,83.79]	0.86 [0.44,0.95]

Skin Tone	**I-II**	21 (51.2)	402.6 (49.94)	390.05 (66.52)	Prop Bias = T Normality = F Heterosceda sticity = F	14.29 [-0.647,26.44]	183.25 - (0.44 x PSG)	Intercept = [52.05, 301.34] Slope = [-0.76, -0.1]	-75.39 [-156.01,-11.29]	97.39 [16.90,161.99]	0.7 [0.35,0.92]
	III-VI	20 (48.8)	396.3 (43.26)	379.65 (55.49)	Prop Bias = T Normality = F Heterosceda sticity = F	15.35 [-6.90,49.20]	186.36 - (0.45 x PSG)	Intercept = [15.78, 321.85] Slope = [-0.8, -0.02]	-64.05 [-100.39,22.02]	90.05 [53.60,175.66]	0.64 [0.2,0.92]
Arm Hair Index	**1**	17 (41.5)	416.71(49.4)	411.09 (57.12)	Prop Bias = T Normality = F Heterosceda sticity = F	5.62 [-17.176,22.12]	171.52 - (0.4 x PSG)	Intercept = [32.85, 385.14] Slope = [-0.93, -0.05]	-70.42 [-162.597,-36.82]	96.42 [3.75,130.13]	0.68 [0.07,0.94]
	2	16 (39.0)	388.34(45.21)	371.75 (55.94)	Prop Bias = F Normality = F Heterosceda sticity = F	16.59 [1.69,38.10]	16.59	[1.69,38.10]	75.00 [-89.40,14.84]	229.00 [64.59,169.47]	0.65 [0.05,0.96]
	3 and 4*	8 (19.51)	385.38 (32.77)	355.94 (63.05)	–	–	–	–	–	–	–

AHI = arm hair index; BMI = body mass index; CI= confidence interval; ICC = intraclass correlation; LOA = limits of agreement; SD = standard deviation; TST = total sleep time; * subgroups < 10

Supplement Table 5.

Summary of performance metrics for WASO according to subgroups.

Variable	Subgroup	n (%)	Device Mean (SD)	PSG Mean (SD)	Assumptions	Bias [95% CI]	Proportional Bias	Proportional Bias 95% CI	LOA lower [95% CI]	LOA upper [95% CI]	ICC [95% CI]
Age	**18-40 yrs**	25 (61)	44.68 (47.55)	48.2 (44.41)	Prop Bias = T Normality = F Heteroscedasticity = F	-3.52 [-16.48, 13.60]	11.76 - (0.32 x PSG)	Intercept = [0.07, 29.58] Slope = [-0.47, -0.04]	-82.91 [-128.63,-0.20]	70.91 [25.05, 153.48]	0.63 [0.32,0.89]
	>40 yrs	16 (39)	60.75 (33.81)	85.41 (51.02)	Prop Bias = F Normality = F Heteroscedasticity = F	-24.66 [-47.06,-7.09]	-24.66	[-47.06,-7.09]	-71.81 [-186.0,-63.26]	91.81 [-22.82, 100.21]	0.41 [0,0.83]
Sex	**Female**	23 (56.1)	49.11 (43.77)	57.63 (49.57)	Prop Bias = T Normality = F Heteroscedasticity = F	-8.52 [-28.91, 12.24]	27.95 - (0.63 x PSG)	Intercept = [4.05, 54.09] Slope = [-0.86, -0.14]	-105.34 [-194.29,-18.91]	93.34 [3.81, 180.63]	0.4 [0.13,0.83]
	Male	18 (43.9)	53.31 (43.13)	69.22 (51.15)	Prop Bias = T Normality = T Heteroscedasticity = F	-15.92 [-27.17,-5.42]	2.17 - (0.26 x PSG)	Intercept = [-11.33, 12.63], Slope = [-0.45, -0.01]	-55.42 [-102.84,-30.56]	41.42 [-5.90, 66.27]	0.81 [0.64,0.9]
BMI	**<25**	30 (73.2)	55.87 (46.02)	68.58 (55.18)	Prop Bias = T Normality = F Heteroscedasticity = F	-12.72 [-28.62, 4.02]	21.65 - (0.5 x PSG)	Intercept = [2.68, 45.33] Slope = [-0.74, -0.2]	-96.34 [-183.94,-22.25]	84.34 [-3.85, 159.51]	0.56 [0.28,0.86]

