Accuracy of 11 Wearable, Nearable, and Airable Consumer Sleep Trackers: A Prospective and Multicenter Validation Study

Taeyoung Lee, Younghoon Cho, Kwang Su Cha, Jinhwan Jung, Jungim Cho, Hyunggug Kim, Daewoo Kim, Joonki Hong, Dongheon Lee, Moonsik Keum, Clete A. Kushida, In-Young Yoon, Jeong-Whun Kim

Submitted to: JMIR mHealth and uHealth
on: July 18, 2023

Table of Contents

Accuracy of 11 Wearable, Nearable, and Airable Consumer Sleep Trackers: A Prospective and Multicenter Validation Study

Taeyoung Lee[1*] MS; Younghoon Cho[1, 2*] MD; Kwang Su Cha[1] MS; Jinhwan Jung[1] PhD; Jungim Cho[1]; Hyunggug Kim[1] MS; Daewoo Kim[1] PhD; Joonki Hong[1] PhD; Dongheon Lee[1] MS; Moonsik Keum[2]; Clete A. Kushida[3]; In-Young Yoon[4]; Jeong-Whun Kim[4]

[1]ASLEEP Seoul KR
[2]Clionic Lifecare Clinic Seoul KR
[3]Stanford University Redwood City US
[4]Seoul National University Bundang Hospital Seongnam-si KR
[*]these authors contributed equally

Corresponding Author:
Jeong-Whun Kim
Seoul National University Bundang Hospital
Department of Otorhinolaryngology-Head and Neck Surgery, Seoul national University Bundang Hospital,
82, Gumi-ro 173beon-gil, Bundang-gu,
Seongnam-si
KR

Abstract

Background: Consumer sleep trackers (CSTs) have gained significant popularity, because they enabled individuals to conveniently monitor and analyze their sleep. However, limited studies have comprehensively validated the performance of widely used CSTs. Our study therefore investigated popular CSTs which based on various biosignals and algorithms by assessing the agreement with polysomnography.

Objective: This study aims to validate the accuracy of various types of CSTs through a comparison with PSG. Additionally, by including widely-used CSTs and conducting a multicenter study with a large sample size, the research seeks to provide comprehensive insights into the performance and applicability of these CSTs for sleep monitoring.

Methods: The study analyzed 11 commercially available CSTs including five Wearables (Google Pixel Watch, Galaxy Watch 5, Fitbit Sense 2, Apple Watch 8, and Oura Ring 3), three Nearables (Withings Sleep Tracking Mat, Google Nest Hub 2, and Amazon Halo Rise), and three Airables (SleepRoutine, SleepScore, and Pillow). The 11 CSTs were divided into two groups, ensuring maximum inclusion while avoiding interference between the CSTs within each group. Each group (comprising 8 CSTs) was also compared via polysomnography.

Results: The study enrolled 75 participants from a tertiary hospital and a primary sleep-specialized clinic in Korea. Across two centers, we collected a total 3890 hours of sleep sessions based on the 11 CSTs along with 543 hours of PSG recordings at the two centers. Each CST sleep recording covered an average of 353 hours. We analyzed a total of 349,114 epochs from the 11 CSTs compared with PSG, where epoch-by-epoch agreement in sleep stage classification showed substantial performance variation. More specifically, the highest macro f1 score was 0.69, while the lowest macro f1 score was 0.26.

Conclusions: Our study showed that among the 11 CSTs examined, specific CSTs showed substantial agreement with PSG, indicating their potential application in sleep monitoring, while other CSTs were partially consistent with PSG. This study offers insights into the strengths of each CST within the three different classes.

(JMIR Preprints 18/07/2023:50983)
DOI: https://doi.org/10.2196/preprints.50983

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
✔ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.
2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?
✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Authors (Corresponding authors are not included here, separately described below)

Taeyoung Lee**, MS[1] 0000-0002-2182-8654
Younghoon Cho**, MD[1,2] 0000-0002-9542-3916
Kwang Su Cha, MS[1] 0000-0002-7710-8648
Jinhwan Jung, PhD[1] 0000-0002-2455-4206
Jungim Cho[1] 0009-0005-5608-819X
Hyunggug Kim, MS[1] 0000-0002-6579-5313
Daewoo Kim, PhD[1] 0000-0003-2475-6956
Joonki Hong, PhD[1] 0000-0002-8378-3332
Dongheon Lee, MS[1] 0000-0003-4660-6174
Moonsik Keum[2] 0009-0001-4539-1493
Clete A. Kushida[3] 0000-0002-9430-3752
** two authors contributed equally.
[1] ASLEEP, 415, Teheran-ro, Gangnam-gu, Seoul, Republic of Korea, 06160
[2] Clionic Lifecare Clinic, 251, Eulji-ro, Jung-gu, Seoul, Republic of Korea, 04564
[3] Department of Psychiatry & Behavioral Sciences, Stanford University Medical Center, Palo Alto, CA 94305, USA

Corresponding Author(s)

In-Young Yoon, Department of Psychiatry, Seoul National University Bundang Hospital, Seongnam, Republic of Korea, 82, Gumi-ro 173beon-gil, Bundang-gu, Seongnam-si, Gyeonggi-do, Republic of Korea, 13620, ORCID 0000-0002-3995-8238
Jeong-Whun Kim, Department of Otorhinolaryngology-Head and Neck Surgery, Seoul national University Bundang Hospital, 82, Gumi-ro 173beon-gil, Bundang-gu, Seongnam-si, Gyeonggi-do, Republic of Korea, 13620, ORCID 0000-0003-4858-3316

Accuracy of 11 Wearable, Nearable, and Airable Consumer Sleep Trackers: A Prospective and Multicenter Validation Study

Abstract

Background: Consumer sleep trackers (CSTs) have gained significant popularity, because they enabled individuals to conveniently monitor and analyze their sleep. However, limited studies have comprehensively validated the performance of widely used CSTs. Our study therefore investigated popular CSTs which based on various biosignals and algorithms by assessing the agreement with polysomnography.

Objective: This study aims to validate the accuracy of various types of CSTs through a comparison with in-lab polysomnography (PSG). Additionally, by including widely-used CSTs and conducting a multicenter study with a large sample size, the research seeks to provide comprehensive insights into the performance and applicability of these CSTs for sleep monitoring in a hospital environment.

Methods: The study analyzed 11 commercially available CSTs including five Wearables (Google Pixel Watch, Galaxy Watch 5, Fitbit Sense 2, Apple Watch 8, and Oura Ring 3), three Nearables (Withings Sleep Tracking Mat, Google Nest Hub 2, and Amazon Halo Rise), and three Airables (SleepRoutine, SleepScore, and Pillow). The 11 CSTs were divided into two groups, ensuring maximum inclusion while avoiding interference between the CSTs within each group. Each group (comprising 8 CSTs) was also compared via polysomnography.

Results: The study enrolled 75 participants from a tertiary hospital and a primary sleep-specialized clinic in Korea. Across two centers, we collected a total 3890 hours of sleep sessions based on the 11 CSTs along with 543 hours of PSG recordings at the two centers. Each CST sleep recording covered an average of 353 hours. We analyzed a total of 349,114 epochs from the 11 CSTs compared with PSG, where epoch-by-epoch agreement in sleep stage classification showed substantial performance variation. More specifically, the highest macro f1 score was 0.69, while the lowest macro f1 score was 0.26. Various sleep trackers exhibited diverse performances across sleep stages, with SleepRoutine excelling in wake and REM stages, while Wearables like Google Pixel Watch and Fitbit Sense 2 showed superiority in the deep stage. There was a distinct trend in sleep measure estimation according to the type of device. Wearables showed high proportional-bias in sleep efficiency, while nearables exhibited high proportional-bias in sleep latency. Subgroup analyses of sleep trackers revealed variations in macro F1 scores based on factors such as BMI, sleep efficiency, and AHI, while the differences between male and female subgroups were minimal.

Conclusions: Our study showed that among the 11 CSTs examined, specific CSTs showed substantial agreement with PSG, indicating their potential application in sleep monitoring, while other CSTs were partially consistent with PSG. This study offers insights into the strengths of each CST within the three different classes to individuals interested in wellness who wish to understand and proactively manage their own sleep.

Keywords: Consumer sleep Trackers (CST); Wearables; Nearables; Airables; Sleep monitoring; Sleep stage; Comparative study; Polysomnography (PSG); Multicenter study; Deep learning; AI

Introduction

With the growing recognition of the importance of sleep for overall health,[1] there has been a significant rise in public interest in monitoring sleep patterns using consumer sleep trackers (CSTs).[2–4] While the laboratory monitoring of sleep using the traditional sleep analyzing tool polysomnography (PSG) has several limitations associated with the need for cumbersome sensors,[5] CSTs facilitate individual monitoring of sleep at home using minimal equipment. Recently many big-tech companies, including Apple, Samsung, and Google, as well as healthcare startups like Withings, and Oura have released their own CSTs. These companies have made significant contributions to enhancing the performance of CSTs by integrating deep learning algorithms and biosignal sensing technology.[6,7] As a result, CSTs have emerged as accessible solutions for home sleep monitoring.[2,6,8–11] CSTs are widely used not only by individuals interested in wellness who wish to understand and proactively manage their own sleep, but also by those who want to self-check and screen for sleep disorders.

This study classified CSTs into three types: Wearables, Nearables, and Airables. Wearable devices or Wearables, such as smartwatches and ring-shaped devices, are generally worn by users to track sleep using sensors like PPGs and accelerometers.[6,12–16] Nearable devices or Nearables, placed near the body without direct contact, carry radar or mattress pads to detect subtle movements during sleep.[9,17] Airable devices or Airables utilize mobile phones to analyze sleep via built-in microphones or environmental sensors.[2,3,8] This classification is based on the measurement methods and biological signals utilized in each category.

Given the surge of diverse CSTs, it is necessary to conduct comprehensive and objective evaluation of their performance of these CSTs available in the market.[4,15,18–21] Some studies compared CSTs and alternative tools available for sleep analysis, such as electroencephalography headbands[4] or subjective sleep diaries[12] (without employing the gold standard PSG), which failed to validate the consistency between CSTs and PSG. Chinoy et al.[11] compared the performance of seven CSTs with PSG; Fatigue Science Readiband, Fitbit Alta HR, Garmin Fenix 5S, Garmin Vivosmart 3, EarlySense Live, ResMed S+, and SleepScore Max. However, this study showed limitations of recruitment from a single institution and exclusion of widely used CSTs available commercially.

To address these limitations, we conducted a multicenter study comparing widely used or newly released CSTs with in-lab PSG in a hospital setting. By simultaneously assessing multiple CSTs, we aimed to minimize bias and evaluate their performance across various metrics. Subgroup analysis was also performed to assess the impact of demographic factors on performance, including sex assigned at birth, apnea-hypopnea index (AHI), and body mass index (BMI). By performing the most extensive simultaneous comparison of widely-used CSTs, and conducting a multicenter study with diverse demographic groups, this study offers comprehensive insights into the performance and applicability of these CSTs for sleep monitoring.

