Sleep Medicine 75 (2020) 54–61

Contents lists available at ScienceDirect

Sleep Medicine

journal homepage: www.elsevier.com/locate/sleep

Original Article

Validation of sleep stage classification using non-contact radar technology and machine learning (Somnofy®)

Ståle Toften [a, *], Ståle Pallesen [b, c], Maria Hrozanova [d], Frode Moen [e], Janne Grønli [f]

[a] Department of Data Science, VitalThings AS, Tønsberg, Norway
[b] Department of Psychosocial Science, University of Bergen, Bergen, Norway
[c] Norwegian Competence Center for Sleep Disorders, Haukeland University Hospital, Bergen, Norway
[d] Department of Neuromedicine and Movement Science, Faculty of Medicine and Health Sciences, Centre of Elite Sport Research, NTNU, Trondheim, Norway
[e] Department of Education and Lifelong Learning, Centre of Elite Sport Research, NTNU, Trondheim, Norway
[f] Department of Biological and Medical Psychology, University of Bergen, Bergen, Norway

ARTICLE INFO

Article history:
Received 8 November 2019
Received in revised form
11 February 2020
Accepted 23 February 2020
Available online 6 March 2020

Keywords:
Polysomnography
Radar
Sleep stages
Somnofy
Validation

ABSTRACT

Objective: To validate automatic sleep stage classification using deep neural networks on sleep assessed by radar technology in the commercially available sleep assistant Somnofy® against polysomnography (PSG).

Methods: Seventy-one nights of overnight sleep in healthy individuals were assessed by both PSG and Somnofy at two different institutions. The Somnofy unit was placed in two different locations per room (nightstand and wall). The sleep algorithm was validated against PSG using a 25-fold cross validation technique, and performance was compared to the inter-rater reliability between the PSG sleep scored by two independent sleep specialists.

Results: Epoch-by-epoch analyses showed a sensitivity (accuracy to detect sleep) and specificity (accuracy to detect wake) for Somnofy of 0.97 and 0.72 respectively, compared to 0.99 and 0.85 for the PSG scorers. The sleep stage differentiation for Somnofy was 0.75 for N1/N2, 0.74 for N3 and 0.78 for R, whilst PSG scorers ranged between 0.83 and 0.96. The intraclass correlation coefficient revealed excellent and good reliability for total sleep time and sleep efficiency, while sleep onset and R latency had poor agreement. Somnofy underestimated total wake time by 5 min and N1/N2 by 3 min. N3 was overestimated by 4 min and R by 3 min. Results were independent of institution and sensor location.

Conclusion: Somnofy showed a high accuracy staging sleep in healthy individuals and has potential to assess sleep quality and quantity in a sample of healthy, mostly young adults. More research is needed to examine performance in children, older individuals and those with sleep disorders.

1. Introduction

The gold standard for objective sleep assessment, polysomnography (PSG), is typically performed in a sleep laboratory and data needs to be manually scored by a sleep technician, making it obtrusive, costly and less suitable for longitudinal studies. Hence, there is a need for validated low-cost equipment for the assessment of sleep, which is user-friendly, accurate and non-intrusive. From a clinical and research perspective, the capacity to obtain longitudinal sleep–wake data may individualize treatment decision and health optimization and improve disease phenotyping.

Radar technology has a great potential for home sleep assessment as it is completely non-intrusive. Technology based on impulse radio ultra-wideband (IR-UWB) radar sensor technique has been shown to reliably monitor vital signs in real-time as respiratory and cardiac events, as well as limb movements [1–4]. Respiration and movement have been shown to correlate with REM and non-REM sleep [5,6]. Recently, an IR-UWB radar was developed for the purposes of sleep assessment [7]. In the pilot validation study against PSG, the IR-UWB radar quantified sleep and wakefulness by an algorithm integrating movements from all body parts. Results of the study revealed a small overall discrepancy between PSG estimates for total sleep time, and the mean sensitivity (radar = sleep when PSG = sleep) and specificity (radar = wake when PSG = wake) were higher or comparable to that reported for actigraphic studies [8]. Despite the promising results of the IR-UWB

* Corresponding author. Department of Data Science, VitalThings AS, Jarlsøveien 48, Tønsberg, Vestfold 3124, Norway.
E-mail address: st@vitalthings.com (S. Toften).

https://doi.org/10.1016/j.sleep.2020.02.022
1389-9457/

radar technique, the algorithm tended to overestimate wake after sleep onset and underestimate sleep onset latency [7].

The radar technology and algorithms for tracking sleep have since been under continuous improvement. These efforts have resulted in the commercially available sleep assistant Somnofy. Somnofy utilizes respiration and movement data derived from radar technology to classify sleep stages using machine learning. In addition to sleep stage classification, Somnofy has built-in sensors for collecting data from the sleeping environment (light intensity and colour composition, audible noise level, room temperature, air quality, air pressure, air humidity). It is also possible to track other relevant data through the Somnofy app (exercise, diet, medication, etc), which can be coupled with other Bluetooth devices to collect relevant data such as heart rate and oxygen saturation.