	>=25	11 (26.8)	37.55 (31.35)	46.73 (27.85)	Prop Bias = F Normality = T Heterosced asticity = T	-9.18 [-22.18, 4.41]	-9.18	[-22.18,4.41]	bias - 2.46(0.72 + 0.33 x PSG) Intercept = [-11.4, 12.84] Slope = [0, 0.66]	bias + 2.46(0.72 + 0.33 x PSG Intercept = [-11.4, 12.84] Slope = [0, 0.66]	0.63 [0,0.83]
Skin Tone	**I-II**	21 (51.2)	53.64 (49.59)	60.10 (44.02)	Prop Bias = T Normality = F Heterosced asticity = F	-7.55 [-18.92, 6.73]	13.9 - (0.34 x PSG)	Intercept = [-6.89, 45.32] Slope = [-0.64, -0.04]	-83.44 [-136.77,-4.91]	84.44 [31.01,163.01]	0.58 [0.21,0.88]
	III-VI	20 (48.8)	48.12 (35.86)	65.47 (56.59)	Prop Bias = T Normality = F Heterosced asticity = T	-24.85 [-58.4,-1.85]	18.13 - (0.54 x PSG)	Intercept = [0.75, 33.01] Slope = [-0.75, -0.15]	bias - 2.46(7.55 + 0.16 x PSG) Intercept = [2.13, 14.65], Slope = [0.02, 0.27]	bias + 2.46(7.55 + 0.16 x PSG Intercept = [2.13, 14.65] Slope = [0.02, 0.27]	0.6 [0.28,0.88]
Arm Hair Index	**1***	17 (41.5)	50.88 (47.57)	59.79 (43.54)	Prop Bias = T Normality = F Heterosced asticity = F	-8.91[-26.118 ,15.35]	18.8 - (0.46 x PSG)	Intercept = [-1.23, 57.06] Slope = [-0.76, -0.26]	-96.33 [-142.43,-10.65]	84.33 [38.12, 170.67]	0.47 [0.1,0.84]
	2*	16 (39.0)	48.53 (43.85)	58.38 (55.36)	Prop Bias = F Normality = F Heterosced asticity = T	-9.84[-32.31, 5.85]	-9.84	[-32.31,5.85]	bias - 2.46(2.37 + 0.34 x PSG) Intercept = [-3.77, 9.5] Slope = [0.08, 0.5]	bias + 2.46(2.37 + 0.34 x PSG) Intercept = [-3.77, 9.5] Slope = [0.08, 0.5]	0.63 [0.28,0.95]
	3 and 4*	8 (19.51)	55.94 (35.04)	77.62 (55.25)	–	–	–	–	–	–	–

AHI = arm hair index; BMI = body mass index; CI= confidence interval; ICC = intraclass correlation; LOA = limits of agreement; SD = standard deviation; WASO = wake after sleep onset; * subgroups < 10

Supplement Table 6.

Summary of performance metrics for SE according to subgroups.

Variable	Subgroup	n (%)	Device Mean (SD)	PSG Mean (SD)	Assumptions	Bias [95% CI]	Proportional Bias	Proportional Bias 95% CI	LOA lower [95% CI]	LOA upper [95% CI]	ICC [95% CI]
Age	**18-40 yrs**	25 (61)	85.65 (9.98)	83.52 (11.91)	Prop Bias = T Normality = F Heteroscedasticity = F	2.13 [-0.94,5.10]	32.84 - (0.37 x PSG)	Intercept = [13.73, 49.15] Slope = [-0.56, -0.14]	-12.53 [-27.21,0.27]	18.35 [3.61,31.15]	0.73 [0.46,0.92]
	>40 yrs	16 (39)	83.58 (7.29)	78.84 (11.16)	Prop Bias = T Normality = F Heteroscedasticity = T	4.74 [1.13,9.04]	50.05 - (0.57 x PSG)	Intercept = [7.85, 69] Slope = [-0.8, -0.07]	bias - 2.46(17.57 + -0.17 x PSG) Intercept = [3.88, 32.5] Slope = [-0.35, -0.01]	bias + 2.46(17.57 + -0.17 x PSG) Intercept = [3.88, 32.5] Slope = [-0.35, -0.01]	0.5 [0.15,0.84]
Sex	**Female**	23 (56.1)	85.18 (9.59)	83.51 (11.14)	Prop Bias = F Normality = F Heteroscedasticity = F	1.68 [-1.86,5.24]	1.68	[-1.86,5.24]	-14.43 [-30.52,-0.34]	20.25 [4.11,34.21]	0.63 [0.29,0.93]
	Male	18 (43.9)	84.41 (8.41)	79.38 (12.33)	Prop Bias = T Normality = T Heteroscedasticity = F	5.03 [2.04,8.18]	38.52 - (0.42 x PSG)	Intercept = [21.41, 52.15] Slope = [-0.58, -0.22]	-12.40 [-16.19,3.63]	14.45 [10.69,30.50]	0.69 [0.46,0.86]