Methods

Participants

Table 1 presents the demographic information of the study participants. A total of 75 individuals were recruited from Seoul National University Bundang Hospital (SNUBH) and Clionic Lifecare Clinic (CLC). Thirty-seven participants (27 males, 10 females) with scheduled PSG for sleep disorders at the hospital were recruited from SNUBH, while at CLC, 38 participants (12 males, 26 females) were recruited through an online platform. Both institutions used the same inclusion criteria, including participants aged between 19 years and 70 years experiencing subjective sleep discomfort. Individuals with uncontrolled acute respiratory conditions were excluded. Participant demographics revealed that the sample consisted of 52% males, with a mean age of 43.59 ± 14.10 years and a mean BMI of 23.90 ± 4.07 kg/□. Significant differences in sleep measures were observed between the two institutions, including time in bed (TIB), total sleep time (TST), wake after sleep onset (WASO), and AHI. Supplementary Table S1 presents the number of measurements and data collection success rate for each CST.

Table 1. Comparative analysis of participant demographics: SNUBH vs. CLC.

	Total	**SNUBH**	**CLC**	***P***
Number of participants	75	37	38	-
Male, % (n)	52.00 (39)	72.97 (27)	31.58 (12)	-
Age, years	43.59 ± 14.10	53.49 ± 11.96	33.95 ± 8.07	**<.001**
Body mass index, kg/m²	23.90 ± 4.07	24.64 ± 4.07	23.18 ± 3.98	.12
Time in bed, hours	7.24 ± 0.92	8.01 ± 0.31	6.49 ± 0.66	**<.001**
Total sleep time, hours	5.82 ± 1.33	6.26 ± 1.2	5.40 ± 1.33	**.0045**
Sleep latency, hours	0.27 ± 0.37	0.27 ± 0.38	0.26 ± 0.38	.93
Wake After Sleep Onset, hours	1.15 ± 1.2	1.48 ± 1.22	0.83 ± 1.23	**.02**
Sleep efficiency, %	81.00 ± 17.5	78.40 ± 15.54	83.54 ± 19.1	.20
Apnea hypopnea index	18.18 ± 20.39	26.56 ± 24.25	10.02 ± 10.99	**<.001**

Percentage (%) of total population (number of participants), mean ± standard deviation. All p-values were obtained using two-sample independent t-tests. The bold values represent statistical significance (*P* < .05). Abbreviations: SNUBH Seoul National Bundang Hospital, CLC Clionic Life Center.

Commercial Sleep Trackers

We evaluated 11 different CSTs in this study. The Wearables included ring-type devices (Oura Ring 3, Oura, Finland) and watch-type devices (Apple Watch 8 Apple Inc., USA; Galaxy Watch 5, Samsung Electronics Co. Ltd., South Korea; Fitbit Sense 2, Fitbit Inc., USA; Google Pixel Watch, Google LLC., USA). Nearables included pad-type devices (Withings Sleep Tracking Mat, Withings, France) and motion sensor devices (Amazon Halo Rise, Amazon.com Inc., USA; Google Nest Hub 2, Google LLC., USA). Mobile applications (SleepRoutine, ASLEEP, South Korea; SleepScore App, SleepScore Labs, USA; and Pillow, Neybox Digital Ltd., USA) with iPhone 12s and Galaxy S21s represented Airables. The selection of these devices was based on their popularity and availability in the market at the time of the study. The methods of usage and application for each sleep tracker were based on user instructions provided by the respective manufacturers. To mitigate a possible learning curve for each device, the researchers educated participants on how to use each device before measurements, and in the case of wearable CSTs, and ensured that the devices were properly fitted. During the study, software updates of all devices were performed on March 1st, 2023, to ensure that they were up-to-date, and automatic updates were disabled.

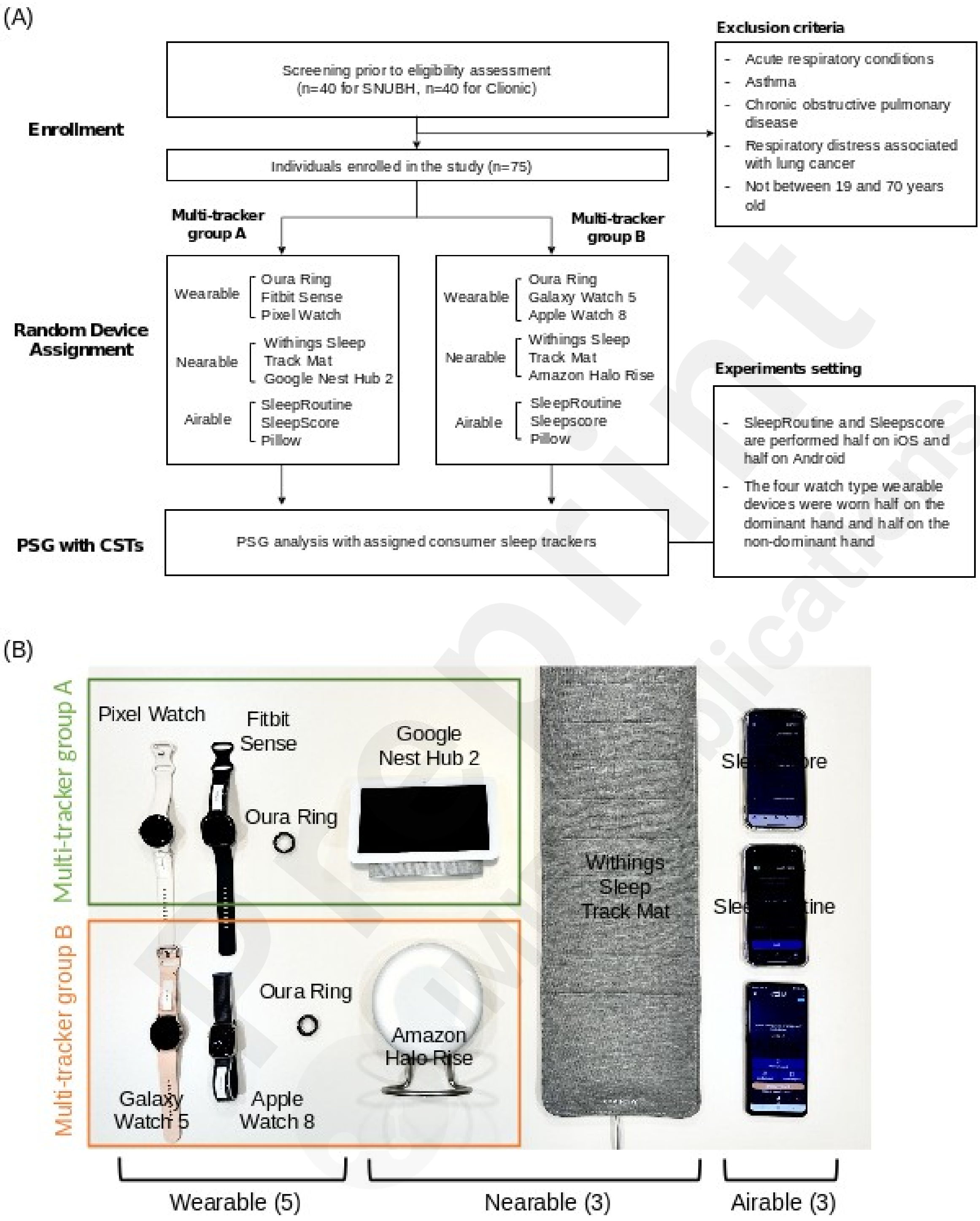
Study Design

This was a prospective cross-sectional study conducted to investigate the accuracy of various CSTs and PSG in analyzing sleep stages. It was conducted at two independent medical institutions in South Korea, namely SNUBH, a tertiary care hospital, and CLC, a primary care clinic. Ethical approval was obtained from the respective Institutional Review Boards (IRBs) of each institution (IRB No.: B-2302-908-301 from SNUBH, P01-202302-01-048 from CLC).

All participants were contacted by phone at least two days prior to participating in the PSG study and provided with instructions. On the day before and the day of the test, they were advised to abstain from alcohol and caffeine consumption, refrain from engaging in strenuous exercise, and were informed of the designated test time. These measures were taken to standardize participant behaviors and minimize the influence of potential confounding factors. On the designated test days, participants visited the hospitals and received detailed explanations about the study. They provided written informed consent and underwent PSG at each institution. PSG recordings were conducted in a controlled sleep laboratory environment in accordance with the guidelines recommended by American Academy of Sleep Medicine (AASM). [22] Two technicians independently interpreted the results, followed by a review by sleep physicians.

To address the issue of interference due to multiple CSTs sharing the same biosignals, the participants were divided into two groups in both of two medical institutions: multi-tracker group A and multi-tracker group, as illustrated in Figure 1. Specifically, at the SNUBH, multi-tracker group A consisted of 18 individuals, while multi-tracker group B comprised 19 individuals. Similarly, at CLC, each group was composed of 19 individuals. Across both institutions, the demographic statistics for participants in multi-tracker groups A and B demonstrated no significant differences across all metrics as provided in supplementary Table S2. Each group consisted of a combination of non-interfering CSTs. Specifically, Nearables Google Nest Hub2, and Amazon Halo Rise, which utilize similar radar sensors to detect motion, were allocated to different groups. In the case of Wearables, participants were allowed to wear a maximum of two watch devices, one on each wrist, which are a type of wearable, simultaneously. Consequently, Fitbit Sense 2 and Pixel Watch were assigned to group A, while Galaxy Watch 5 and Apple Watch 8 were assigned to group B. As a result, these devices were expected to yield approximately half of the intended measurements. Airables, which were available on both iOS and Android devices (SleepRoutine and SleepScore), were analyzed: half of them on iOS and the other half on Android. We used the Pillow on iOS, as it is not available on Android. The PSG and CST results were then compared and analyzed.

Figure 1. Flowchart outlining the experimental design.



(A) Experimental procedures involving subject enrollment, CST assignment, and experimental settings for simultaneous measurement of both CSTs and PSG. (B) Configuration of CSTs used in the experiment. Abbreviations: *CST,* consumer sleep tracker; *PSG,* polysomnography.

Statistical Methods and Evaluation Metrics

Two-sample independent t-test was employed to compare demographic information (Table 1) and sleep measures, and the significance was determined based on p-values below 0.05. The average sleep measurements were compared, and any proportional bias was assessed using Pearson correlation and Bland-Altman plot. The study used sensitivity, specificity, and F1 scores as evaluation metrics for sleep stage classification. Macro F1 score, weighted F1 score, and kappa values were used to summarize the results of evaluation, considering the imbalance in data classes, such as sleep stages. All statistical analysis and visualizations were conducted using Python 3 (version 3.9.16) and utilized of scikit-learn, matplotlib, and scipy libraries.

Data Preprocessing

Three main steps were followed in the data processing stage. First, raw sleep score data were extracted from each CST device, either through direct download via the manufacturer's app or the web portal, or by requesting raw data from the SleepRoutine device manufacturer. The sleep score codes were standardized across devices, with the wake stage assigned 0; the light stage denoted1; the deep stage, 2; and the REM stage, 3. The Apple Watch 8 used alternative expressions, such as “core-sleep” instead of light sleep.

The extracted data were synchronized in time to compare CST the results of sleep tracking with PSG accurately. The sleep stages measured by devices earlier than PSG scores were discarded. Conversely, for devices that started measuring after the PSG scoring, the sleep stages were marked as the wake stage until their measurement began. The end point of all device measurements was aligned with the end of PSG, resulting in consistent measurement of total time in bed (TIB) across the devices.

