SLEEPJ, 2020, 1–13

doi: 10.1093/sleep/zsaa097
Advance Access Publication Date: 20 May 2020
Original Article

ORIGINAL ARTICLE

The Dreem Headband compared to polysomnography for electroencephalographic signal acquisition and sleep staging

Pierrick J. Arnal[1,*], Valentin Thorey[2], Eden Debellemaniere[1], Michael E. Ballard[1], Albert Bou Hernandez[2], Antoine Guillot[2], Hugo Jourde[2], Mason Harris[2], Mathias Guillard[3,4], Pascal Van Beers[3,4], Mounir Chennaoui[3,4] and Fabien Sauvet[3,4]

[1]Dreem, Science Team, New York, NY, [2]Dreem, Algorithm Team, Paris, France, [3]French Armed Forces Biomedical Research Institute (IRBA), Fatigue and Vigilance Unit, Bretigny-sur-Orge, France, [4]EA 7330 VIFASOM, Paris Descartes University, Paris, France

*Corresponding author. Pierrick J. Arnal, Dreem, Science Team, 450 Park Ave S, New York, NY 10016. Email: research@dreem.com.

Abstract

Study Objectives: The development of ambulatory technologies capable of monitoring brain activity during sleep longitudinally is critical for advancing sleep science. The aim of this study was to assess the signal acquisition and the performance of the automatic sleep staging algorithms of a reduced-montage dry-electroencephalographic (EEG) device (Dreem headband, DH) compared to the gold-standard polysomnography (PSG) scored by five sleep experts.

Methods: A total of 25 subjects who completed an overnight sleep study at a sleep center while wearing both a PSG and the DH simultaneously have been included in the analysis. We assessed (1) similarity of measured EEG brain waves between the DH and the PSG; (2) the heart rate, breathing frequency, and respiration rate variability (RRV) agreement between the DH and the PSG; and (3) the performance of the DH's automatic sleep staging according to American Academy of Sleep Medicine guidelines versus PSG sleep experts manual scoring.

Results: The mean percentage error between the EEG signals acquired by the DH and those from the PSG for the monitoring of α was 15 ± 3.5%, 16 ± 4.3% for β, 16 ± 6.1% for λ, and 10 ± 1.4% for θ frequencies during sleep. The mean absolute error for heart rate, breathing frequency, and RRV was 1.2 ± 0.5 bpm, 0.3 ± 0.2 cpm, and 3.2 ± 0.6%, respectively. Automatic sleep staging reached an overall accuracy of 83.5 ± 6.4% (F1 score: 83.8 ± 6.3) for the DH to be compared with an average of 86.4 ± 8.0% (F1 score: 86.3 ± 7.4) for the 5 sleep experts.

Conclusions: These results demonstrate the capacity of the DH to both monitor sleep-related physiological signals and process them accurately into sleep stages. This device paves the way for, large-scale, longitudinal sleep studies.

Clinical Trial Registration: NCT03725943.

Statement of Significance

The development of ambulatory technologies able to monitor physiological signals during sleep at home and longitudinally is rising. These technologies advancements have the potential to move forward the field of sleep medicine, but the poor validation of current wearable technology inhibits widespread use. This validation study of a reduced-montage dry-electroencephalographic (EEG) device showed that this device is able to acquire EEG, heart rate, and breathing frequency and automatically analyze these signals using machine learning approach to provide sleep stages with an accuracy close to the consensus of five sleep scorers.

Key words: sleep; EEG; machine learning; sleep stages; device

Submitted: 4 December, 2019; **Revised:** 23 April, 2020

Introduction

Sleep disorders and insufficient sleep negatively impact hundreds of millions of people across the world and constitute a growing public health epidemic with grave consequences, including increased risk of cardiovascular and neurodegenerative diseases and psychiatric disorders [1]. The most prevalent sleep disorders include insomnia, which affects ~20% of the general population, and obstructive sleep apnea, which affects ~10% of the general population [2]. Despite their high prevalence, sleep disorders remain largely unidentified and/or untreated with less than 20% of patients estimated to be accurately diagnosed and treated [3].

Today, the gold standard to study or diagnose sleep disorders is nocturnal polysomnography (PSG). A PSG sleep study is typically a single overnight assessment, usually taking place in a sleep center, during which physiological signals including electroencephalographic (EEG), electromyographic (EMG), and electrooculographic (EOG) activity, breathing effort, airflow, pulse, and blood oxygen saturation are recorded. Analysis of these signals relies on trained sleep experts to visually inspect and manually annotate and recognize specific EEG, EOG, EMG patterns on 30-s segments (epochs) of the full PSG recording to score sleep stages (Wake, sleep stages 1 [N1], 2 [N2], and 3 [N3], and REM sleep), according to the American Academy of Sleep Medicine's (AASM) guidelines [4].

However, the gold-standard PSG suffers from several limitations. From a practical standpoint, a PSG is complicated and time-consuming to set-up, requiring up to 1 h to install by a trained sleep technician; it is also quite expensive (typically $1,500–$2,000 per night in the United States). Furthermore, a clinical PSG may not reliably capture a patient's typical sleep because it is cumbersome, and the clinical setting often generates stress for the patient. Moreover, because a PSG is generally performed over only one night, it does not capture intra-individual variability across nights and the final diagnosis is often rendered on an unrepresentative night of sleep [5, 6]. From a clinical standpoint, performing a PSG exam requires extensive training, it is time-consuming, and the sleep staging suffers from low inter-rater reliability. For instance, one study conducted on the AASM ISR data set found that sleep stage agreement across experts averaged 82.6% using data from more than 2,500 scorers, most with 3 or more years of experience, who scored 9 record fragments, representing 1,800 epochs (i.e. more than 3,200,000 scoring decisions). The agreement was highest for the REM sleep stage (90.5%) and slightly lower for N2 and Wake (85.2% and 84.1%, respectively), while the agreement was far lower for stages N3 and N1 (67.4% and 63.0%, respectively), placing constraints on the reliability of manual scoring [7]. Critically, studies also indicate that agreement varies substantially across different sleep pathologies and sleep centers [7, 8].

Automatic PSG analysis in sleep medicine has been explored and debated for some time but has yet to be widely adopted in clinical practice. In recent years, dozens of algorithms have been published that achieve expert-level performance for automated analysis of PSG data [9–12]. Indeed, scientists and engineers have used artificial intelligence methods to develop automated sleep stage classifiers and EEG pattern detectors, thanks to open access sleep data sets such as the National Sleep Research Resource (https://sleepdata.org). Regarding sleep staging, Biswal *et al.* proposed the SLEEPNET algorithm [13], a deep recurrent neural network trained on 10,000 PSG recordings from the Massachusetts General Hospital Sleep Laboratory. The algorithm achieved an overall accuracy comparable to human-level performance of 85.76% (N1: 56%, N2: 88%, N3: 85%, REM: 92%, and Wake: 85%). Another important collaborative study recently published an algorithm validated on ~3,000 normal and abnormal sleep recordings [12]. They showed that their best model using a deep neural network performed better than any individual scorer (overall accuracy: 87% compared to the consensus of 6 scorers). The problem of low interscorer reliability of sleep stages is addressed by using a consensus of multiple trained sleep scorers instead of relying on a single expert's interpretation [8, 12, 14]. Regarding the topic of sleep EEG event detection, deep learning methods have shown state-of-the-art performance for automatic detection of sleep events such as spindles and k-complexes in PSG records [15].

With the rise of wearable technology over the last decade, consumer sleep trackers have seen exponential growth [16]. For many years, these devices used only movement analysis, called actigraphy, before incorporating measures of pulse oximetry. Actigraphy has been extensively used in sleep research for sleep–wake cycle assessment at home. However, this measure has very low specificity for differentiating sleep from motionless wakefulness, resulting in an overestimation of total sleep time (TST) and underestimation of wake after sleep onset (WASO) time [17, 18]. Thus, actigraphy is still quite far from being a reliable alternative to PSG for sleep staging. And though the addition of pulse oximetry improves analysis over actigraphy alone, it still only enables rough estimations of sleep efficiency (SE) and stages. This is because the essential component of monitoring brain electrical activity with EEG sensors was still lacking.