BMI	**<25**	30 (73.2)	83.99 (9.6)	80.78 (13)	Prop Bias = T Normality = F Heteroscedasticity = F	3.21 [0.03,6.49]	41.55 - (0.47 x PSG)	Intercept = [19.13, 57.45] Slope = [-0.67, -0.21]	-15.01 [-29.86,4.03]	20.83 [6.01,39.66]	0.65 [0.41,0.86]
	>=25	11 (26.8)	87.17 (6.9)	84.18 (7.02)	Prop Bias = F Normality = T Heteroscedasticity = F	3.00 [0.52,5.63]	3.00	[0.52,5.63]	-4.78 [-11.48,0.87]	13.12 [6.42,18.7]	0.7 [0,0.92]
Skin Tone	**I-II**	21 (51.2)	84.90 (9.74)	82.04 (11.95)	Prop Bias = T Normality = F Heteroscedasticity = F	3.13 [0.42,5.68]	37.54 - (0.42 x PSG)	Intercept = [18.54, 58.62] Slope = [-0.69, -0.19]	-14.66 [-29.27,-0.58]	18.81 [4.09,32.92]	0.66 [0.29,0.89]
	III-VI	20 (48.8)	84.79 (8.37)	81.33 (11.76)	Prop Bias = F Normality = F Heteroscedasticity = F	3.22 [-1.19,9.95]	3.45	[0.46,7.15	-12.48 [-19.83,4.37]	18.30 [10.98,35.31]	0.65 [0.38,0.91]
Arm Hair Index	**1**	17 (41.5)	85.43 (10.67)	83.95 (9.78)	Prop Bias = F Normality = F Heteroscedasticity = F	1.48[-2.56,4.57]	1.48	[-2.56,4.57]	-12.44 [-29.52,-5.81]	18.25 [1.28,24.90]	0.7 [0.22,0.92]
	2	16 (39.0)	84.91 (8.21)	81.56 (12.6)	Prop Bias = T Normality = F Heterosce	3.34 [0.37,7.59]	42.7 - (0.48 x PSG)	Intercept = [7.99, 63.87] Slope = [-0.72, -	bias - 2.46(20.08 + -0.2 x PSG) Intercept = [11.08, 24.8] Slope = [-0.26, -0.1]	bias + 2.46(20.08 + -0.2 x PSG) Intercept = [11.08, 24.8]	0.68 [0.37,0.95]

				dasticity = T			0.08]		Slope = [-0.26, -0.1]	
3 and 4*	8 (19.51)	83.48 (7.32)	77.16 (13.73)	–	–	–	–	–	–	–

AHI = arm hair index; BMI = body mass index; CI= confidence interval; ICC = intraclass correlation; LOA = limits of agreement; SD = standard deviation; SE = sleep efficiency; * subgroups < 10

Supplement Table 7.

Summary of performance metrics for SOL according to subgroups.