Because the sleep stages change at every epoch, the results may be inaccurate if the start time of the 30-second epoch differed. Therefore, the results of sleep stage involving all devices, including PSG, were segmented into 1-second intervals, and compared every second. This approach enabled a more precise comparison of sleep stage results between PSG and CST measurements and eliminated potential bias.

Results

Epoch-by-Epoch Analysis: overall performance

Table 2 presents the results of epoch-by-epoch agreements between PSG and each of the 11 CSTs under the sleep stage classifications. SleepRoutine (Airable) demonstrated the highest macro F1 score of 0.6863, which was closely followed by the Amazon Halo Rise (Nearable) with a macro F1 score of 0.6242. In terms of Cohen's kappa, a measure of inter-rater agreement, three Wearables (Google Pixel Watch, Galaxy Watch 5, and Fitbit Sense 2), one Nearable (Amazon Halo Rise), and one Airable (SleepRoutine), demonstrated moderate agreement with sleep stage classification (kappa 0.4 to 0.6). On the other hand, two Wearables (Apple Watch 8, and Oura Ring 3), one Nearable (Withings Sleep Tracking Mat), and an Airable SleepScore showed a fair level of agreement (kappa 0.2 and 0.4). Finally, Google Nest Hub 2 (Nearable) and Pillow (Airable) exhibited only a slight level of agreement across sleep stage classifications. The performance of CSTs was assessed in two distinct institutions, where the macro F1 scores, averaged over all devices in each institution, were 0.4973 and 0.4876 at SNUBH and CLC, respectively. It showed no significant difference in performance between these two locations. Among the 11 CSTs evaluated, five of them (Galaxy Watch 5, Apple Watch 8, Amazon Halo Rise, Pillow, and Sleep Routine) exhibited better performance at SNUBH, while the remaining CSTs demonstrated superior performance at CLC.

Table 2. Epoch-by-epoch agreement: classification of 4 sleep stages

	ALL				SNUBH	CLC
	Accuracy	**Weighted f1**	**Cohen`s Kappa**	**Macro F1**	**Macro F1**	**Macro F1**
Airable						
SleepRoutine (67)	**0.7106**	**0.7166**	**0.5565**	**0.6863**	**0.7188**	**0.6551**
SleepScore (38)	0.4329	0.4472	0.2065	0.4049	0.4408	0.3094
Pillow (74)	0.2830	0.2906	0.0741	0.2588	0.2604	0.2564
Nearable						
Withings Sleep Tracking Mat (75)	0.4921	0.5007	0.2455	0.4496	0.4205	0.4837
Google Nest Hub 2 (33)	0.4121	0.4089	0.0644	0.3009	0.2676	0.3299
Amazon Halo Rise (28)	0.6634	0.6706	0.4807	0.6242	0.6231	0.6031
Wearable						
Google Pixel Watch (30)	0.6355	0.6143	0.4044	0.5669	0.5381	0.5925
Galaxy Watch 5 (22)	0.6494	0.6499	0.4177	0.5761	0.6261	0.5651
Fitbit Sense 2 (26)	0.6464	0.6296	0.4185	0.5814	0.5130	0.6268
Apple Watch 8 (26)	0.5640	0.5731	0.2976	0.4910	0.5436	0.4203
Oura Ring 3 (53)	0.5427	0.5518	0.3492	0.5186	0.5187	0.5211

The number in the parenthesis indicates the number of participants tested with each device. Values for the top-performing CSTs are shown in bold. Abbreviations: *CST,* consumer sleep tracker; *SNUBH,* Seoul National Bundang Hospital; *CLC,* Clionic Life Center.

Epoch-by-Epoch Analysis: performance as per sleep stages

The performance of various sleep trackers across different sleep stages is presented in Table 3. For wake and REM stages, the SleepRoutine (Airable) achieved the highest macro F1 scores of 0.7065 and 0.7596, respectively. These scores substantially surpassed those of the second-best tracker, Amazon Halo Rise (Nearable), by a margin of 0.1098 for the wake stage and 0.0313 for the REM stage. For the deep stage, Google Pixel Watch and Fitbit Sense 2, which are Wearables, exhibited superior performance with macro F1 scores of 0.5933 and 0.5564, respectively. The Google Pixel Watch achieved the highest performance with a substantial margin. It surpassed the Fitbit Sense 2 by a margin of 0.0368 and outpaced SleepRoutine, which is the sleep tracker with the third highest score with an even larger margin of 0.0567. For light stage, an array of sleep trackers, including three Wearables (Google Pixel Watch, Galaxy Watch 5, Fitbit Sense 2), one Nearable (Amazon Halo Rise), and one Airable (SleepRoutine), demonstrated similarly high levels of performance, with a macro F1 score ranging from 0.7142 to 0.7436. Additional detailed assessments of sleep stage performance, including accuracy, weighted F1, and AUROC metrics, are presented in Supplementary Tables S3 to S6.

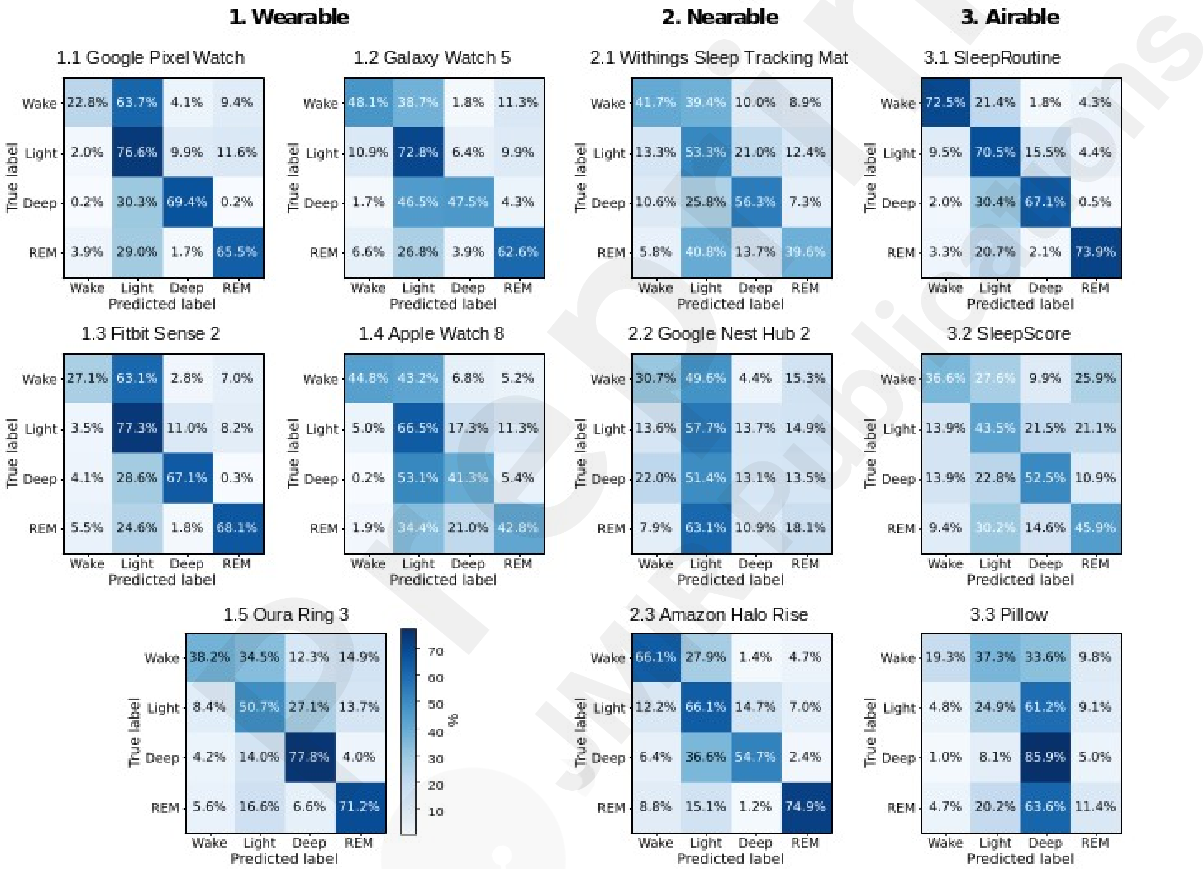
Table 3. Epoch-by-epoch agreement: two-class classifications for detecting individual sleep stage

	Wake			Light			Deep			REM		
	F1	Sen	Spe	F1	Sen	Spe	F1	Sen	Spe	F1	Sen	Spe
Airable												
SleepRoutine (67)	**0.7065**	**0.7246**	0.9269	**0.7436**	0.7054	**0.7665**	0.5355	0.6712	0.8973	**0.7596**	0.7394	**0.9609**
SleepScore (38)	0.4057	0.3665	0.8696	0.5147	0.4355	0.7272	0.3574	0.5247	0.8264	0.3418	0.4587	0.7895
Pillow (74)	0.2828	0.1934	0.9572	0.3409	0.2490	0.7534	0.2673	**0.8594**	0.4449	0.1440	0.1140	0.9126
Nearable												
Withings Sleep Tracking Mat (75)	0.4419	0.4172	0.8854	0.5764	0.5328	0.6336	0.3800	0.5633	0.8270	0.4001	0.3964	0.8906
Google Nest Hub 2 (33)	0.3296	0.3068	0.8649	0.5619	0.5772	0.4518	0.1245	0.1308	0.8883	0.1876	0.1805	0.8514
Amazon Halo Rise (28)	0.5967	0.6612	0.8921	0.7142	0.6609	0.7484	0.4575	0.5467	0.9018	0.7283	**0.7490**	0.9401
Wearable												
Google Pixel Watch (30)	0.3456	0.2277	**0.9784**	0.7150	0.7657	0.5620	**0.5922**	0.6937	0.9290	0.6146	0.6548	0.9029
Galaxy Watch 5 (22)	0.4755	0.4814	0.9104	0.7346	0.7280	0.6412	0.4963	0.4752	**0.9481**	0.5982	0.6265	0.9058
Fitbit Sense 2 (26)	0.3807	0.2714	0.9602	0.7262	**0.7734**	0.5727	0.5564	0.6710	0.9247	0.6623	0.6812	0.9297
Apple Watch 8 (26)	0.5493	0.4481	0.9624	0.6680	0.6649	0.5737	0.3073	0.4130	0.8412	0.4394	0.4276	0.9070
Oura Ring 3 (53)	0.4527	0.3822	0.9264	0.5953	0.5072	0.7630	0.4272	0.7784	0.7974	0.5993	0.7118	0.8716

The number in the parenthesis indicates the number of participants tested with each device. Individual sleep-stage classification was used to categorize each class and the remaining classes. Values for the top-performing CSTs are shown in bold. Abbreviations: *CST,* consumer sleep tracker; *REM,* rapid-eye movement sleep.