More recently, a new group of devices has emerged for home sleep monitoring that uses EEG electrodes to measure brain activity. These include headbands [19–22] and devices placed around the ear [23, 24]. These compact devices are usually cheap, less burdensome, designed to be worn for multiple nights at home to enable longitudinal data collection, and require minimal or no expert supervision. However, as a recent International Biomarkers Workshop on Wearables in Sleep and Circadian Science reported: "Given the state of the current science and technology, the limited validation of wearable devices against gold standard measurements is the primary factor limiting large-scale use of wearable technologies for sleep and circadian research" [25]. Indeed, only a few of these device makers have published their performance compared to PSG; and those that have often only report aggregated metrics rather than raw data, and do not permit open access to the data set so that results can be independently verified.

In this study, we introduce the Dreem headband (DH) which is intended as an affordable, comfortable, and patient-friendly EEG-reduced montage with a high level of accuracy regarding both physiological signal acquisition and automatic sleep stage analysis using a deep learning algorithm along with five dry-EEG electrodes (O1, O2, FpZ, F7, and F8), a 3D accelerometer, and a pulse oximeter embedded in the device. To this end, we recorded data from 31 subjects over a single night using the DH and a clinical PSG simultaneously. We assessed: (1) the ability of the DH to monitor brain sleep frequencies during the night; (2) the accuracy of heart rate, breathing frequency, and respiration rate variability (RRV) during sleep; and (3) the performance of the automatic sleep stage classification algorithm of

the DH compared to a consensus of five sleep experts' manual scoring of the PSG. The data set of the current study is available in open access here: https://dreem-octave-irba.s3.eu-west-3.amazonaws.com/index.html

Methods

Subjects

A total of 31 volunteers were recruited without regard to gender or ethnicity from the local community by study advertisement flyers. Volunteers were eligible if they were between the ages of 18 and 65 years and capable of providing informed consent. Exclusion criteria included current pregnancy or nursing; severe cardiac, neurological, or psychiatric comorbidity in the last 12 months; morbid obesity (BMI ≥ 40); or use of benzodiazepines, non-benzodiazepines (Z-drugs), or γ-hydroxybutyrate on the day of the study.

Each participant provided one night of data; with the exception of two participants who completed a second night each due to data loss related to PSG battery issues on their first nights of study making a total of 33 nights. Finally, eight nights of data were excluded from the final analysis data set: five due to poor signal or system malfunction including battery issues on PSG, two due to the discovery of asymptomatic Apnea–Hypopnea Index (AHI) > 5 during the course of the study, and one due to an unusually short TST (4.5 h).

The final analysis data set consisted of one-night record from each of 25 participants; demographics are summarized in Table 1. The sample included individuals with self-reported sleep quality ranging from no complaints to sub-threshold insomnia symptoms, according to the Insomnia Severity Index (ISI) [26] and moderate to severe daytime sleepiness, according to the Epworth Sleepiness Scale (ESS) [27]. Only one met the Insomnia Symptom Questionnaire diagnostic threshold of insomnia. All had at worst mild symptoms of anxiety, according to the Generalized Anxiety Disorder-7 scale [28] or depression according to the Patient Health Questionnaire-9 (PHQ-9) [29]. Most reported moderate consumption of alcohol (less than a drink per day) and caffeine (less than 2 coffee per day), moderate frequency of exercise (1–3 sessions per week), and only occasional naps. Six were current nicotine users, 14 reported using nicotine less than 100 times total, and 1 was a former nicotine user.

Table 1. Demographics of the sample

	Mean ± SD	Min–Max
#Female/male	6/19	
Age	35.32 ± 7.51	23–50
BMI (kg m^{-2})	23.81 ± 3.43	17.44–31.6
ISI	5.00 ± 3.67	0–14
ESS	7.76 ± 3.77	1–19
PHQ-9	1.84 ± 1.95	0–6
GAD-7	2.00 ± 2.57	0–10
N Naps/week	0.79 ± 1.13	0–4
N Exercise/week	1.77 ± 1.72	0–6

BMI, body mass index; GAD, general anxiety disorder; ESS, Epworth Sleepiness Scale; PHQ-9, Patient HealthQuestionnaire-9; ISI, Insomnia Severity Questionnaire.

Protocol

Potential participants first completed a brief phone screen with study staff followed by an in-person interview at the French Armed Forces Biomedical Research Institute's (IRBA) Fatigue and Vigilance Unit (Bretigny-Sur-Orge, France) during which they provided informed consent and subsequently completed a detailed demographic, medical, health, sleep, and lifestyle survey with a study physician to confirm eligibility. Once consented and eligibility was confirmed, participants were equipped by a sleep technologist to undergo an overnight sleep study at the center with simultaneous PSG and the DH recordings. The beginning and the end of the PSG and DH data collection periods were set based on participants' self-selected lights-off and lights-on times. PSG and DH data recordings were synchronized a posteriori by resampling the DH data on the same timestamps as the PSG data so that records were perfectly aligned. Following the sleep study, technologists removed both devices, participants were debriefed and interviewed to identify any adverse events, and any technical problems were noted. All participants received financial compensation commensurate with the burden of study involvement. The study was approved by the Committees of Protection of Persons (CPP), declared to the French National Agency for Medicines and Health Products Safety, and carried out in compliance with the French Data Protection Act and International Conference on Harmonization (ICH) standards and the principles of the Declaration of Helsinki of 1964 as revised in 2013.

Polysomnographic assessment

The PSG assessment was performed using a Siesta 802 (Compumedics Limited, Victoria, Australia) with the following EEG derivations: F3/M2, F4/M1, C3/M2, C4/M1, O1/M2, O2/M1; 256 Hz sampling rate with a 0.03–35 Hz bandpass filter; bilateral EOG, electrocardiographic (EKG), submental and bilateral leg electromyographic recordings were also performed. Airflow, thoracic movements, snoring, and oxygen saturation were also monitored. EEG cup-electrodes of silver-silver chloride (Ag–AgCl) were attached to participants' scalps with EC2 electrode cream (Grass Technologies, Astro-Med, Inc., West Warwick, RI, USA), according to the international 10–20 system for electrode placement. Auto-adhesive electrodes (Neuroline 720, Ambu A/S, Ballerup, Denmark) were used for EOG and EKG recordings.

Study device

The DH device is a wireless headband worn during sleep which records, stores, and automatically analyzes physiological data in real time with-out any connection (e.g. Bluetooth, Wi-Fi, etc.). Following the recording, the DH connects to a mobile device (e.g. smart phone and tablet) via Bluetooth to transfer aggregated metrics to a dedicated mobile application and via Wi-Fi to transfer raw data to the sponsor's servers. Five types of physiological signals are recorded via three types of sensors embedded in the device: (1) brain cortical activity via five EEG dry electrodes yielding seven derivations (FpZ-O1, FpZ-O2, FpZ-F7, F8-F7, F7-O1, F8-O2, FpZ-F8; 250 Hz with a 0.4–35 Hz bandpass filter); (2–4) movements, position, and breathing frequency via a 3D accelerometer located over the head; and (5) heart rate via a red-infrared pulse oximeter located in the frontal band. The EEG

electrodes are made of high consistency silicone with soft, flexible protrusions on electrodes at the back of the head enabling them to acquire signal from the scalp through hair. An audio system delivering sounds via bone conduction transducers is integrated in the frontal band but was not active in this study. The DH is composed of foam and fabric with an elastic band behind the head making it adjustable such that it is tight enough to be secure, but loose enough to minimize discomfort. Additional details have been published previously in [19].

Data analysis

We divided data analysis into three parts: (1) EEG signal quality; (2) heart rate, breathing frequency, and RRV agreement; and (3) Automatic Sleep Stage classification of the DH compared to scorers' consensus on the PSG.