Variable	Subgroup	n (%)	Device Mean (SD)	PSG Mean (SD)	Assumptions	Bias [95% CI]	Proportional Bias	Proportional Bias 95% CI	LOA lower [95% CI]	LOA upper [95% CI]	ICC [95% CI]
Age	**18-40 yrs**	25 (61)	23.3 (20.48)	29.52 (22.07)	Prop Bias = F Normality = F Heteroscedasticity = T	-6.22 [-11.48,-2.02]	-6.22	[-11.48,-2.02]	bias - 2.46(-0.09 + 0.27 x PSG) Intercept = [-3.28, 2.54], Slope = [0.15, 0.41]	bias + 2.46(-0.09 + 0.27 x PSG Intercept = [-3.28, 2.54], Slope = [0.15, 0.41]	0.78 [0.44,0.97]
	>40 yrs	16 (39)	20.44 (26.62)	19.03 (15.98)	Prop Bias = F Normality = F Heteroscedasticity = F	1.41 [-11.56,17.38]	-3.59	[-23.81,21.06]	-119.60 [-147.86,-12.57]	63.60 [35.26,170.26]	0.04 [0,0.89]
Sex	**Female**	23 (56.1)	23.87 (24.73)	23.24 (17.79)	Prop Bias = T Normality = F Heteroscedasticity = F	0.63 [-7.46,11.24]	10.1-0.41 x PSG	Intercept = [-2.25, 33.15] Slope = [-1.12, -0.02]	-53.60 [-73.00,0.94]	38.60 [19.44,93.00]	0.4 [0,0.95]
	Male	18 (43.9)	20.03 (20.53)	28.22 (23.5)	Prop Bias = F Normality = F Heteroscedasticity = T	-8.19 [-16.86,-1.31]	-8.19	[-16.861,-1.306]	bias - 2.46(-0.62 + 0.38 x PSG) Intercept = [-4.23, 2.17] Slope = [0.24, 0.56]	bias + 2.46(-0.62 + 0.38 x PSG Intercept = [-4.23, 2.17] Slope = [0.24, 0.56]	0.62 [0,0.97]

BMI	**<25**	30 (73.2)	22.48 (22.79)	25.05 (19.3)	Prop Bias = T Normality = F Heteroscedasticity = F	-2.57 [-10.32,6.32]	11.45 + -0.56 x PSG	Intercept = [-1.18, 28.35] Slope = [-1.06, -0.06]	-54.04 [-88.27,-6.33]	39.04 [4.47,87.29]	0.36 [0,0.89]
	>=25	11 (26.8)	21.36 (23.86)	26.45 (24.04)	Prop Bias = F Normality = F Heteroscedasticity = T	-5.09 [-13.82,0.59]	-5.09	[-13.82,0.59]	bias - 2.46(0.18 + 0.29 x PSG) Intercept = [-5.78, 3.39] Slope = [0.14, 0.7]	bias + 2.46(0.18 + 0.29 x PSG Intercept = [-5.78, 3.39] Slope = [0.14, 0.7]	0.82 [0,0.99]
Skin Tone	**I-II**	21 (51.2)	19.71 (19.94)	26.29 (23.49)	Prop Bias = T Normality = F Heteroscedasticity = T	-7.21 [-13.29,-2.26]	4.08 + -0.41 x PSG	Intercept = [-2.53, 12.3] Slope = [-0.92, -0.01]	bias - 2.46(0.45 + 0.35 x PSG) Intercept = [-2.71, 2.97] Slope = [0.22, 0.53]	bias + 2.46(0.45 + 0.35 x PSG Intercept = [-2.71, 2.97] Slope = [0.22, 0.53]	0.65 [0.03,0.96]
	III-VI	20 (48.8)	24.77 (25.71)	24.52 (17.06)	Prop Bias = F Normality = F Heteroscedasticity = F	9.05 [-1.90,29.05]	0.25	[-8.6,12.45]	-56.63 [-76.957, -2.9]	41.63 [21.15,95.77]	0.34 [0,0.96]
Arm Hair Index	**1**	17 (41.5)	23.03 (27.84)	20.21 (16.92)	Prop Bias = F Normality = F Heteroscedasticity = F	2.82 [-5.82,15.77]	2.82	[-5.82,15.765]	-57.12 [-64.47,2.21]	42.12 [34.737,101.8]	0.39 [0,0.95]
	2	16 (39.0)	21.88 (18.5)	29.09 (20.46)	Prop Bias = F Normality = F Heteroscedasticity =	-7.22 [-17.13,0.03]	-7.22	[-17.13,0.03]	bias - 2.46(-1.98 + 0.45 x PSG) Intercept = [-7.99, 2.6] Slope = [0.24, 0.61]	bias + 2.46(-1.98 + 0.45 x PSG Intercept = [-7.99, 2.6] Slope = [0.24,	0.49 [0,0.99]

				T						0.61]	
3 and 4*	8 (19.51)	21(21.42)	29.19(26.58)	–	–	–	–	–	–	–	

AHI = arm hair index; BMI = body mass index; CI= confidence interval; ICC = intraclass correlation; LOA = limits of agreement; SD = standard deviation; SOL = sleep onset latency; * subgroups < 10

Supplement Table 8.

Summary of performance metrics for NAWK according to subgroups.