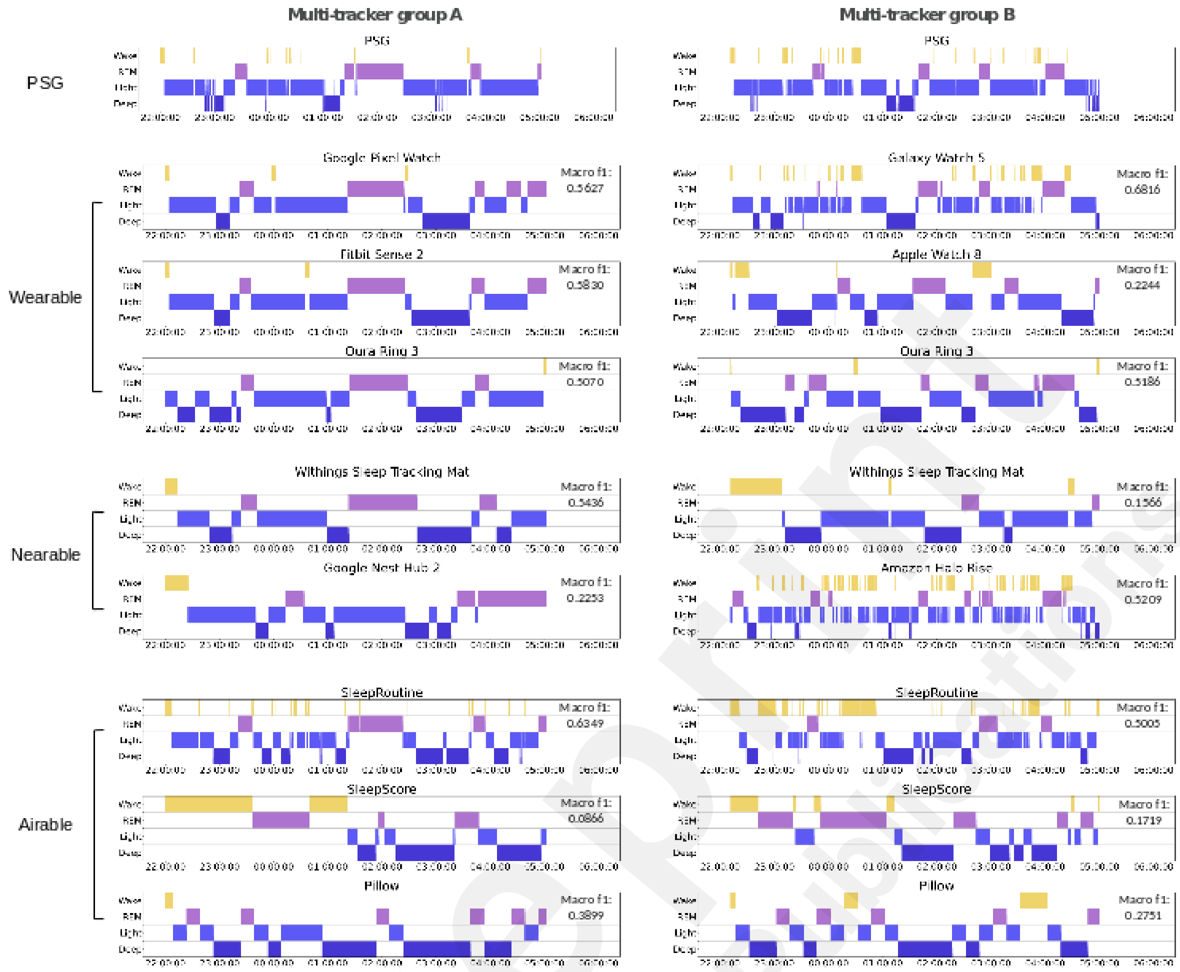
Figure 2 presents the confusion matrices for the sleep stages of the 11 CSTs, providing a clear visual representation of prediction biases and misclassification. Analysis of the average tendencies across all devices reveals that a prediction bias toward the light sleep stage. Google Nest Hub 2 (Nearable) showed the largest bias toward light stage among all the devices. Unlike other devices, Pillow (Airable) was highly biased toward deep stage predicting 59% of epoch as deep, whereas only 10.8% of epochs were deep based on the results of PSG. The confusion matrices also revealed distinct patterns of misclassification in sleep stage prediction for device types. Wearables primarily misclassified wake as light, while Nearables strongly misclassified REM as light. Airables, on the other hand, demonstrated a relatively higher frequency of confusion between light and deep stages. Figure 3 presents a comparison of hypnograms illustrating the epoch-by-epoch agreement at an individual level, which facilitated the evaluation of agreement between CSTs and PSG in a time-series format.

Figure 2. Normalized Confusion Matrices for 11 CSTs.



Four-stage sleep classification confusion matrices, comparing CSTs. Each row in the confusion matrix is the sleep stage annotated by PSG, while each column represents the sleep stage annotated by CST. Abbreviations: *CST,* consumer sleep tracker; *PSG,* polysomnography; *REM* rapid-eye movement sleep.

Figure 3. Sample hypnogram of 11 CSTs involving two subjects in different groups.

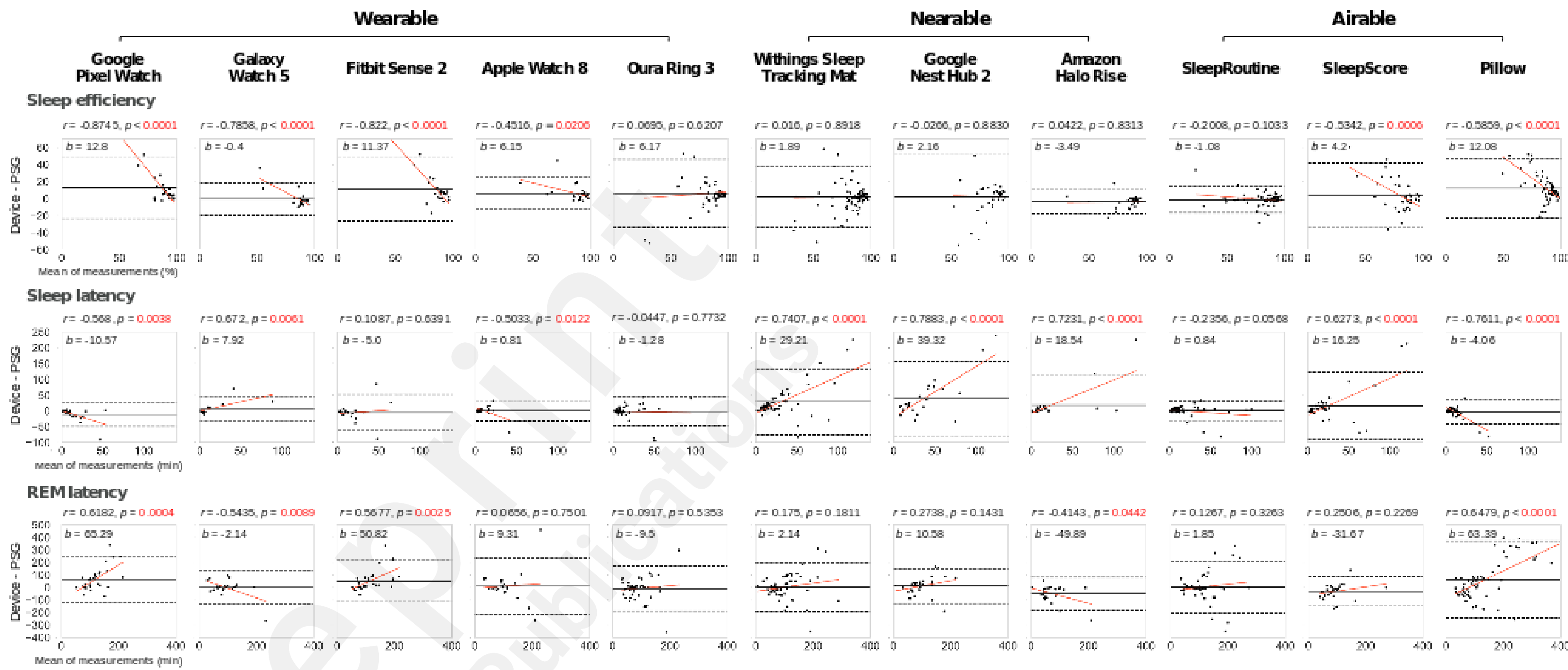


Hypnogram samples for each CST were selected based on the last measured subject in the multi-tracker group A (female, age: 35 years; BMI: 30.1, AHI: 2.9) and multi-tracker group B (female, age: 26 years, BMI: 20, AHI: 3.5 for group B). Abbreviations: *CST,* consumer sleep tracker; *PSG,* polysomnography; *BMI,* body mass index; *AHI,* apnea-hypopnea index.
The division of groups was necessary due to the limited number of watches worn simultaneously, as explained in the Methods. As nine devices were utilized simultaneously for each subject, the hypnograms for each device are presented, with the PSG result displayed at the top. As shown in this figure, SleepRoutine, Amazon Halo Rise, and Galaxy Watch 5 exhibited more frequent transition of stages and predicted wake in the middle of sleep more frequently, resulting in better estimation of WASO, as shown in the analysis of sleep parameters.

Sleep Measure Analysis

Figure 4 presents the Bland-Altman plots of CSTs, illustrating the performance of sleep measurements, including sleep efficiency, sleep latency, and REM latency, when compared with PSG. The average value of PSG sleep efficiency ranged from 77.57% to 86.05%, while the bias for each CST varied from -3.4909%p (Amazon Halo Rise) to 12.8035%p (Google Pixel Watch). PSG values for sleep latency ranged from 10.80 min to 19.80 min, with CST biases ranging from -0.81 min (Apple Watch 8) to 39.42 min (Google Nest Hub 2). PSG values for REM latency ranged from 87.00 min to 112.20 min, with CST biases ranging from -49.89 min (Amazon Halo Rise) to 65.29 min (Google Pixel Watch). Devices demonstrated distinct and best performance for each sleep metric. In terms of sleep efficiency, Galaxy Watch 5 (Wearable) achieved a minimal bias of -0.4%. In case of estimation of sleep latency, Apple Watch 8 (Wearable) exhibited a bias of 0.81 min. Lastly, SleepRoutine (Airable) demonstrated the best performance for REM latency with a bias of 1.85 min. The proportional bias, presented as ‘r’ in Figure 4, indicates how consistent the mean bias was regardless of the sleep measure. Oura ring and SleepRoutine showed no proportional bias (i.e., no significant correlation in the Bland-Altman plot) for any sleep measure.

Figure 4. Bland-Altman plots of CST and PSG of sleep efficiency, sleep latency, and REM latency measurements based on CSTs and PSG.



Bland-Altman plots present the mean bias (middle horizontal black solid line), upper (upper horizontal black dashed line), and lower (lower horizontal black dashed line) limits of agreement. In the figure, "b" represents bias, and "r" denotes Pearson's correlation coefficient between the mean of measurements and the difference between the CSTs and PSG. The correlation coefficient is displayed along with its corresponding p-value. The red lines indicate the estimated linear regression line. Abbreviations: PSG, polysomnography; REM, rapid-eye movement sleep; CST, consumer sleep tracker.