Assessing EEG brain wave similarity

Following Sterr *et al.* [23] proposed to use the relative spectral power (RSP) computed in frequency bands relevant for sleep analysis as a proxy for assessing the capacity of the DH to monitor EEG waves. We did not expect the RSP from the DH and the one from the PSG to be strictly equal but they should follow the same trends as the subjects evolve through sleep states. RSP was computed in the λ (0.5–4 Hz), θ (4–8 Hz), α (8–14 Hz), and β (15–30 Hz) bands, every 30 s using Fast Fourier Transform on both devices. Exponential smoothing with $\alpha = 0.7$ was applied to the resulting RSP to avoid abrupt transitions. As a measure of concordance between the DH and the gold standard PSG RSP trends, the mean percentage error (MPE) was computed on the resulting RSP for each 30 s window and averaged for each record. To maximize the time of the night with good signal quality on the DH, we developed a procedure to select a virtual channel which corresponds to the EEG frontal-occipital channel (FpZ-O1, FpZ-O2, F7-O1, or F8-O2) with the best quality signal at any given epoch throughout the night; previously described in [19]. We assessed RSP on the virtual channel for the DH. We excluded periods in which the virtual channel could not be computed on the DH signal because of bad signal quality on all channels of the DH (2.1% of the windows on average across all the recordings). For a fair comparison, we created and used a similar derivation on the PSG: F3-O1 to compute the RSP and hence the final MPE between the DH and the PSG. As a baseline for the order of magnitude of MPE to expect with the same device but using a different electrodes locations (but also frontal occipital), we also computed the MPE between the RSP of the PSG F3-O1 PSG derivation and the PSG F4-O2 derivation for each frequency band using the same process.

Assessing heart rate, breathing frequency, and RRV agreement

The agreement of DH measurements of heart rate frequency (beats per minute) and breathing frequency (in cycle per minute) with PSG measurements of the same variables was also assessed. To do so, values were computed every 15 s (on 30-s sliding windows) on DH data and compared to the respective PSG values using an average of the mean absolute error (MAE) computed for each record. An analogous method was employed to assess the capacity of the DH to retrieve RRV (in percentage), as described in [30].

Heart rate

Heart rate measurement was computed from the oximeter and provided directly by the PSG device in the recorded data. On the DH, heart rate was computed a posteriori. It was derived from the pulse oximeter infrared signal using the following process:

(1) Infrared signal was filtered between 0.4 and 2 Hz and zero crossing was applied to compute the mean heart rate frequency **fs**.
(2) Infrared signal was filtered between *fs*/1.25 and *fs**1.25 and zero crossing was applied to compute the minimum and maximum heart rate frequencies fs_{min} and fs_{max}.
(3) Infrared signal was filtered between $\alpha = 0.3$ was applied on both the DH and PSG heart rates to avoid brutal transitions.

This standard method provides a robust measure of heart rate frequency during sleep. However, it would probably be ill-suited for waking measurement where artifacts and noise are more likely to occur. Of note, one record was excluded from heart rate analyses because the PSG heart rate measurement remained at the same value for the entire duration of the record and was therefore assumed to be inaccurate.

Breathing frequency

Breathing frequency was computed from the z-axis of the accelerometer on the DH and from the external pressure signal on the PSG. To compute breathing frequency, an analogous three-step process to that for the heart rate computation was followed, using a filter between 0.16 and 0.3 Hz in the first step on both the PSG and the DH.

Respiration rate variability

The RRV was computed with the exact same methodology than the one described in [30], except that the method employed here computed RRV throughout the entire night instead of on steady sleep windows. For the PSG, the method was applied to the external pressure channel. For the DH, the RRV was computed on the three axes of the accelerometer and the minimum value between the three was kept for each computed value to reduce noise.

Assessing sleep stages classification performance

Due to known inter-rater variance among even expert sleep scorers [31], using a single rater as the reference point renders comparison vulnerable to unintended bias. Thus, each PSG records were independently scored by five trained and experienced registered sleep technologists from three different sleep centers following the guidelines of the AASM [4]. The DH data was scored by the embedded automatic algorithm of the DH.

Scoring performance metrics

Accuracy (ratio of correct answers) and Cohen's Kappa, $\kappa = (p_j - p_e)/(1 - p_e)$, where p_j is the scorer accuracy and p_e is the baseline accuracy, are provided to measure agreement between two hypnograms on a record. The F1 score was also computed because it takes into account both Precision and Recall, as well as class imbalance, making it a rigorous metric for evaluating performance [32]. It is computed as:

$$F1\ \text{score} = 2 * \left[\frac{\text{Pr} * \text{Re}}{\text{Pr} + \text{Re}}\right]$$

with precision $[TP/(TP + FP)]$ and recall $[TP/(TP + FN)]$, where TP, FP, and FN are the number of true positives, false positives, and false negatives, respectively. This score is computed per-class and averaged taking the weight of each class into account. For "overall" analyses, the average of the respective values from each individual record is calculated.

Scoring performance metrics evaluation. To evaluate scoring performance metrics and benefit from the multiple sleep experts scorings, a similar methodology to [12] was used. Indeed, to evaluate the performance metrics for each scorer, the scoring from each individual scorer was compared to the consensus scoring of the four other scorers. To evaluate the performance metrics of the DH automatic approach, the automatic scoring from the DH was compared to the consensus scoring of the four top-ranked scorers. This method ensures that both the individual scorers and the automatic algorithm running on the DH data were evaluated against a consensus of exactly four scorers. The idea behind using a consensus scoring instead of doing one-by-one evaluations is that a consensus is a more robust point of comparison than a single scorer thanks to the majority vote, especially on epochs that are difficult to score. Performance metrics computed from one-by-one comparison are also provided.

Building a consensus scoring from multiple scorings. Thus, we developed a way to build a unique consensus scoring from multiple scorings on a record. For each epoch, the majority opinion across scorers is chosen. In case of a tie, the sleep stage scored by the top ranked scorer is used (scorer ranking procedure described below, as Soft-Agreement); ties occurred on 7.3 ± 2.4% of the epochs on average across all the records.

Scorer ranking

The previous section highlights the need to rank scorers in order to build a valid consensus scoring. The ranking of a scorer is based on his level of agreement with all the other scorers. To measure this, we introduce below an agreement metric between one scoring against multiple other scorings. We call this metric "Soft-Agreement" as it takes all the scorings into account and does not require any thresholding.

Notations

Let $y_j \in [[4]]^T$ be the sleep staging associated to scorer J taking values in {0, 1, 2, 3, 4} standing, respectively, for Wake, N1, N2, N3 and REM with size $\mathbf{T}$ epochs. Let $\mathbf{N}$ be the number of scorers. Let $\hat{y}_j \in \{0,1\}^{5\times T}$ be the one hot encoding of $\mathbf{y}j$. For each epoch $t \in [[T]]$ its value is 1 for the scored stage and 0 for the other stages.

First, we define a probabilistic consensus $\hat{z}_j$ as

$$\hat{z}_j[t] = \frac{\sum_{i=1;\ i\neq j}^{N} \hat{y}_i[t]}{\max \sum_{i=1;\ i\neq j}^{N} \hat{y}_i[t]}\ \forall t,$$

where $\hat{z}_j$takes values in $[0,1]^{5\times T}$. For each epoch $t \in [[T]]$ the value for each sleep stage is proportional to the number of scorers who scored that sleep stage. A value of '1' is assigned if the chosen stage matches the majority sleep stage or any of the sleep stages involved in a majority tie. We then define the Soft-Agreement for scorer j as:

$$\text{Soft-agreement}_j = \frac{1}{T}\sum_{t=0}^{T} \hat{z}_j[y_j].$$

A Soft-agreement of 1 means that for all epochs, scorer J scored the same sleep stage as the majority and, in case of tie, he scored one of the sleep stages involved in the tie. A Soft-Agreement of 0 would happen if scorer J systematically scores a different stage than *all* of the other scorers. To rank the five scorers in this study, the Soft-agreement was computed for each scorer against the four others on each record and then averaged across all the records. Based on these values, we are able to build unique consensus scorings for comparison with each scorer. To build the consensus scorings for comparison with the DH automatic algorithm, scorings from the top-four scorers were used.

Performance assessment of sleep variables

The following standard sleep variables were calculated: time in bed (TIB), as the number of minutes from lights-out to lights-on; TST (min); SE (%), as TST/TIB×100; sleep onset latency (SOL), as the number of minutes from lights-out to the first three consecutive epochs of any sleep stage; WASO, as the number of minutes awake following the first three consecutive epochs of any sleep stage; and the time (min) and percentage of TST spent in each sleep stage (N1, N2, N3, and REM).