The aim of the present study was to investigate if Somnofy can provide accurate and reliable classification of sleep stages when compared to PSG. The present study was limited to healthy individuals and mostly involved young subjects. Specifically, the study aimed to validate both overall sleep parameters as well as epoch-by-epoch sleep staging of wakefulness (W), non-REM sleep (N1/N2, N3) and REM sleep (R).

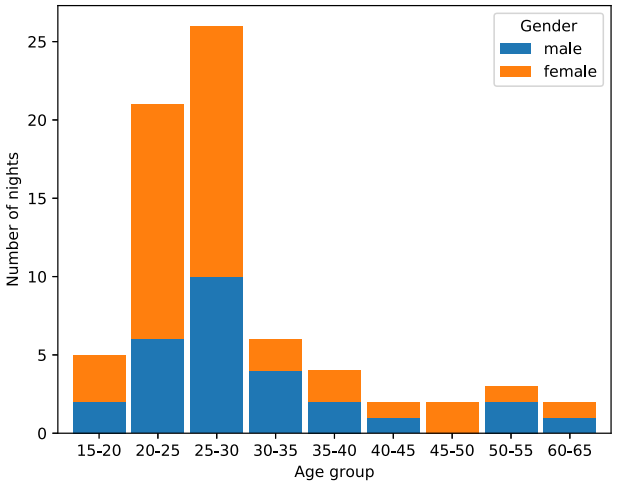

Fig. 1. Distribution of participants age and gender. The x-axis is divided in five years buckets while the y-axis shows the count of participants/nights per bucket. Gender is denoted by orange colour for females and blue for males.

2. Materials and methods

2.1. Participants and data sample

One hundred and two volunteers were recruited through information at lectures among students at the University of Bergen or social media. The inclusion criteria were healthy adults 18 years or above. Twenty-three participants were later excluded from the final analyses as PSG indicated presence of sleep disorders, such as sleep apnoea (AHI > 5, n = 16), periodic limb movement disorder (PLMI > 15, n = 10) and narcolepsy (n = 1). These participants were used for training the algorithm and included in a separate analysis for preliminary results on sleep assessment in participants with sleep disorders. Further, eight recordings were excluded due to missing more than 2 h of Somnofy data. Five nights lacked approximately half an hour of Somnofy data, but these nights were kept and compared to the corresponding PSG recordings. Thus, 71 nights of recordings from 71 different persons (43 females) with a mean age of 28.9 years (SD = 9.7, range 19–61 years) constituted the final data set. The sex and age distribution are shown in Fig. 1.

2.2. Procedure

Assessments took place at two different institutions: at the Colosseum clinic in Oslo, Norway, where participants slept in sound-attenuated bedrooms (n = 37) and at the University of Bergen, Norway, where participants slept at home (n = 34). Lights-on and lights-out times were self-selected. Two Somnofy units were placed per room. One unit was placed at the nightstand (by the head) and one was placed on the wall (above the head). Both units aimed at the participants' chest.

The participants at the sleep clinic were not allowed to drink alcohol 48 h prior to sleep assessment and could not smoke during the assessments. The participants sleeping at home could use tobacco freely but did not consume alcohol the evening before assessments.

2.3. Polysomnography

PSG was performed according to the technical specifications by the American Academy of Sleep Medicine (AASM) [9] with SOMNOscreen plus (SOMNOmedics, Germany). The electrodes included electroencephalography (EEG; F4–M1, F3–M2, C4–M1, C3–M2, O2–M1, O1–M2), bipolar submental electromyography (EMG), and electrooculography (EOG; E1–M2, E2–M1). Additional measurements were used to screen for the presence of any sleep disorder: EMG anterior tibialis, electrocardiogram (ECG), and respiration sensors (nasal cannula, thermistor, thoracic and abdominal respiratory inductance plethysmography and pulse oximetry). Sleep stages (W, N1, N2, N3 and R) were scored in 30 s epochs according to the AASM criteria [9].

A total of five sleep specialists (hereafter named Europe_1, Europe_2, Europe_3, Europe_4 and USA_1 for geographical location) took part in manual scoring, and each recording was scored separately by two sleep specialists.

2.4. Somnofy

Somnofy (version 0.7, VitalThings AS, Norway) with sleep algorithm version 1.0 was used. Somnofy is certified according to the Federal Communication Commission (FCC) and “Conformité Européene” (CE). Movement and respiration from the sleeping person were derived from the IR-UWB radar. Simply put, the radar emitted pulses that were reflected by objects and returned to a receiver. The average sampling rate was 23.8 GHz. Through configuration, the samples were transformed into a 3-m-long frame of 5 cm bins which was updated with a frequency of approximately 17 Hz. Time-of-flight (the time it takes for a signal to return to the receiver), was used to put the signals into the different bins, denoted by the distance (range) between the object and the radar. The radar can therefore detect multiple objects and separate them by their distance. This allows for precise measurements of behaviour of a specific person even in the presence of, eg multiple persons in the bed, a moving fan, or moving curtains. The received signals within each bin were analysed using the Doppler effect and Fast Fourier Transformation (FFT). Respiration rate was derived every second using FFT with a 20-s Hanning window, ie every rate had 19 s overlap with the previous rate. Movement was calculated as the change in the received signal over time and was divided in fast movements and slow movements analysing changes over 6 s and 20 s, respectively.