Variable	Subgroup	n (%)	Device Mean (SD)	PSG Mean (SD)	Mean Difference [95%CI]	Kappa [95% CI]	ICC [95% CI]
Age	**18-40 yrs**	25 (61)	1.48(1.96)	1.28(1.54)	0.2[-0.52,0.96]	0.29[0.09,0.52]	0.43[0.13,0.7]
	>40 yrs	16 (39)	2.81(1.91)	2.69(1.89)	0.12[-0.56,0.81]	0.49[0.17,0.66]	0.7[0.44,0.84]
Sex	**Female**	23 (56.1)	2(1.76)	1.74(1.71)	0.26[-0.39,0.96]	0.41[0.21,0.61]	0.55[0.23,0.78]
	Male	18 (43.9)	2(2.38)	1.94(1.95)	0.06[-0.72,0.89]	0.44[0.09,0.65]	0.65[0.25,0.85]
BMI	**<25**	30 (73.2)	2.4(2.08)	1.93(1.96)	0.47[-0.13,1.13]	0.39[0.22,0.58]	0.58[0.29,0.78]
	BMI >=25	11 (26.8)	0.91(1.45)	1.55(1.29)	-0.64[-1.18,-0.09]	0.47[0.05,0.79]	0.62[-0.27,0.87]
Skin Tone	**I-II**	21 (51.2)	2(2.1)	1.62(1.6)	0.38[-0.38,1.14]	0.36[0.09,0.61]	0.49[0.07,0.77]
	III-VI	20 (48.8)	2(2)	2.05(2.01)	-0.05[-0.7,0.65]	0.56[0.3,0.73]	0.71[0.46,0.87]

Arm Hair Index	**1**	17 (41.5)	2.24(1.95)	2(1.87)	0.24[-0.53,1.12]	0.39[0.12,0.62]	0.57[0.15,0.84]
	2	16 (39.0)	1.81(2.26)	1.5(1.79)	0.31[-0.38,1.06]	0.56[0.36,0.81]	0.7[0.42,0.89]
	3 and 4*	8 (19.51)	1.88(1.89)	2.12(1.81)	–	–	–

CI= confidence interval; ICC = intraclass correlation; LOA = limits of agreement; NAWK = number of awakenings; SD = standard deviation; * subgroups < 10

Supplement Table 9.

Summary of performance metrics for duration of light sleep according to subgroups.

Variable	Subgroup	n (%)	Device Mean (SD)	PSG Mean (SD)	Assumptions	Bias [95% CI]	Proportional Bias	Proportional Bias 95% CI	LOA lower [95% CI]	LOA upper [95% CI]	ICC [95% CI]
Age	**18-40 yrs**	25 (61)	242.26 (45.22)	233.76 (49.14)	Prop Bias = T Normality = F Heterosceda sticity = F	8.50 [-2.60,19.52]	66.42 - (0.25 x PSG)	Intercept = [20.15, 127.66] Slope = [-0.52, -0.06]	-49.38 [-94.59,-7.81]	63.38 [18.07,104.89]	0.8 [0.56,0.91]
	>40 yrs	16 (39)	247.81 (45.29)	251.41 (49.06)	Prop Bias = F Normality = F Heterosceda sticity = F	-3.59 [-23.81,21.06]	-3.59	[-23.81,21.06]	-119.60 [-147.86,-12.57]	63.60 [35.26,170.26]	0.51 [0,0.83]
Sex	**Female**	23 (56.1)	248.43 (48.06)	245.7 (57.12)	Prop Bias = T Normality = T Heterosceda sticity = F	2.74 [-12.94,20.22]	102.16 - (0.4 x PSG)	Intercept = [18.41, 186.98] Slope = [-0.76, -0.09]	-73.52 [-137.47,-0.31]	87.52 [23.41,160.73]	0.7 [0.32,0.89]
	Male	18 (43.9)	239.31 (40.94)	234.19 (37.55)	Prop Bias = F Normality = T Heterosceda sticity = F	5.11 [-9.44,19.11]	5.11	[-9.44,19.11]	-59.73 [-106.30,-13.76]	63.73 [17.10,109.64]	0.67 [0.11,0.86]
BMI	**<25**	30 (73.2)	244.13 (46.71)	236.55 (54.63)	Prop Bias = T Normality = F Heterosceda sticity = F	7.58 [-4.62,20.88]	91.71 - (0.36 x PSG)	Intercept = [29.06, 167.53] Slope = [-0.68, -0.11]	-64.22 [-116.71,6.78]	78.22 [25.88,149.09]	0.73 [0.39,0.89]