DH algorithm

The DH embedded automatic algorithm works in two stages: (1) feature extraction and (2) classification. It is able to provide real-time sleep staging predictions. (1) Feature extraction is performed for each new epoch of 30 s. Features extracted from the various sensors are concatenated to go through the classification layer. EEG features include power frequency in the λ, α, θ, and β bands and ratio of relative powers as described in 23. Sleep patterns (e.g. slow oscillations, α rhythm, spindles, and K-complexes) are detected using an expert approach. The accelerometer provides breathing, movement, and position features. The pulse oximeter provides cardiac features. A total of 79 features are extracted from each raw DH record. (2) The classification module is built from two layers of Long-Short Term Memory [33] and a Softmax function outputting the final probability prediction that the epoch belongs to each sleep stage. It relies on the features extracted from the last 30 epochs to predict the current one. Hence, it takes into account the past temporal context to make a prediction, as a sleep expert would do. This classification module is trained using backpropagation. The training has been done on a dataset composed of previously recorded

Table 2. Mean percentage error for α, β, λ, and θ EEG relative spectral power between the DH and PSG1 (F3–O1)

	DH/PSG F3–O1	PSG F3–O1/PSG F4–O2
Alpha	15 ± 3.5	7.7 ± 2.4
Beta	16 ± 4.3	8.6 ± 3.7
Delta	16 ± 6.1	7.2 ± 2.7
Theta	10 ± 1.4	7.4 ± 2.1

Values are also provided for PSG1 (F3–O1) versus PSG2 (F4–O2) as a baseline.

internal Dreem records. A total of 423 records were used for training and presented several times to the network. A total of 213 validation records from other subjects were used to stop the training when the performance metrics computed on this validation set were not improving anymore. None of the records of the current study were used to train or validate the network. We used the framework provided by Pytorch [34] and trained on a single Nvidia Titan X GPU (~1 h of training, ~1 s for inference).

Results

EEG Brain Waves Similarity. The quality of the EEG signal assessed through the MPE of the relative spectral power between DH and PSG for α, β, λ, and θ frequencies is presented in Table 2. Results indicate a MPE around 15% for α, β, and λ and 10% for θ between the DH and PSG. As expected, the MPE between the DH and the PSG are higher than the baseline MPE between the two derivations from the same PSG record (F3-O1 and F4-O2) which are around 7%–8% for all the frequency bands. Figure 1 shows a sample of raw signals recorded by the DH and a PSG on the same record during each sleep stage (N1, N2, N3, REM, and Wake). Figure 2 shows the relative spectral power between DH and PSG for each EEG frequency examined throughout a representative record as well as the corresponding MPE.

Heart rate and breathing agreement

The agreements of heart rate, breathing frequency, and respiratory rate variability measured by the DH are not statistically different as the one measured by the PSG are presented (Table 3).

Figure 2 shows an example of heart rate, breathing frequency, and RRV measured on the PSG throughout the night.

Sleep stage classification

Results show a soft-Agreement scores of 88.6%, 90.7%, 91.7%, 84.2%, and 91.6% for scorers 1, 2, 3, 4, and 5, respectively (overall soft-agreement score = 89.4 ± 2.79%). With these values, we were able to develop consensuses with which to compare each scorer and the predictions of the DH automatic algorithm for the purpose of evaluating the metrics presented in Table 4. The overall accuracy of the five scorers is of 86.4 ± 7.4%, with scorer 1 = 86.3 ± 10.5%, scorer 2 = 88.2 ± 4.2%, scorer 3 = 88.9 ± 5.1%, scorer 4 = 82.0 ± 8.1%, and scorer 5 = 88.9 ± 4.6%. Notably, these accuracies are above the average performance of other certified scorers reported in the literature [8], indicating the scorers in this study were well-trained. Across the manual scorers, accuracy was highest for REM sleep (87.8 ± 13.6%) and followed closely by N2 (85.9 ± 10.7%) and N3 sleep stages (84.2 ± 20.6%). The accuracy for wake was slightly lower (82.5 ± 17.5%). The accuracy was the lowest for N1 (54.2 ± 16.8%).

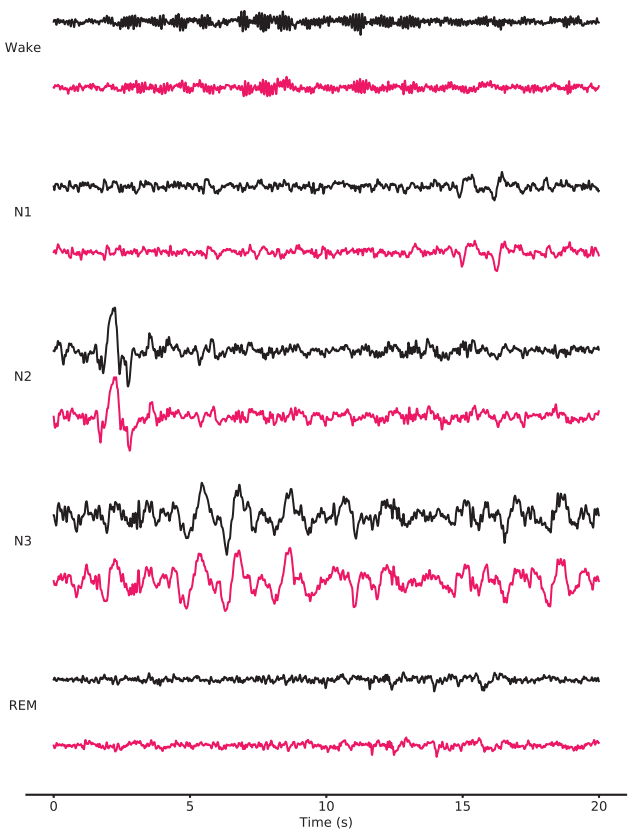

Figure 1. 20-s samples of raw signals recorded by DH (pink) and PSG (black) on the same record during each sleep stages (N1, N2, N3, REM, and Wake). The derivations are F7–O1 for the DH and F3–O1 for the PSG. The signals are presented between –150 and 150 μV.

The overall accuracies of the DH automated algorithm using the DH data for sleep staging compared to the scorer consensuses are presented in Table 4 (overall accuracy = 83.8 ± 6.8%). The classification accuracy per stage of the DH parallels the order of the manual scorers using PSG data: highest accuracy for REM sleep (84.5 ± 13.3%) followed by N2 (82.9 ± 8.1%) and N3 sleep stages (82.6 ± 20.6%). The accuracy for wake was lower (74.0 ± 18.1%) with the lowest accuracy similarly obtained for the N1 sleep stage (47.7 ± 15.6%). Performance metrics computed using one-by-one comparison are presented in supplementary. They do not impact the relative performance of the scorers and the DH but present lower values and higher variance. This confirms the assumption that evaluation against a consensus is a more robust way of measuring performance than comparing one-by-one.

The confusion matrices (Figure 3) show the classifications per stage of both the DH and scorer averages versus the respective consensuses. According to the matrices, both in the case of the DH and the PSG scoring, Wake is most often misclassified as N1 (12.4% and 10.5% of epochs, respectively), and N3 is most often misclassified as N2 (16.4% and 20.5% of epochs for DH and PSG, respectively). Figure 4 shows five representative hypnograms computed from the DH classifications and the corresponding scorer consensus hypnograms.