Subgroup Analysis

Subgroup analyses were conducted for all devices, considering factors including sex assigned at birth, AHI, sleep efficiency, and BMI. The Macro F1 scores for each subgroup are presented in Table 4 and S6. The average performance of overall CSTs showed a comprehensive relationship between sleep tracker performance and these parameters. In terms of BMI, the average macro F1 score was 0.5043 for individuals with BMI < = 25, whereas it dropped to 0.4790 for those with BMI > 25, indicating a gap of 0.0253. Similarly, for sleep efficiency, the scores were 0.4757 for individuals with sleep efficiency < = 85% and 0.4902 for those with sleep efficiency > 85%, with a difference of 0.0145. In the case of AHI, the scores were 0.4905 for AHI < = 15 and 0.5024 for AHI > 15, resulting in a difference of 0.0119. In contrast, the difference between male and female subgroups was minimal, with a macro F1 score of 0.4926 for males and 0.4932 for females, resulting in a negligible difference of 0.0006. In each subgroup, the highest variations were observed with the Airable SleepScore for AHI (difference: 0.0929), the Nearable Google Pixel Watch for sleep efficiency (difference: 0.1067), the Wearable Galaxy Watch 5 for BMI (difference: 0.0785), and again the Airable SleepScore for sex assigned at birth (difference: 0.0872). Supplementary Tables S9-S10 represent the subgroup analysis results of epoch-by-epoch agreement conducted within distinct institutions. Additionally, Tables S11-S12 provide an overview of the average macro F1 scores, individually calculated for each participant.

Table 4. Epoch-by-epoch agreement: Subgroup analysis of AHI and demographic characteristics

	Apnea-Hypopnea Index		Sleep Efficiency		Body Mass Index		Gender	
	≤ 15	> 15	≤ 85%	> 85%	≤ 25	> 25	Male	Female
Airable								
SleepRoutine (67)	**0.6536**	**0.7320**	**0.6971**	**0.6490**	**0.6840**	**0.6889**	**0.7137**	**0.6568**
SleepScore (38)	0.3636	0.4565	0.4107	0.3808	0.4118	0.3937	0.4431	0.3559
Pillow (74)	0.2602	0.2567	0.2472	0.2567	0.2601	0.2548	0.2670	0.2446
Nearable								
Withings Sleep Tracking Mat (75)	0.4644	0.4225	0.3766	0.4653	0.4760	0.3964	0.4587	0.4355
Google Nest Hub 2 (33)	0.3000	0.3059	0.3115	0.2762	0.3209	0.2517	0.2889	0.3059
Amazon Halo Rise (28)	0.6160	0.6389	0.6297	0.5857	0.6414	0.5801	0.6075	0.6491
Wearable								
Google Pixel Watch (30)	0.5670	0.5626	0.5035	0.6102	0.5653	0.5791	0.5235	0.5956
Galaxy Watch 5 (22)	0.5701	0.5790	0.6029	0.5547	0.5521	0.6306	0.5655	0.5867
Fitbit Sense 2 (26)	0.5839	0.5753	0.5325	0.6090	0.5910	0.5541	0.5320	0.6129
Apple Watch 8 (26)	0.4861	0.4950	0.4326	0.4804	0.5093	0.4561	0.5263	0.4414
Oura Ring 3 (53)	0.5302	0.5021	0.4882	0.5245	0.5354	0.4830	0.4926	0.5405

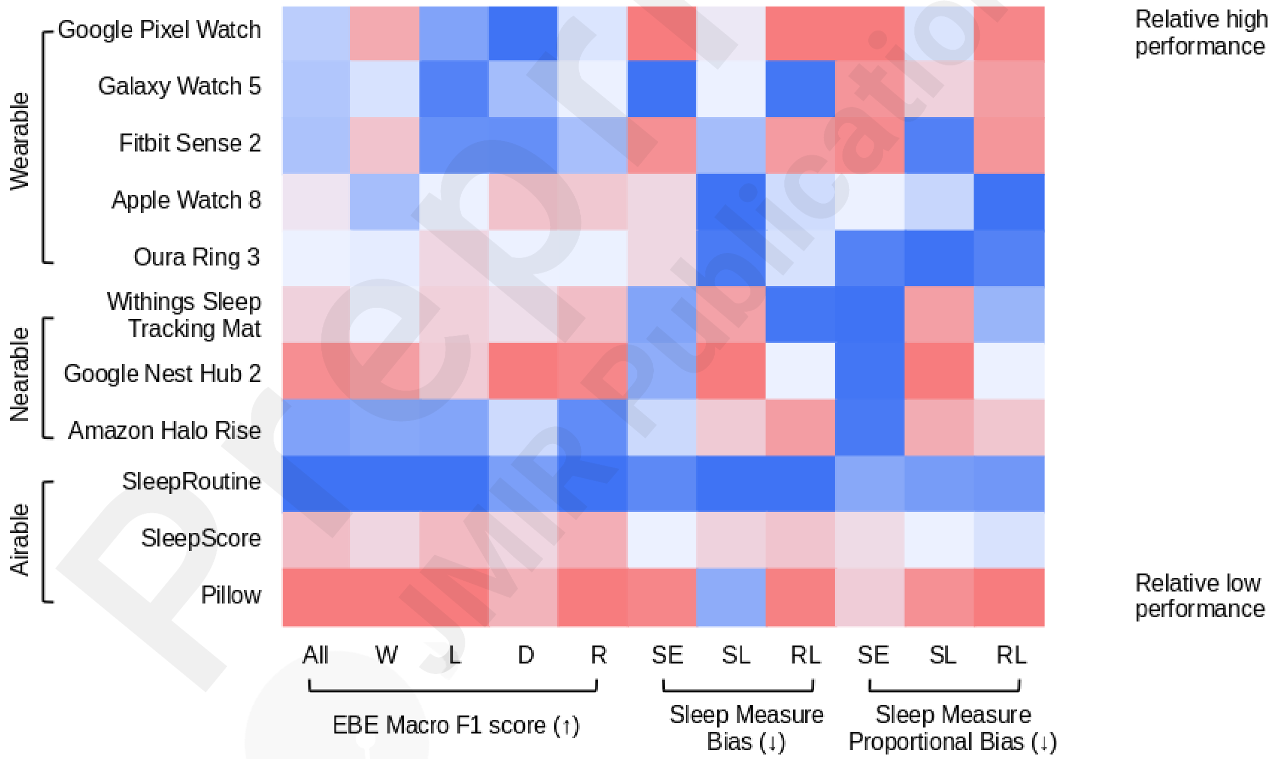
The number in the parenthesis indicates the number of participants tested with each device. Values for the top-performing consumer sleep trackers are shown in bold. Abbreviations: AHI, apnea-hypopnea index.

Discussion

Principal Results

We conducted an extensive analysis of 11 CSTs involving 75 subjects, which, to the best of our knowledge, represents the largest number of devices simultaneously evaluated in the literature.[3,4,10,11,13,18] The findings are illustrated in the heatmap (Figure 5), which presents a relative performance of the 11 CSTs in estimating sleep stages and sleep measures. Our findings revealed that Google Pixel Watch, Galaxy Watch 5, and Fitbit Sense 2 demonstrated competitive performance among Wearables, while Amazon Halo Rise and SleepRoutine stood out among Nearables and Airable, respectively.

Figure 5. Relative Performance Rank Heatmap of 11 CSTs.



Heatmap of the relative performance of sleep-stage and classification of sleep measures normalized to the highest and the lowest macro f1 value in each CST. Abbreviations: *CST*, consumer sleep tracker; *EBE*, epoch-by-epoch agreement

Wearables

Wearables, including watch and ring-type sleep trackers, represent the most prevalent CSTs in the market.[23][24] They employ PPG and accelerometer sensors to measure cardiac activity (e.g., heart rate variability) and body movements. Given their reliance on similar biosignals for sleep tracking, Wearables exhibit consistent patterns in estimating sleep stages. First, most Wearables generally overestimate sleep by misclassifying wake stages, leading to a substantial negative proportional bias in estimating sleep efficiency, which results in worse performance for individuals with low sleep efficiency (Figure 4). This bias was specifically observed when actigraphy was used to measure sleep efficiency and WASO.[5,11] This can be attributed to the dependence of actigraphy and Wearables on body movement to determine sleep-wake states. Insomniacs often lie still in bed while trying to sleep, even though they are actually awake.[24] As a result, these periods of wakefulness can be misinterpreted as sleep. Nevertheless, the Oura ring showed negligible proportional bias, potentially owing to its use of additional features, such as body temperature and circadian rhythm, for sleep staging.[6] Second, Wearables comprising top three CSTs demonstrated substantial alignment in the classification of deep stages (Supplementary Table S5). In particular, the results from Oura Ring 3 and Fitbit Sense 2 in this study showed improved accuracy in sleep stage detection compared to previous studies that focused on earlier versions of the Oura Ring and Fitbit in assessing the accuracy of wearable sleep evaluations. [25,26] Thus, Wearables may facilitate accurate detection of different stages of deep sleep, given their unique association with autonomic nervous system stabilization. Heart rate variability (HRV), a key indicator of autonomic nervous system activity can be directly measured by PPG.[27] Therefore, Wearables are effective in monitoring deep sleep stages.

Nearables

Nearables, encompassing pad and motion sensor-type sleep trackers, utilize overall body movements and respiratory efforts (thorax and abdominal) for sleep monitoring. Similar to Wearables, Nearables also exhibit aligned tendencies. First, all Nearables tend to overestimate sleep onset latency, resulting in significant mean bias (29.02 min for Nearable, -2.71 and 4.34 for Wearable and Airable) and a significant positive proportional bias in sleep latency measurement. This indicates that Nearables may probably overestimate sleep latency, particularly in individuals with prolonged sleep latency. During extended periods of attempting to fall asleep in bed, users may experience increased restlessness and movement, which makes it challenging for Nearables to estimate the sleep stage using radar-like sensors.[28,29] Second, unlike Wearables, Nearables demonstrated the least sensitivity in deep stage classification (as shown in Figure 2). Distinguishing stages of deep sleep from light sleep based on variations in respiratory patterns requires precise monitoring of respiratory activity. However, the radar-like sensors employed by Nearables, while efficient at detecting larger body movements, have difficulty capturing smaller, fidgeting movements, which are a challenge in accurately identifying the stages of deep sleep.

Airables

Airable excels in terms of its accessibility by not requiring purchase of additional hardware. However, its clinical validation is not well-established as noted in previous studies up until 2022 which highlighted the limited agreement between Airable and PSG as a notable limitation.[30,31] In this study, our aim was to validate the latest Airable CSTs. One distinguishing feature of Airable CSTs is their utilization of diverse sensor types, and the variation in performance is substantially influenced by the specific sensor type and accompanying algorithm. Thus, we chose three types of Airable CSTs considering diversity such that microphone, ultrasound, and accelerometer-based applications. These methodological distinctions contribute to pronounced variations in of the determination of sleep stage. Pillow requires placement on the mattress and uses the smartphone's accelerometer sensor to detect user movements through the mattress. Notably, Pillow showed a prediction bias towards the deep stage, suggesting that movement information during sleep was insufficient for accurate determination of sleep stages. SleepScore utilizes a sonar biomotion sensor, directing the smartphone's speaker towards the chest area to emit ultrasonic signals above 18 kHz, tracking thoracic respiratory effort. Depending on the biosignal used, SleepScore showed similar tendencies with Nearables, demonstrating a substantial mean bias and positive proportional bias in estimating sleep onset latency. SleepRoutine analyzes the sound recorded during sleep.[2] Sleep sounds provide a wealth of sleep-related information, including changes in breathing regularity linked to autonomic nervous system stabilization, changes in breathing sound characteristics (such as tone, pitch, and amplitude) due to altered respiratory muscle tension, and noise from body movements. Among all CSTs, SleepRoutine exhibited the highest accuracy in predicting wake and REM stages.