	>=25	11 (26.8)	245.23 (41.07)	251.82 (29.45)	Prop Bias = F Normality = T Heterosceda sticity = F	-6.59 [-27.64,14.05]	-6.59	[-27.637,14.045]	-53.71 [-132.14,-33.56]	93.71 [15.37,113.91]	0.43 [0,0.73]
Skin Tone	**I-II**	21 (51.2)	246.29 (41.91)	238.57 (51.44)	Prop Bias = T Normality = T Heterosceda sticity = F	4.63 [-5.76,14.66]	86.05 - (0.33 x PSG)	Intercept = [36.38, 163.34] Slope = [-0.68, -0.13]	-67.57 [-93.60,-8.87]	46.57 [20.64,105.21]	0.79 [0.43,0.9]
	III-VI	20 (48.8)	242.48 (48.6)	242.82 (48.11)	Prop Bias = F Normality = T Heterosceda sticity = F	1.15 [-27.95,36.15]	-0.35	[-18.05,19.08]	-78.82 [-145.68,-4.78]	92.82 [25.83,167.11]	0.59 [0.11,0.86]
Arm Hair Index	**1**	17 (41.5)	253.26 (45.06)	258.68 (47.84)	Prop Bias = F Normality = T Heterosceda sticity = F	-5.41 [-17.41,6.00]	-5.41	[-17.41,6.0]	-43.42 [-98.39,-16.77]	57.42 [2.65,84.26]	0.84 [0.56,0.94]
	2	16 (39.0)	228.31 (43.2)	226.09 (48.96)	Prop Bias = T Normality = T Heterosceda sticity = F	2.22 [-18.85,26.16]	130.7 - (0.57 x PSG)	Intercept = [4.13, 237.95] Slope = [-1.05, -0.04]	38.21 [-149.79,-8.33]	221.79 [34.18,174.80]	0.48 [0,0.84]
	3 and 4*	8 (19.51)	257.88 (42.66)	231.44 (46.19)	–	–	–	–	–	–	–

AHI = arm hair index; BMI = body mass index; CI= confidence interval; ICC = intraclass correlation; LOA = limits of agreement; SD = standard deviation; * subgroups < 10

Supplement Table 10.

Summary of performance metrics for duration of deep sleep according to subgroups.

Variable	Subgroup	n (%)	Device Mean (SD)	PSG Mean (SD)	Assumptions	Bias [95% CI]	Proportional Bias	Proportional Bias 95% CI	LOA lower [95% CI]	LOA upper [95% CI]	ICC [95% CI]
Age	**18-40 yrs**	25 (61)	66.28(22.36)	71(26.31)	Prop Bias = T Normality = T Heterosceda sticity = F	-4.72[-14.96,5.08]	40.75 - (0.64 x PSG)	Intercept = [21.44, 72.14] Slope = [-1.06, -0.37]	-41.65 [-103.31,-17.85]	61.65 [0.02,85.46]	0.4 [0,0.68]
	>40 yrs	16 (39)	69(18.38)	50.7(24.43)	Prop Bias = T Normality = T Heterosceda sticity = F	18.3[6.13,31.33]	57.87 - (0.78 x PSG)	Intercept = [37.93, 78.68] Slope = [-1.08, -0.52]	-53.83 [-61.94,10.96]	47.83 [39.65,112.64]	0.07 [0,0.59]
Sex	**Female**	23 (56.1)	68.63(21.91)	70.54(26.76)	Prop Bias = T Normality = T Heterosceda sticity = F	-1.91[-13.70,9.57]	52.44 - (0.77 x PSG)	Intercept = [25.89, 89.85] Slope = [-1.26, -0.43]	-47.73 [-109.74,-9.54]	67.73 [5.38,105.91]	0.27 [0,0.65]
	Male	18 (43.9)	65.5(19.58)	53.71(25.39)	Prop Bias = T Normality = T Heterosceda sticity = F	11.79[0.59, 23.71]	49.04 - (0.69 x PSG)	Intercept = [29.61, 69.19] Slope = [-1.04, -0.35]	-66.81 [-72.82,3.96]	31.81 [25.79,102.49]	0.29 [0,0.64]
BMI	**<25**	30 (73.2)	67.84(22.35)	64.69(26.35)	Prop Bias = T Normality = T Heterosceda sticity = F	3.16[-6.38,12.88]	45.74 - (0.66 x PSG)	Intercept = [26.49, 69.43] Slope = [-1, -0.37]	-42.58 [-98.55,-4.45]	62.58 [6.64,100.74]	0.39 [0,0.69]