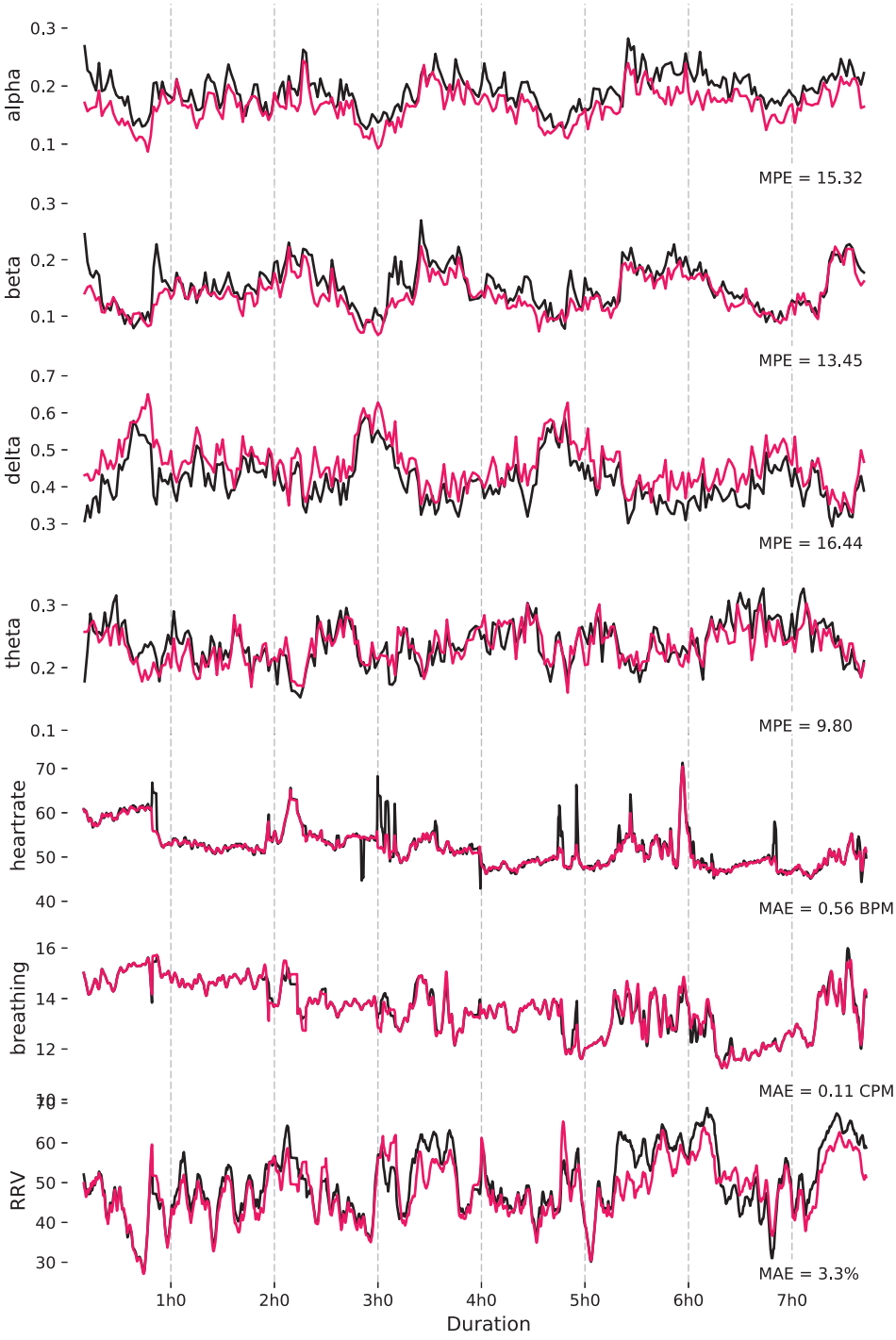

Figure 2. Relative spectral power (α, β, λ, and θ frequencies, AU), heart rate (beats per minute, BPM), breathing frequency (cycles per minute, CPM), and respiratory rate variability (RRV, %) for a representative record (i.e. with a MPE similar to the mean of the group). These signals are presented for a whole record for both the DH (pink) and PSG (black).

Table 3. The DH and PSG columns presents mean values ± SD computed across all records for heart rate, breathing frequency, and respiratory rate variability (RRV) for both devices. Average mean absolute error (MAE) by record is given in the last column.

	PSG	DH	MAE
Heart rate (bpm)	61.3 ± 6.8	60.6 ± 6.5	1.2 ± 0.5
Breathing (cpm)	14.9 ± 1.9	14.8 ± 1.8	0.3 ± 0.2
RRV (%)	53.5 ± 2.2	52.2 ± 2.8	3.2 ± 0.6

The averages of sleep variables are not statistically different when comparing the consensus (top-four ranked scorers), the DH, the differential DH (average per-record difference observed between the DH and the scorer consensus), and the overall differential scorers (average per-record difference observed between each scorer and the scorer consensus formed by the four other scorers; Table 5).

The evaluation of the agreements between the DH and experts scoring on the PSG for each sleep stage are reported by the

Table 4. Performance metrics for each scorer and the automatic approach of the DH computed by comparison against their consensus

		DH	Overall scorers	Scorer 1	Scorer 2	Scorer 3	Scorer 4	Scorer 5
All	F1 (%)	83.8 ± 6.3	86.8 ± 7.4	86.3 ± 10.5	88.2 ± 4.2	88.9 ± 5.1	82.0 ± 8.1	88.9 ± 4.6
	Accuracy (%)	83.5 ± 6.4	86.4 ± 8.0	85.7 ± 12.1	87.5 ± 4.5	88.9 ± 4.6	81.2 ± 8.8	88.9 ± 4.2
	Cohen Kappa (%)	74.8 ± 10.4	79.8 ± 11.4	78.9 ± 15.7	81.2 ± 7.0	83.2 ± 7.2	72.5 ± 13.2	83.0 ± 7.2
Wake	F1 (%)	76.7 ± 14.3	84.1 ± 13.6	85.9 ± 10.1	86.5 ± 12.3	87.6 ± 9.9	74.9 ± 18.0	85.6 ± 11.9
	Accuracy (%)	74.0 ± 18.1	82.5 ± 17.5	80.2 ± 14.0	78.1 ± 19.3	90.2 ± 12.9	82.4 ± 21.0	81.6 ± 16.5
	Cohen Kappa (%)	74.1 ± 15.2	82.2 ± 15.0	84.4 ± 10.7	85.2 ± 12.6	86.3 ± 10.4	71.1 ± 20.5	84.1 ± 12.8
N1	F1 (%)	46.5 ± 12.4	49.7 ± 14.5	49.3 ± 13.8	51.2 ± 11.4	53.7 ± 13.7	39.7 ± 16.3	54.5 ± 11.6
	Accuracy (%)	47.7 ± 15.6	54.2 ± 16.8	58.2 ± 14.7	60.3 ± 12.6	59.1 ± 14.4	38.3 ± 15.5	54.9 ± 16.3
	Cohen Kappa (%)	43.5 ± 12.6	47.0 ± 15.1	46.6 ± 14.0	48.5 ± 11.8	51.4 ± 13.9	36.4 ± 17.3	52.3 ± 12.1
N2	F1 (%)	87.5 ± 5.5	89.0 ± 7.3	88.2 ± 11.8	90.3 ± 4.6	90.7 ± 4.1	84.3 ± 6.9	91.3 ± 3.5
	Accuracy (%)	82.9 ± 8.1	85.9 ± 10.7	87.8 ± 13.6	87.6 ± 8.7	89.3 ± 6.0	75.8 ± 10.9	89.0 ± 5.7
	Cohen Kappa (%)	75.4 ± 10.8	78.8 ± 13.2	77.5 ± 20.5	81.1 ± 8.6	81.5 ± 8.5	71.1 ± 12.9	82.9 ± 7.1
N3	F1 (%)	76.4 ± 22.9	78.3 ± 23.8	81.0 ± 24.1	79.6 ± 22.8	76.8 ± 25.0	75.2 ± 22.1	78.9 ± 24.7
	Accuracy (%)	82.6 ± 20.6	84.2 ± 20.6	89.1 ± 14.0	89.1 ± 11.1	66.9 ± 27.6	92.7 ± 12.0	84.1 ± 21.3
	Cohen Kappa (%)	74.0 ± 22.6	76.6 ± 23.1	79.2 ± 24.1	77.2 ± 22.8	74.8 ± 25.0	74.9 ± 17.9	76.8 ± 24.3
REM	F1 (%)	82.9 ± 12.3	90.9 ± 10.2	85.0 ± 17.0	91.4 ± 4.2	94.4 ± 4.6	91.5 ± 8.8	92.1 ± 8.1
	Accuracy (%)	84.5 ± 13.3	87.8 ± 13.6	76.3 ± 19.4	85.8 ± 6.5	92.1 ± 9.5	89.5 ± 13.6	95.3 ± 3.1
	Cohen Kappa (%)	79.0 ± 13.7	89.0 ± 11.1	82.5 ± 17.6	89.4 ± 5.3	93.1 ± 5.1	89.7 ± 10.5	90.5 ± 8.9

Overall column presents mean ± SD observed for the five scorers. Results are given for each sleep stages.