REM Sleep Stage Estimation Performance

REM is the stage where most CSTs demonstrated relatively higher agreement with the PSG stage compared with the other stages. Among the top 5 CSTs with the highest macro f1 scores (SleepRoutine, Amazon Halo Rise, Fitbit Sense 2, Galaxy Watch 5, and Google Pixel Watch), the REM stage showed a substantially higher average f1 score of 0.672, compared with 0.501 for wake and 0.528 for deep sleep. This can be attributed to the unique characteristics of REM sleep, which include increased irregularity in heart rate and breathing, minimal muscle movement, and rapid variations in blood pressure and body temperature.[32] These features allow easy detection of different types of biosignals and accurate classification of REM sleep.

Cost Effectiveness

We evaluated the costs of 11 sleep tracking technologies, based on the costs analyzed in Figure S1 of the Supplementary Materials. Wearables, averaging $386, offer a wide range of function, including messaging and various apps, beyond sleep tracking. The Oura ring, while lacking these supplementary features, provides a broad spectrum of health tracking functions. Nearables, with an average price of $123, include a variety of features across different models. The Google Nest Hub and Amazon Halo Rise offer extra features such as an IoT hub and wake-up light, whereas the Withings sleep mat is exclusively designed for sleep tracking. Airable, an app-based technology, harnesses smartphone sensors for sleep tracking, requiring only a $53 subscription fee and no additional hardware. This economical and flexible option, which can be easily canceled, represents a cost-effective solution for sleep tracking.

Standardized Validation and Data Transparency

Standardized methods of validation and data transparency are crucial for comparing sleep trackers,[33] particularly due to the increasing use of deep learning algorithms whose inner workings are often opaque. In our study, we adhered to established frameworks for standardized validation,[15,34] while also conducting multi-center evaluations based on diverse demographic factors. Regarding data transparency, we provided comprehensive details of validation; However, obtaining access to the training data for each CST was challenging. Transparency in both training and validation data is essential for building trustworthy AI models and can also contribute to a better understanding of CSTs. [35]

Limitations

It is important to note the limitations of our study. First, data collection rates significantly varied between the two institutions as the study was independently implemented. Issues such as battery management, account management, and human errors resulted in data omissions. Second, demographic differences were detected between the institutions, including disparities in time spent in bed and total sleep time. Operational issues led to slightly earlier waking of participants in the CLC. Third, this study focused solely on the Korean population, with limited ability for analysis of performance differences among various races. Future studies should incorporate multiracial comparisons and evaluate CST performance across diverse home environments for realistic assessments.

Conclusions

In conclusion, our study represents a comprehensive and comparative analysis of 11 CSTs and their accuracy in tracking sleep in a sleep lab setting. The objective of this study was to gain insight into the performance and capabilities of these CSTs. Personalized sleep health management is necessary to enable individuals to make informed choices for monitoring and improving sleep quality. Further, our findings emphasize the importance of understanding the characteristics and limitations of these devices. It lays the foundation for guiding the development of sleep trackers in the future. Accordingly, future studies should focus on developing accurate sleep stage classification systems by integrating different types of biosignals in a home environment.

Acknowledgements

SleepRoutine as one of the consumer sleep trackers (CSTs) included in this study, was developed and operated by Asleep, and the deep learning algorithm used in SleepRoutine was trained using data from SNUBH, to which some of the authors are affiliated. The entities played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript. The authors conducted the study independently.

Funding

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT)(No.RS-2022-00156561).

Conflicts of Interest

YC is the director of Clionic Lifecare Clinic.
Authors affiliated with Asleep and CAK hold stocks or stock options in Asleep.

Data Availability

Access to datasets from Seoul National University Bundang Hospital and Clionic Lifecare Clinic, used with permission for this study, should be requested directly from these institutions. Subject to the institutional review boards' ethical approval, the corresponding author agrees to share de-identified individual participant data, the study protocol, and the statistical analysis plan with academic researchers following completion of a data use agreement. Proposals should be directed to corresponding authors. All experimental and implementation details that can be shared are described in detail in this Article and its appendix.

Authors' Contributions

TL and YC contributed to conceptualization, data curation, formal analysis, investigation, methodology, writing the original draft, and review & editing. KSC and JJ performed software coding, validation, visualization, and reviewed the draft. JJ, JC, HK, DK, JH, and DL contributed to conceptualization, software coding, and manuscript revision. MK conducted data curation and validation. CAK supervised the project, contributed to conceptualization, and provided guidance. I-YY and J-WK are the principal investigators and contributed to conceptualization, design, interpretation, writing, and project administration. All authors have read and approved the final version of the report.

Abbreviations

AHI: apnea-hypopnea index
BMI: body mass index
CSTs: Consumer sleep trackers
CLC: Clionic Lifecare Clinic
IRB: Institutional Review Boards
PSG: Polysomnography
AHI: apnea-hypopnea index
BMI: body mass index
SNUBH: Seoul National University Bundang
CLC: Clionic Lifecare Clinic

TIB: time in bed
TST: total sleep time
WASO: wake after sleep onset
IRB: Institutional Review Boards

References

1. Watson NF, Badr MS, Belenky G, Bliwise DL, Buxton OM, Buysse D, Dinges DF, Gangwisch J, Grandner MA, Kushida C, Malhotra RK, Martin JL, Patel SR, Quan S, Tasali E. Recommended Amount of Sleep for a Healthy Adult: A Joint Consensus Statement of the American Academy of Sleep Medicine and Sleep Research Society. Sleep 2015 Jun 1; doi: 10.5665/sleep.4716
2. Tran HH, Hong JK, Jang H, Jung J, Kim J, Hong J, Lee M, Kim J-W, Kushida CA, Lee D, Kim D, Yoon I-Y. Prediction of Sleep Stages Via Deep Learning Using Smartphone Audio Recordings in Home Environments: Model Development and Validation. J Med Internet Res 2023 Jun 1;25:e46216. doi: 10.2196/46216
3. Fino E, Plazzi G, Filardi M, Marzocchi M, Pizza F, Vandi S, Mazzetti M. Smart sleep tracking through the phone: Findings from a polysomnography study testing the reliability of four sleep applications. J Sleep Res 2020 Feb 31;29(1). doi: 10.1111/jsr.12935
4. Chinoy ED, Cuellar JA, Jameson JT, Markwald RR. Performance of Four Commercial Wearable Sleep-Tracking Devices Tested Under Unrestricted Conditions at Home in Healthy Young Adults. Nat Sci Sleep 2022 Mar;Volume 14:493–516. doi: 10.2147/NSS.S348795
5. Marino M, Li Y, Rueschman MN, Winkelman JW, Ellenbogen JM, Solet JM, Dulin H, Berkman LF, Buxton OM. Measuring Sleep: Accuracy, Sensitivity, and Specificity of Wrist Actigraphy Compared to Polysomnography. Sleep 2013 Nov 1;36(11):1747–1755. doi: 10.5665/sleep.3142
6. Altini M, Kinnunen H. The Promise of Sleep: A Multi-Sensor Approach for Accurate Sleep Stage Detection Using the Oura Ring. Sensors 2021, Vol 21, Page 4302 Multidisciplinary Digital Publishing Institute; 2021 Jun 23;21(13):4302. PMID:34201861
7. Hong JK, Lee T, Delos Reyes RD, Hong J, Tran HH, Lee D, Jung J, Yoon I-Y. Confidence-Based Framework Using Deep Learning for Automated Sleep Stage Scoring. Nat Sci Sleep 2021 Dec;Volume 13:2239–2250. doi: 10.2147/NSS.S333566
8. Fino E, Mazzetti M. Monitoring healthy and disturbed sleep through smartphone applications: a review of experimental evidence. Sleep and Breathing 2019 Mar 23;23(1):13–24. doi: 10.1007/s11325-018-1661-3
9. Tuominen J, Peltola K, Saaresranta T, Valli K. Sleep Parameter Assessment Accuracy of a Consumer Home Sleep Monitoring Ballistocardiograph Beddit Sleep Tracker: A Validation Study. Journal of Clinical Sleep Medicine 2019;15(3):483–487. doi: 10.5664/jcsm.7682
10. Mantua J, Gravel N, Spencer RMC, Leonhardt S, Teichmann D. Reliability of Sleep Measures from Four Personal Health Monitoring Devices Compared to Research-Based Actigraphy and Polysomnography. doi: 10.3390/s16050646
11. Chinoy ED, Cuellar JA, Huwa KE, Jameson JT, Watson CH, Bessman SC, Hirsch DA,

Cooper AD, Drummond SPA, Markwald RR. Performance of seven consumer sleep-tracking devices compared with polysomnography. Sleep Oxford Academic; 2021 May 14;44(5):1–16. PMID:33378539