	>=25	11 (26.8)	65.86(16.65)	59.95(30.35)	Prop Bias = F Normality = T Heterosceda sticity = F	5.91[-11.96,24.91]	62.14 - (0.94 x PSG)	Intercept = [36.88, 104.01] Slope = [-1.5, -0.59]	-73.53 [-97.55,-11.46]	55.53 [31.68,117.70]	0.08 [0,0.55]
Skin Tone	**I-II**	21 (51.2)	64.67(23.22)	64.14(32)	Prop Bias = T Normality = T Heterosceda sticity = F	2.55[-7.17,12.32]	43.13 - (0.66 x PSG)	Intercept = [24.02, 62.88] Slope = [-0.97, -0.38]	-22.00 [-107.11,-13.15]	94.00 [8.94,102.86]	0.44 [0.02,0.71]
	III-VI	20 (48.8)	70.21(17.8)	62.55(21.52)	Prop Bias = T Normality = T Heterosceda sticity = F	8[-9.25,26.4]	66.43 - (0.94 x PSG)	Intercept = [38.87, 97.06] Slope = [-1.4, -0.55]	-42.73 [-83.96,2.22]	62.73 [21.54,107.62]	0.03 [0,0.55]
Arm Hair Index	**1**	17 (41.5)	69.32(23.96)	63.06(28.68)	Prop Bias = T Normality = T Heterosceda sticity = F	6.26[-7.97,19.65]	50.99 - (0.71 x PSG)	Intercept = [19.02, 85.72] Slope = [-1.23, -0.18]	-49.39 [-105.76,-10.55]	69.39 [12.59,108.0]	0.32 [0,0.72]
	2	16 (39.0)	70.56(18.71)	67.47(26.98)	Prop Bias = F Normality = T Heterosceda sticity = F	3.09[-10.72,17.59]	61.76 - (0.87 x PSG)	Intercept = [33.47, 93.91] Slope = [-1.27, -0.53]	52.42 [-93.42,-4.15]	64.42 [23.42,112.52]	0.17 [0,0.58]
	3 and 4*	8 (19.51)	54.93(13.23)	54.86(25.73)	–	–	–	–	–	–	–

CI= confidence interval; ICC = intraclass correlation; LOA = limits of agreement; REM = rapid eye movement; SD = standard deviation; * subgroups < 10

Supplement Table 11.

Summary of performance metrics for duration of REM sleep according to subgroups.

Variable	Subgroup	n (%)	Device Mean (SD)	PSG Mean (SD)	Assumptions	Bias [95% CI]	Proportional Bias	Proportional Bias 95% CI	LOA lower [95% CI]	LOA upper [95% CI]	ICC [95% CI]
Age	**18-40 yrs**	25 (61)	89.04 (26.48)	83.54 (29.34)	Prop Bias = T Normality = T Heteroscedasticity = F	5.50 [-2.96,13.62]	35.64 - (0.36 x PSG)	Intercept = [15.04, 59.16] Slope = [-0.64, -0.11]	-46.09 [-74.33,-0.58]	38.09 [10.06,83.62]	0.69 [0.4,0.86]
	>40 yrs	16 (39)	90.03 (14.11)	80.84 (18.61)	Prop Bias = T Normality = T Heteroscedasticity = F	9.19 [-0.66,18.31]	73.37 - (0.79 x PSG)	Intercept = [42.41, 100.79] Slope = [-1.16, -0.4]	-17.84 [-64.80,-3.11]	60.84 [13.84,75.53]	0.17 [0,0.52]
Sex	**Female**	23 (56.1)	90.46 (22.18)	83.87 (24.4)	Prop Bias = T Normality = T Heteroscedasticity = F	6.59 [-1.52,13.96]	41.32 - (0.41 x PSG)	Intercept = [21.41, 77.66] Slope = [-0.81, -0.14]	-42.69 [-70.30,-2.90]	34.69 [7.19,74.15]	0.61 [0.18,0.82]
	Male	18 (43.9)	88.11 (22.93)	80.72 (27.37)	Prop Bias = T Normality = T Heteroscedasticity = F	7.39 [-2.69,17.33]	46.87 - (0.49 x PSG)	Intercept = [15.36, 77.75] Slope = [-0.93, -0.14]	-23.74 [-71.96,-1.68]	64.74 [16.54,87.00]	0.56 [0.01,0.81]
BMI	**<25**	30 (73.2)	88.95 (21.32)	84.77 (27.83)	Prop Bias = T Normality = T Heteroscedasticity = F	4.18 [-3.43,11.63]	47.14 - (0.51 x PSG)	Intercept = [27.05, 70.74] Slope = [-0.79, -0.27]	-46.26 [-78.77,-1.94]	38.26 [5.84,82.90]	0.61 [0.27,0.81]
	>=25	11 (26.8)	90.73 (25.72)	76.27 (17)	Prop Bias = F Normality = T Heteroscedasticity = F	14.45 [4.64,24.23]	14.45	[4.64,24.23]	-24.44 [-41.42,4.39]	42.44 [25.49,71.27]	0.51 [0,0.79]