Figure 3. Confusion matrix for the DH versus PSG scoring consensuses (top) and the overall confusion matrix for scorers versus the other scorers' consensuses (bottom). Values are normalized by row with the number of epochs in parentheses.

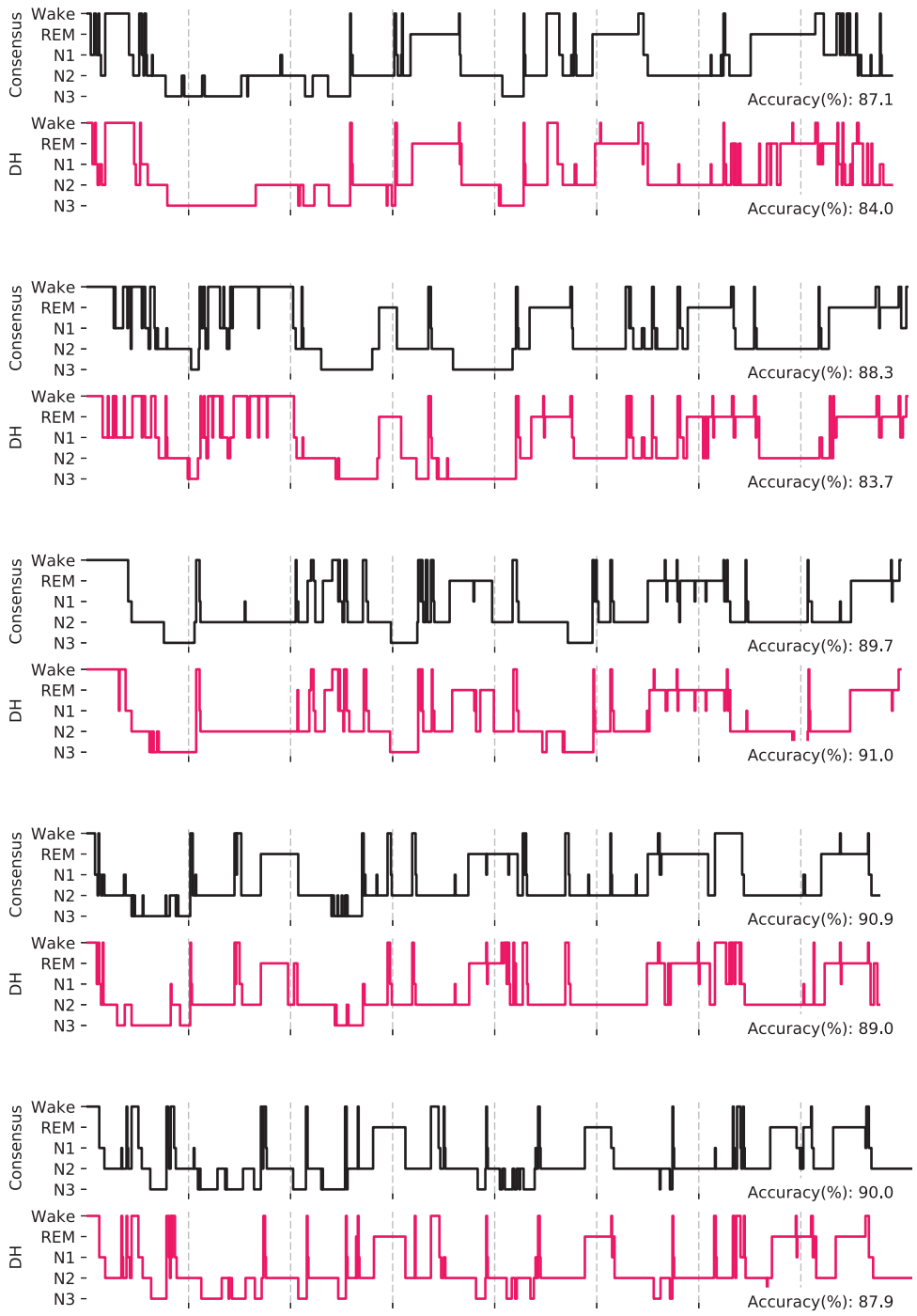

Figure 4. Hypnograms for the five first participants showing both the consensuses of the four top-ranked scorers (gray) and the DH automated sleep stage classifications. Accuracies are presented as average obtained by the five scorers on the consensus hypnogram, and scores obtained for the DH versus the consensuses.

Bland Altman plots (Figure 5). This figure shows an evenly dispersion around the mean with no value beside the significance threshold.

Discussion

While considerable value could be derived from longitudinal sleep EEG monitoring, until recently, (wet) EEG electrodes were too impractical to be easily used on a regular basis and without assistance. In broader domains than sleep, it has been shown that emerging technologies using dry electrodes were able to accurately monitor EEG, paving the way to meaningful physiological monitoring at home under various conditions [35, 36].

In this study, we first showed that the DH that includes dry-EEG electrodes embedded in a headband could measure EEG frequencies in a similar way than a PSG.

It has to be noted that typical EEG patterns such as α rhythm spindles, λ waves, or K complexes have not been analyzed here.

Table 5. The consensus column presents sleep variables computed with the scorer's consensus of the top-four ranked scorers

	Consensus	DH	Differential DH	Overall differential scorers
TST (min)	430.4 ± 66.8 [402.2, 458.5]	431.8 ± 68.4 [402.9, 460.6]	1.4 ± 19.0 [−6.6, 9.4]	−2 ± 28.5 [−7.1, 3.0]
Sleep efficiency (%)	87 ± 8.1 [83.5, 90.4]	87.3 ± 8.4 [83.8, 90.8]	0.3 ± 4.0 [−1.4, 2.0]	−0.4 ± 5.6 [−1.4, 0.6]
SOL (min)	18.9 ± 18.7 [11.0, 26.8]	20 ± 18.2 [12.4, 27.7]	1.2 ± 3.6 [−0.4, 2.7]	−1.7 ± 15.6 [−4.5, 1.1]
WASO (min)	44.1 ± 27.8 [32.4, 55.8]	41.5 ± 31.7 [28.1, 54.9]	−2.6 ± 19.5 [−10.8, 5.7]	3.7 ± 28.4 [−1.3, 8.8]
Stage wake duration (min)	61.8 ± 38.8 [45.5, 78.1]	60.2 ± 40.7 [43.0, 77.3]	−1.7 ± 19.1 [−9.7, 6.4]	2 ± 28.5 [−3.1, 7.0]
Stage Wake (%)	12.8 ± 8.1 [9.4, 16.2]	12.4 ± 8.3 [8.9, 15.9]	−0.4 ± 4.0 [−2.1, 1.3]	0.4 ± 5.6 [−0.6, 1.4]
Stage N1 duration (min)	31 ± 14.4 [25.0, 37.1]	31.9 ± 17.9 [24.4, 39.4]	0.8 ± 12.9 [−4.6, 6.3]	6 ± 25.2 [1.5, 10.5]
Stage N1 (%)	6.3 ± 2.9 [5.1, 7.5]	6.5 ± 3.7 [5.0, 8.1]	0.2 ± 2.6 [−0.9, 1.3]	1.2 ± 4.4 [0.4, 1.9]
Stage N2 duration (min)	244.3 ± 59.2 [219.4, 269.2]	227.9 ± 57.0 [203.9, 251.9]	−16.4 ± 21.8 [−25.5, −7.2]	−9.7 ± 38.7 [−16.6, −2.8]
Stage N2 (%)	49.2 ± 9.3 [45.3, 53.1]	46.1 ± 9.9 [41.9, 50.2]	−3.1 ± 4.3 [−5.0, −1.3]	−1.8 ± 7.5 [−3.2, −0.5]
Stage N3 duration (min)	61.4 ± 30.4 [48.6, 74.2]	70.5 ± 37.1 [54.8, 86.1]	9.1 ± 19.5 [0.8, 17.3]	6.3 ± 32.8 [0.5, 12.1]
Stage N3 (%)	12.5 ± 6.1 [9.9, 15.0]	14.2 ± 6.9 [11.3, 17.1]	1.7 ± 3.8 [0.1, 3.3]	1.3 ± 6.1 [0.2, 2.3]
REM sleep duration (min)	93.6 ± 32.3 [80.0, 107.2]	101.5 ± 41.7 [83.9, 119.1]	7.9 ± 24.9 [−2.5, 18.4]	−4.7 ± 16.7 [−7.7, −1.7]
REM sleep (%)	19 ± 6.2 [16.4, 21.6]	20.5 ± 7.9 [17.1, 23.8]	1.5 ± 5.0 [−0.6, 3.6]	−1 ± 3.3 [−1.5, −0.4]