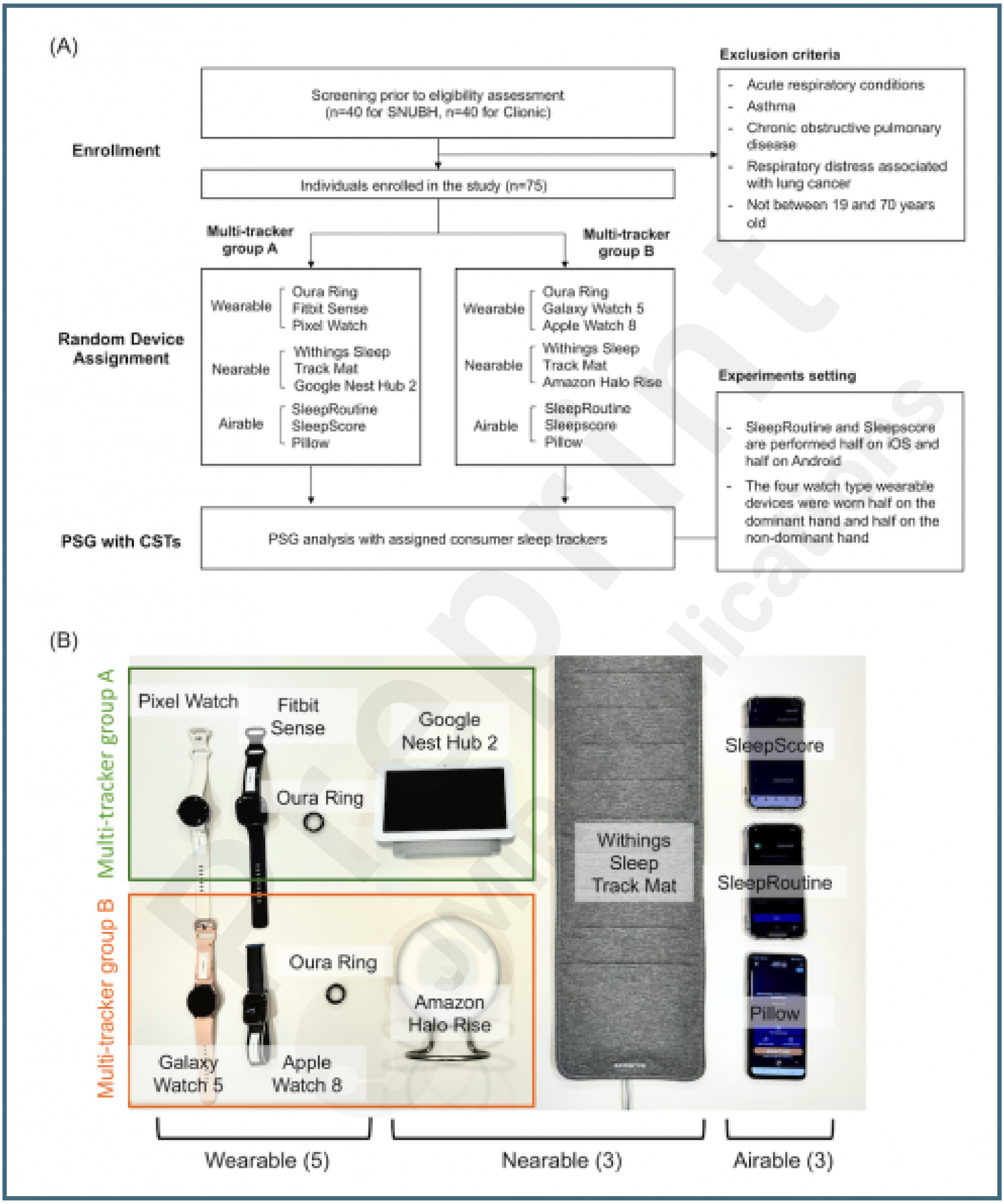
12. Klier K, Wagner M. Agreement of Sleep Measures—A Comparison between a Sleep Diary and Three Consumer Wearable Devices. Sensors 2022, Vol 22, Page 6189 Multidisciplinary Digital Publishing Institute; 2022 Aug 18;22(16):6189. PMID:36015949
13. Haghayegh S, Khoshnevis S, Smolensky MH, Diller KR, Castriotta RJ. Accuracy of Wristband Fitbit Models in Assessing Sleep: Systematic Review and Meta-Analysis. J Med Internet Res JMIR Publications Inc.; 2019 Nov 1;21(11). PMID:31778122
14. Shcherbina A, Mattsson C, Waggott D, Salisbury H, Christle J, Hastie T, Wheeler M, Ashley E. Accuracy in Wrist-Worn, Sensor-Based Measurements of Heart Rate and Energy Expenditure in a Diverse Cohort. J Pers Med 2017 May 24;7(2):3. doi: 10.3390/jpm7020003
15. Nguyen QNT, Le T, Huynh QBT, Setty A, Vo T V, Le TQ. Validation Framework for Sleep Stage Scoring in Wearable Sleep Trackers and Monitors with Polysomnography Ground Truth. 2021; doi: 10.3390/clockssleep3020017
16. 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 NIH Public Access; 2021;38(7):1010. PMID:33792456
17. Schade MM, Bauer CE, Murray BR, Gahan L, Doheny EP, Kilroy H, Zaffaroni A, Montgomery-Downs HE. Sleep validity of a non-contact bedside movement and respiration-sensing device. Journal of Clinical Sleep Medicine American Academy of Sleep Medicine; 2019;15(7):1051–1061. PMID:31383243
18. Kolla BP, Mansukhani S, Mansukhani MP. Consumer sleep tracking devices: a review of mechanisms, validity and utility. https://doi.org/101586/1743444020161171708 Taylor & Francis; 2016 May 3;13(5):497–506. PMID:27043070
19. Robbins R, Seixas A, Walton Masters L, Chanko N, Diaby F, Vieira D, Jean-Louis G. Sleep tracking: A systematic review of the research using commercially available technology. Curr Sleep Med Rep NIH Public Access; 2019 Sep 15;5(3):156. PMID:33134038
20. Mouritzenid NJ, Larsen LH, Lauritzen MH, Kjaer TW. Assessing the performance of a commercial multisensory sleep tracker. 2020; doi: 10.1371/journal.pone.0243214
21. Ananth S. Sleep apps: current limitations and challenges. Sleep Sci 2021;14(1):83–86. PMID:34104344
22. Berry RB, Brooks R, Gamaldo CE, Harding SM, Marcus C, Vaughn B V, others. The AASM manual for the scoring of sleep and associated events. Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine 2012;176:2012.
23. Global Market Insights. Sleep Tech Devices Market - By Product (Wearables, Non-wearables), By Application (Insomnia, Obstructive Sleep Apnea, Narcolepsy) and By Distribution Channel (Sleep centers and fitness centers, Pharmacy and retail stores) & Forecast, 2023 -2032. 2023.
24. Schutte-Rodin S, Broch L, Buysse D, Dorsey C, Sateia M. Clinical guideline for the evaluation and management of chronic insomnia in adults. J Clin Sleep Med 2008 Oct 15;4(5):487–504. PMID:18853708

25. De Zambotti M, Cellini N, Goldstone A, Colrain IM, Baker FC. Wearable Sleep Technology in Clinical and Research Settings. Med Sci Sports Exerc NIH Public Access; 2019 Jul 1;51(7):1538. PMID:30789439
26. de Zambotti M, Goldstone A, Claudatos S, Colrain IM, Baker FC. A validation study of Fitbit Charge 2™ compared with polysomnography in adults. https://doi.org/101080/0742052820171413578 Taylor & Francis; 2017 Apr 3;35(4):465–476. PMID:29235907
27. Vanoli E, Adamson PB, Ba-Lin, Pinna GD, Lazzara R, Orr WC. Heart Rate Variability During Specific Sleep Stages. Circulation 1995 Apr;91(7):1918–1922. doi: 10.1161/01.CIR.91.7.1918
28. Halo Rise Placement Tip.
29. Quick Installation Guide - Withings Sleep Analyzer. Withings; 2020.
30. Ibáñez V, Silva J, Navarro E, Cauli O. Sleep assessment devices: types, market analysis, and a critical view on accuracy and validation. https://doi.org/101080/1743444020191693890 Taylor & Francis; 2019 Dec 2;16(12):1041–1052. PMID:31774330
31. De Zambotti M, Cellini N, Goldstone A, Colrain IM, Baker FC. Wearable Sleep Technology in Clinical and Research Settings. Med Sci Sports Exerc NIH Public Access; 2019 Jul 1;51(7):1538. PMID:30789439
32. Shrivastava D, Jung S, Saadat M, Sirohi R, Crewson K. How to interpret the results of a sleep study. J Community Hosp Intern Med Perspect 2014;4(5):24983. PMID:25432643
33. Lehmann CU, Miller MR. Standardization and the Practice of Medicine. Journal of Perinatology 2004 Mar 1;24(3):135–136. doi: 10.1038/sj.jp.7211060
34. 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 Oxford University Press; 2021 Feb 1;44(2). PMID:32882005
35. Good Machine Learning Practice for Medical Device Development: Guiding Principles. 2021 Oct.