Skin Tone	**I-II**	21 (51.2)	91.64 (22.17)	87.33 (28.24)	Prop Bias = T Normality = T Heteroscedasticity = F	7.18 [-0.40,14.28]	44.56 - (0.46 x PSG)	Intercept = [20.83, 77.09] Slope = [-0.83, -0.19]	-55.10 [-73.93,-4.93]	26.10 [7.46,76.40]	0.65 [0.19,0.86]
	III-VI	20 (48.8)	87.1 (22.68)	77.4 (21.74)	Prop Bias = T Normality = T Heteroscedasticity = F	6.2 [-4.75,18.20]	42.3 - (0.42 x PSG)	Intercept = [14.08, 72.37] Slope = [-0.83, -0.04]	-45.12 [-66.89,0.02]	37.12 [15.17,82.26]	0.48 [0.09,0.73]
Arm Hair Index	**1**	17 (41.5)	94.12 (22.7)	89.35 (28.54)	Prop Bias = T Normality = T Heteroscedasticity = F	4.76 [-6.30,14.82]	49.68 - (0.5 x PSG)	Intercept = [25.33, 99.76] Slope = [-1.08, -0.23]	-48.67 [-79.71,-9.34]	40.67 [9.36,79.98]	0.59 [0,0.84]
	2	16 (39.0)	89.47 (23.72)	78.19 (20.24)	Prop Bias = F Normality = T Heteroscedasticity = F	11.28 [4.13,18.84]	11.28	[4.13,18.84]	-13.97 [-39.30,6.50]	45.97 [20.61,66.55]	0.64 [0.29,0.84]
	3 and 4*	8 (19.51)	79.38 (16.7)	76.5 (27.7)	–	–	–	–	–	–	–

AHI = arm hair index; BMI = body mass index; CI= confidence interval; ICC = intraclass correlation; LOA = limits of agreement; SD = standard deviation; * subgroups < 10

medRxiv preprint doi: https://doi.org/10.1101/2024.09.10.24313425; this version posted September 11, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

Supplement Table 12.

As an additional metric to evaluate the performance of the VNW algorithm, we calculated intra-class correlation coefficients between the mean values of each measure in both devices.

Measure	PSG mean (SD)	VNW Mean (SD)	ICC [95% CI]
TST (min)	384.98 (60.85)	399(46.33)	0.68 [0.43, 0.88]
WASO (min)	62.72 (49.97)	50.95 (42.99)	0.59 [0.34, 0.84]
SE (%)	81.69 (11.71)	84.84 (8.99)	0.66 [0.45, 0.85]
SOL (min)	25.43 (20.37)	22.18 (22.79)	0.50 [0.11, 0.89]
NAWK (count)	2.17 (1.96)	2.14 (2.13)	0.61 [0.39, 0.77]
Light (min)	240.65 (49.27)	244.43 (44.76)	0.69 [0.40, 0.85]
Deep (min)	63.39 (27.19)	67.30 (20.75)	0.31 [0.00, 0.57]
REM (min)	82.49 (25.46)	89.43 (22.26)	0.59 [0.31, 0.77]

ICC = intraclass correlation; PSG = polysomnography; REM = rapid eye movement; SD = standard deviation; SE = sleep efficiency; SOL = sleep onset latency; TST = total sleep time; VNW=Verily Numetric Watch; WASO = wake after sleep onset