The DH column present the sleep variables computed on the DH. Differential DH column presents the average per-record difference observed between the DH and the scorer consensus. Overall Differential Scorers presents the average per-record difference observed between each scorer and the scorer consensus formed by the four other scorers. Results are presented as Mean ± SD [0.95CI].

This could be integrated in future work where a trained human could both score the DH and the PSG.

HR and breathing frequency are two key physiological signals that are closely looked at in PSG recordings. These two signals are critical for deeper phenotyping of healthy and pathological sleep and might therefore be of interest in broader longitudinal sleep studies. Therefore, these data were analyzed and presented in this article. We showed that the method used for detecting breathing frequency and RRV using an accelerometer had strong agreement with the gold standard. The position of the 3D accelerometer, located over the head, appeared to be a sensitive location for detecting small movements. The agreement for heart rate is similar to other studies showing that an infrared pulse oximeter positioned against the forehead can be used to reliably monitor heart rate. However, we were unable to provide a heart rate variability on most of the records due to insufficient resolution, similar to other studies [37].

Finally, we showed that the DH was able to perform real-time sleep staging using data collected by the DH with an accuracy in the range of individual scorers using PSG data and comparable to the accuracy between PSG scorers in other studies [8, 31]. To our knowledge, this performance on a dry-EEG wearable has never been achieved with another device. Sleep variables are macro-metrics computed on the hypnogram and are less impacted than sleep staging metrics by local differences. For instance, wake is slightly underestimated but that does not significantly impact sleep variables related to wake (WASO, sleep latency, and sleep efficiency). Even though the inter-scorer reliability achieved with PSG by our 5 scorers was high, it highlights the need for such validation studies to rely on a consensus of multiple sleep experts when analyzing sleep staging performance [31]. Mixing sleep experts from different sleep centers provides a more realistic analysis than is typically obtained in a standard clinical sleep study where records are scored by only a single individual, which strengthens our results. To evaluate these individual scorers, we introduced an objective methodology to build a consensus from the other scorers. This enables a fair evaluation of both individual scorers and the automated algorithmic approach of the DH.

The main limitation of this study is that the sample was somewhat small and homogeneous in age and sleeper profile; even though this is consistent with the majority of similar validation studies [20, 22, 23]. A larger sample of more diverse sleepers would have provided more reliability and generalizability to the general population. Therefore, further studies should be run with this investigational device on specific sub-populations (e.g. patients with sleep apnea, psychiatric or neurodegenerative diseases, etc.) either to confirm the performance of the automatic sleep stage algorithm but also to evaluate the ease of use of the device and comfort in these specific targeted populations.

In this study, 2.1% of the windows were excluded on average across all the recordings because the *virtual channel* could not be computed on the DH signal due to bad signal quality on every channel. Since the DH was wore as long as with the PSG, in the lab setting. It can then be argued that in the home-environment, a more important proportion of the signals would have been of bad quality due to the negligence of the subject. However, it has to be put in line with the fact that in a home setting, it would have been possible to run multiple nights, which is not easy to do in the lab with a PSG. Also, this issue can be easily coped with an appropriate training of the subject to use the device. Also, this study includes only one night of data per subject with no habituation night, which may lead to a non-reliable representation of typical sleep in the natural home environment, and particularly because sleeping with a full PSG in a clinical sleep lab, which often leads to sleep being shorter and more fragmented. However, our sample did achieve 87% SE on average, suggesting that sleep was not substantially disrupted on a wide scale in this study.

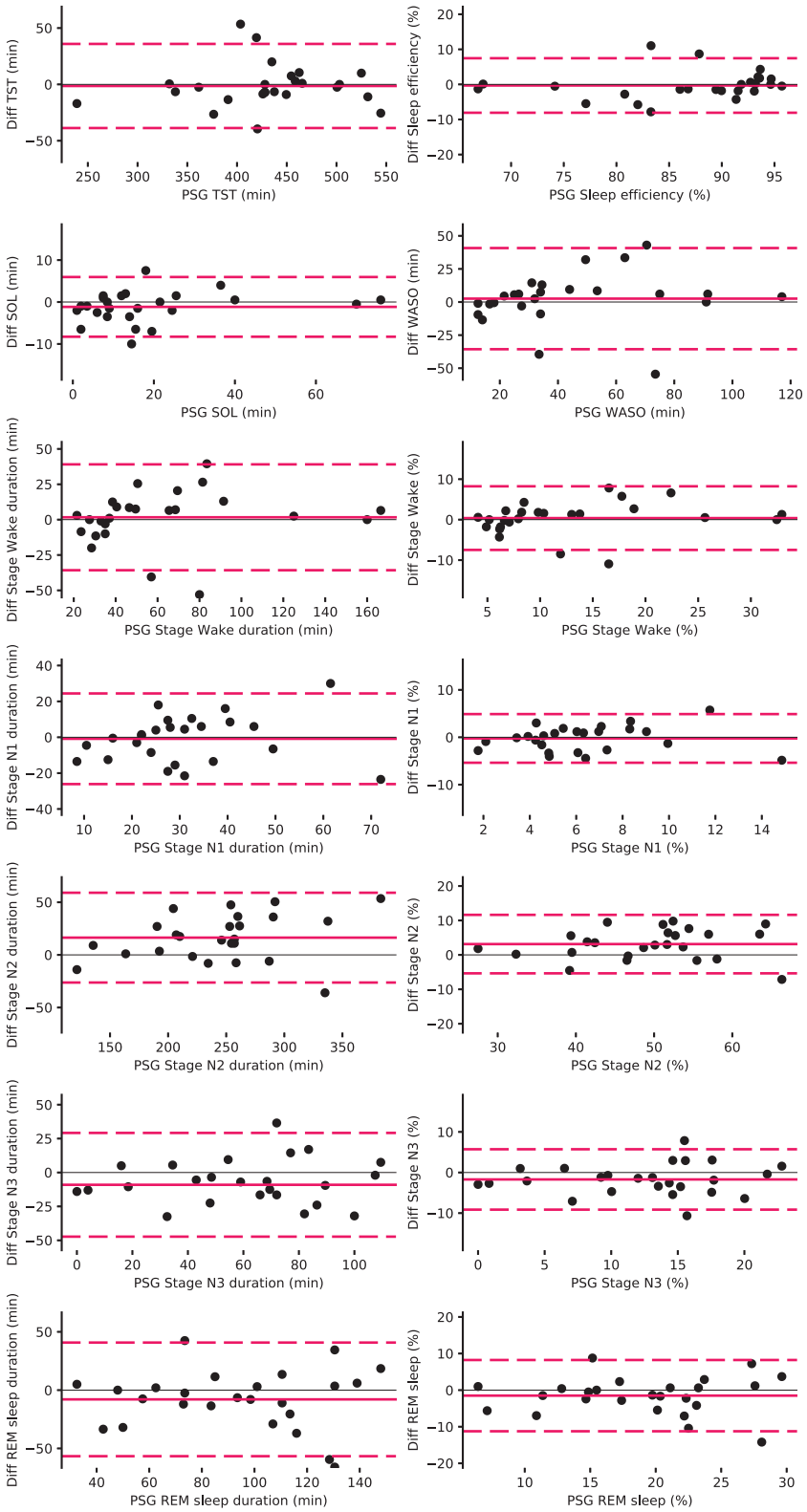

Figure 5. Bland Altman plots for each sleep variable measures by the DH versus the consensus sleep metrics computed for each record.