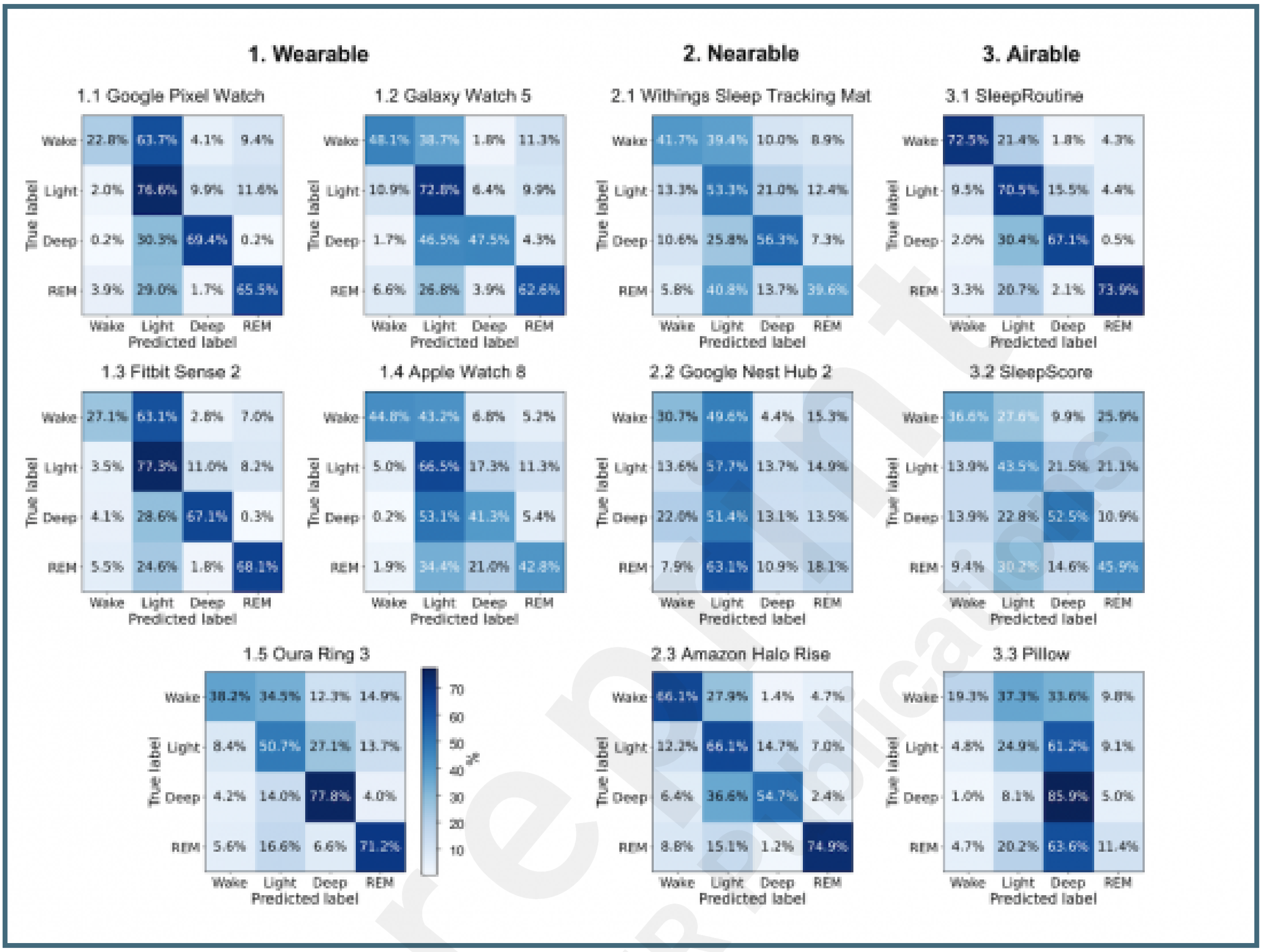
Supplementary Files

Figures

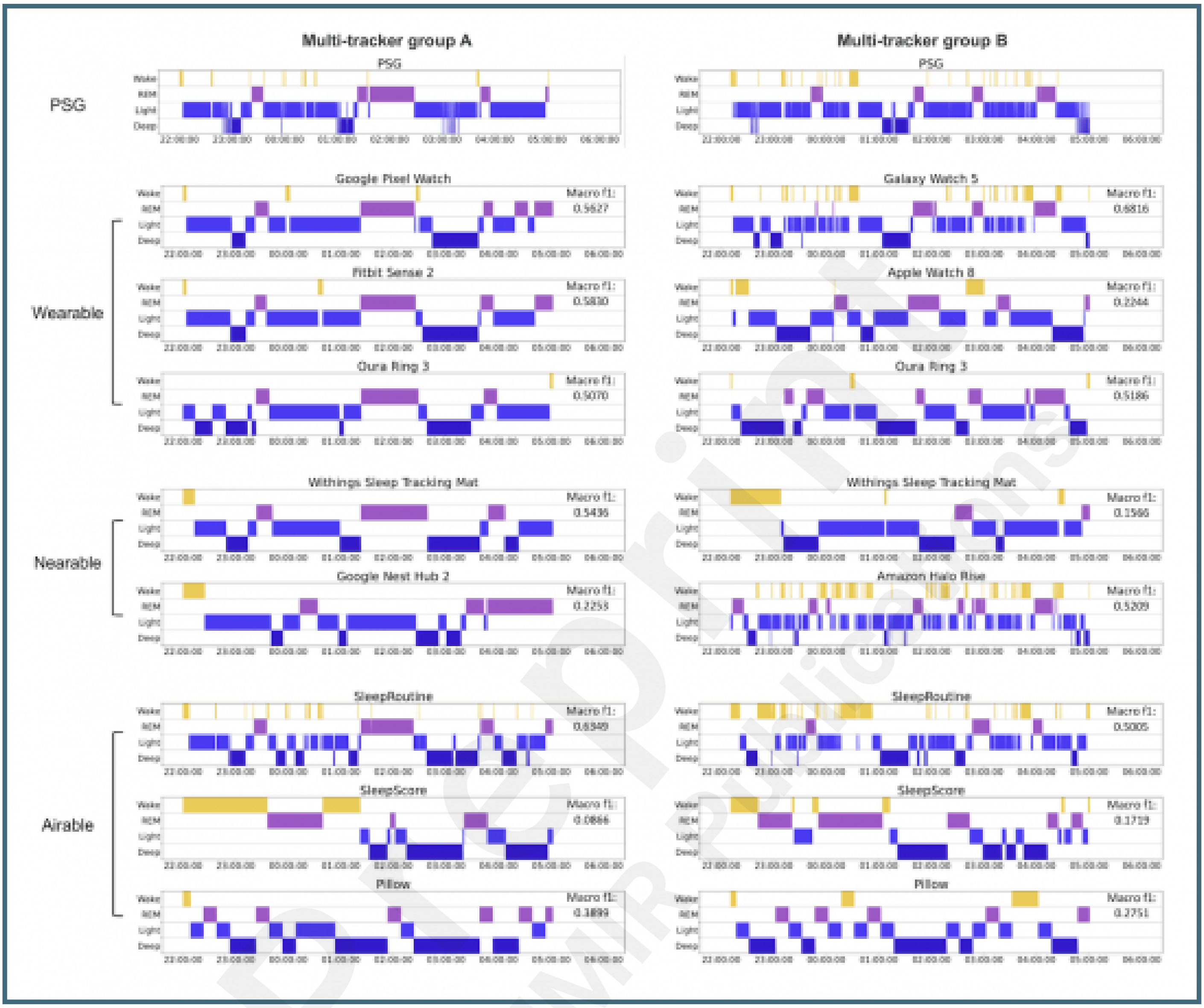
Flowchart outlining the experimental design.



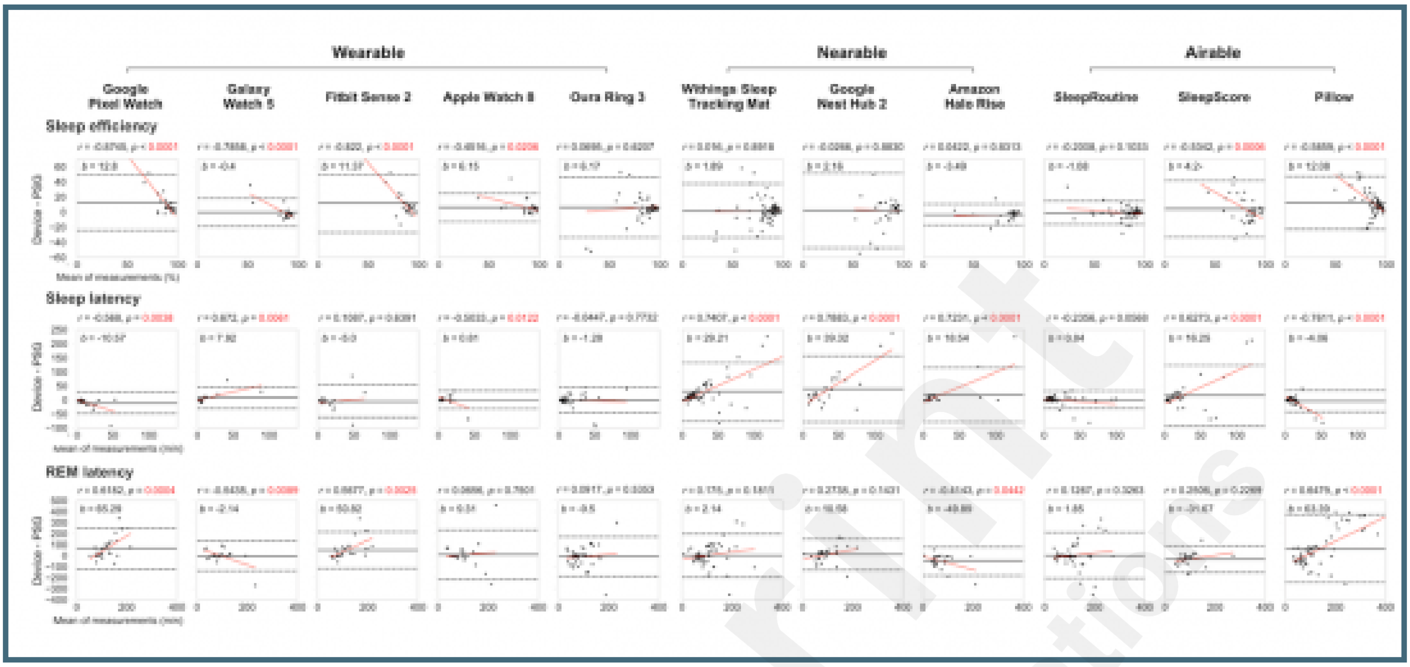
Normalized Confusion Matrices for 11 CSTs.



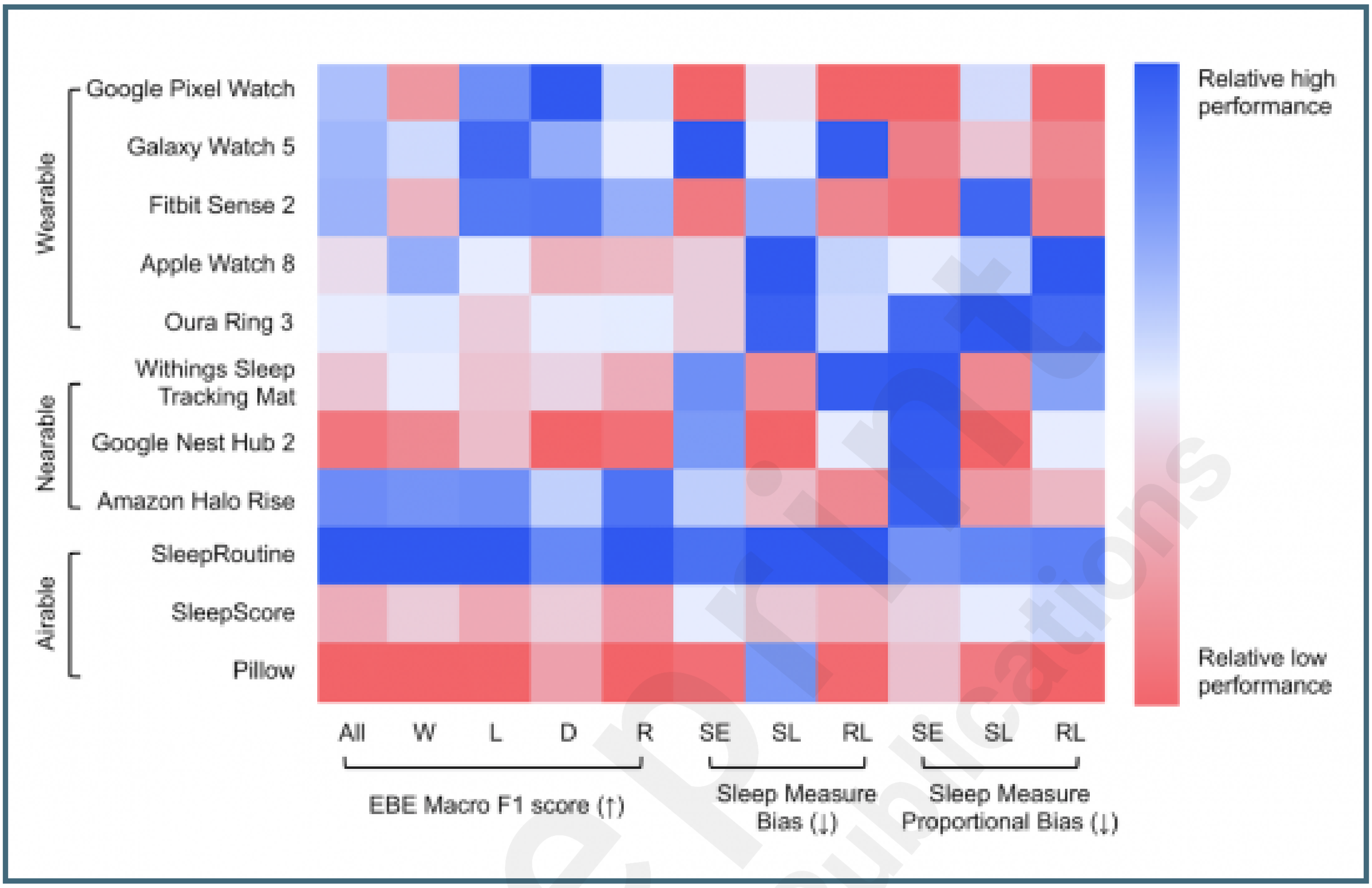
Sample hypnogram of 11 CSTs involving two subjects in different groups.



Bland-Altman plots of CST and PSG of sleep efficiency, sleep latency, and REM latency measurements based on CSTs and PSG.



Bland-Altman plots of CST and PSG of sleep efficiency, sleep latency, and REM latency measurements based on CSTs and PSG.



Multimedia Appendixes

Supplementary Material.
URL: http://asset.jmir.pub/assets/30f4cf8ef0b867851387c622aa09a336.docx

Normalized Confusion Matrices: Mean and Variance of Predicted Values across Participants.
URL: http://asset.jmir.pub/assets/ad9db596e291f70ef0300c343e457619.pdf

Additional Hypnogram example for 11 CSTs of two subjects in different groups.
URL: http://asset.jmir.pub/assets/febdefa4129034546fd9b5e785ba7d62.pdf

Cost-effectiveness of consumer sleep trackers.
URL: http://asset.jmir.pub/assets/01a6b1878d9e0624e3339a2328df70b4.pdf