Conclusion

In this study, we showed that using an ambulatory wireless dry-EEG device, the DH, it was possible to: (1) acquire EEG signals that correlate with the EEG signals recorded with a PSG; (2) reliably measure breathing frequency and heart rate continuously during sleep; and (3) perform automatic sleep staging classification according to AASM criteria with performance similar to that of a consensus of five scorers using medical-grade PSG data.

These results, together with the price, ease of use and the availability of raw signals, pave the way for such a device to be an ideal candidate for high-quality large-scale longitudinal sleep studies in the home or laboratory environment. As such, this technology can enable groundbreaking advancements in sleep research and medicine. For instance, the resulting database can ultimately be integrated with other types of data collection devices and used to identify unknown patient subgroups, detect early disease biomarkers, personalize therapies, and monitor neurological health and treatment response.

Acknowledgments

We would like to thank the Fatigue and Vigilance team including Drogou C., Erblang M., Dorey R., Quiquempoix M., Gomez-Merino D., and Rabat A. for their help in the study. We would like to thank Mignot E. for his help on the manuscript. We also would like to thank the Dreem team for their commitment to working on the Dreem headband over these years.

Disclosure Statements

Non-financial disclosure: F.S. was the principal investigator of the study. F.S., M.G., M.C., and P.V.B. declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Financial disclosure: This study was supported by Dreem sas. P.J.A. and M.E.B. are employees of Dreem, Inc. and V.T., A.B.H., A.G., M.H., E.D., and H.J. of Dreem sas.

Author contributions

Study concept and design: P.J.A., M.E.B., M.G., M.C., F.S., E.D. Data acquisition: H.J., M.H., P.V.B., M.G., F.S. Data analysis: V.T., A.B.H., A.G. Data interpretation: P.J.A., V.T., F.S. Writing the manuscript: P.J.A., V.T., F.S. Revising the manuscript: V.T., E.D.

References

1. National Center on Sleep Disorders Research and others. *National Institutes of Health Sleep Disorders Research Plan*. Bethesda, MD: National Institutes of Health; 2011. Report No. DOT HS 808 707.
2. Sateia MJ. International classification of sleep disorders—third edition: highlights and modifications. *Chest*. 2014;**146**:1387–1394.
3. Ohayon MM. Epidemiological overview of sleep disorders in the general population. *Sleep Med Res*. 2011;**2**(1):1–9.
4. Iber C, *et al*. *The AASM Manual for the Scoring of Sleep and Associated Events : Rules, Terminology, and Technical Specifications*. Westchester, IL: American Academy of Sleep Medicine; 2007.
5. White LH, *et al*. Night-to-night variability in obstructive sleep apnea severity: relationship to overnight rostral fluid shift. *J Clin Sleep Med*. 2015;**11**(2):149–156.
6. Bittencourt LR, *et al*. The variability of the apnoea-hypopnoea index. *J Sleep Res*. 2001;**10**:245–251.
7. Rosenberg RS, *et al*. The American Academy of Sleep Medicine Inter-scorer Reliability program: respiratory events. *J Clin Sleep Med*. 2014;**10**:447–454.
8. Danker-Hopfe H, *et al*. Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. *J Sleep Res*. 2009;**18**:74–84.
9. Sun H, *et al*. Large-scale automated sleep staging. *Sleep*. 2017;**40**(10). doi:10.1093/sleep/zsx139
10. Patanaik A, *et al*. An end-to-end framework for real-time automatic sleep stage classification. *Sleep*. 2018;**41**:1–11.
11. Chambon S, *et al*. A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series. *IEEE Trans Neural Syst Rehabil Eng*. 2018;**26**:758–769.
12. Stephansen JB, *et al*. Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nat Commun*. 2018;**9**:5229.
13. Biswal S, *et al*. Automated sleep staging system via deep learning. SLEEPNET. [Submitted on 26 Jul 2017] http://arxiv.org/abs/1707.08262.
14. Kuna ST, *et al*. Agreement in computer-assisted manual scoring of polysomnograms across sleep centers. *Sleep*. 2013;**36**(4):583–589.
15. Chambon S, *et al*. DOSED: a deep learning approach to detect multiple sleep micro-events in EEG signal. *J Neurosci Methods*. 2019;**321**:64–78.
16. Gandhi M, *et al*. *The Future of Biosensing Wearables*. *Rock Health*. https://rockhealth.com/reports/the-future-of-biosensing-wearables/ Published 2014. Accessed May 28, 2020.
17. Marino M, *et al*. Measuring sleep: accuracy, sensitivity, and specificity of wrist actigraphy compared to polysomnography. *Sleep*. 2013;**36**:1747–1755.
18. Montgomery-Downs HE, *et al*. Movement toward a novel activity monitoring device. *Sleep Breath*. 2012;**16**:913–917.
19. Debellemaniere E, *et al*. Performance of an ambulatory dry-EEG device for auditory closed-loop stimulation of sleep slow oscillations in the home environment. *Front Hum Neurosci*. 2018;**12**:88.
20. Shambroom JR, *et al*. Validation of an automated wireless system to monitor sleep in healthy adults. *J Sleep Res*. 2012;**21**:221–230.
21. Garcia-Molina G, *et al*. Closed-loop system to enhance slow-wave activity. *J Neural Eng*. 2018;**15**:066018.
22. Finan PH, *et al*. Validation of a wireless, self-application, ambulatory electroencephalographic sleep monitoring device in healthy volunteers. *J Clin Sleep Med*. 2016;**12**:1443–1451.
23. Sterr A, *et al*. Sleep EEG derived from behind-the-ear electrodes (cEEGrid) compared to standard polysomnography: a proof of concept study. *Front Hum Neurosci*. 2018;**12**:452.
24. Mikkelsen KB, *et al*. EEG Recorded from the Ear: characterizing the Ear-EEG Method. *Front Neurosci*. 2015;**9**:438.
25. Depner CM, *et al*. Wearable technologies for developing sleep and circadian biomarkers: a summary of workshop discussions. *Sleep*. 2020;**43**(2). doi:10.1093/sleep/zsz254
26. Bastien CH, *et al*. Validation of the insomnia severity index as an outcome measure for insomnia research. *Sleep Med*. 2001;**2**(4):297–307.
27. Johns MW. A new method for measuring daytime sleepiness: The Epworth sleepiness scale. *Sleep*. 1991;**14**(6):540–545.
28. Spitzer RL, *et al*. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Arch Intern Med*. 2006;**166**:1092–1097.

29. Kroenke K, *et al*. The PHQ-9 : a new depression measure. *Psychiatr Ann*. 2002;**32**(9):509–515.
30. Gutierrez G, *et al*. Respiratory rate variability in sleeping adults without obstructive sleep apnea. *Physiol Rep*. 2016;**4**(17):e12949.
31. Van Hout S. The American Academy of Sleep Medicine inter-scorer reliability program: sleep stage scoring Richard S. Rosenberg1. *J Clin Sleep Med*. 2013;**9**(1):81–87.
32. Powers DMW. Evaluation: from precision, recall and F-factor to ROC, informedness, markedness & correlation. *J Mach Learn Technol*. 2011.
33. Hochreiter S, *et al*. Long short-term memory. *Neural Comput*. 1997;**9**:1735–1780.
34. Paszke A, *et al*. Automatic differentiation in PyTorch. NIPS 2017 Workshop Autodiff Decision Program Chairs. 2017;**22**:2–8.
35. Lopez-Gordo MA, *et al*. Dry EEG electrodes. *Sensors (Switzerland)*. 2014;**14**(7):12847–12870.
36. Srinivasa MG, *et al*. Dry electrodes for bio-potential measurement in wearable systems. In: 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore, 2017; pp. 270–276.
37. Lu G, *et al*. Limitations of oximetry to measure heart rate variability measures. *Cardiovasc Eng*. 2009;**9**:119–125.