This radar technology is harmless to human beings as the high sampling rate and large bandwidth enables the use of waves that transmit less energy than tolerable background noise (<FCC Part 15

limit). The frequencies used allowed the pulses to travel through soft material such as bed sheets and clothes and only reflect on denser materials like the human body.

The sleep algorithm was mainly based on non-causal temporal neural networks like Temporal Convolutional Networks (TCN) and Long-Short-Term-Memory (LSTM) recurrent neural networks (RNN) that are fed with respiration and movement data from the radar. Non-causal means that the network can use information from the future if available, but maximum 2 h ahead. In other words, the algorithm will go back in time and update sleep stages if necessary. The network was trained to reduce the categorical cross entropy of the sleep stages W, N1, N2, N3 and R when compared to the PSG nights, ie sleep specialists acted as the supervisor in supervised learning. When the sleep stages were classified, they passed a state transition filter before N1 and N2 were merged to N1/N2 (light sleep). Somnofy and PSG were synchronized in time by maximizing the cross-correlation between the movement from Somnofy and movements from PSG. The output of the algorithm comprises sleep stages classified in 30-s epochs to mirror standard PSG scorings.

2.5. Validation technique

The sleep algorithm was validated using a k-fold cross validation technique. This technique enables validation of machine learning models on the same data set it was trained on, as a neural network can "remember" a night it has seen before. The total data set of 94 PSG nights were split in k-groups. One group was taken out of the data set, and the algorithm was trained on the remaining $k - 1$ groups. After being trained, the algorithm was validated on the group that was originally taken out. This process was performed k times until all the groups had been validated. In the present study k was set to 25, resulting in 3–4 PSG recordings per group.

To assure that the results were generalizable (to prevent overfitting), the following measures were taken: 1) All PSG nights were scored by two sleep specialists. Only scorings from Europe_1, Europe_2 and Europe_3 were used for training the algorithm (*PSGTrain*), while Europe_4 and USA_1 were used for validation (*PSGValidate*). 2) For each bedroom, one sensor location (nightstand or wall) was used for training the algorithm and the remaining sensor location was used for validation. The sensor locations were picked randomly for each bedroom for each fold in the cross validation, assuring that the sleeping environment, from Somnofy's point of view, was unseen at the validation.

2.6. Statistical analysis

For each nightly recording, the precision of Somnofy in terms of quantifying sleep compared to PSG was calculated for parameters central to clinical sleep medicine and sleep research. These parameters were: time in bed (TIB; minutes from lights-out to lights-on), total sleep time (TST; minutes asleep within TIB), sleep onset latency (SOL; minutes from lights-out to the first epoch of any sleep stage), R latency (minutes from SOL and to the first epoch of R), wake after sleep onset (WASO; minutes of wakefulness between sleep onset and final wake up), sleep efficiency (SE expressed in percentage; TST/TIB * 100), total wake time (TWT; minutes awake within TIB), and time spent in each stage of sleep (minutes in N1/N2, N3 and R).

Sleep staging across trained PSG-scorers does not necessarily agree, both due to the interpretation of scoring rules set by AASM, the quality of the signals and any pathology during sleep, which may complicate the defining of a specific sleep stage [10,11]. Hence, in addition to comparing Somnofy to PSG, Somnofy was also compared to the inter-rater variability of PSG scorers. Here, '*Somnofy*' will be used as agreement between Somnofy and *PSGValidate* (Europe_4 and USA_1), and '*PSG*' will be used for the agreement between *PSGTrain* (Europe_1, Europe_2 and Europe_3) and *PSGValidate*. The agreement on the quantitative sleep parameters was assessed by the intra-class correlation (ICC) parameter [12], which was calculated with the one-way random effects model [13] using the ANOVA module in the NAG Numerical Library (website: https://www.nag.com/numeric/py/nagdoc_latest/readme.html). ICC values less than 0.50 indicate poor reliability, values between 0.50 and 0.75 indicate moderate reliability, values between 0.75 and 0.90 indicate good reliability, whereas values greater than 0.90 indicate excellent reliability [13].

Further, the mean absolute disagreement (MAD) was calculated to estimate the expected disagreement, and standard deviation (SD) was calculated to estimate the expected variance of the disagreement. Subsequently, Bland-Altman plots [14,15] were made in order to investigate if Somnofy had any tendency to underestimate or overestimate any given sleep parameter. The mean difference (or bias), and lower and upper agreements limits (mean difference ± 1.96 * SD) were calculated. Biases were tested against zero for significance.

Finally, scorings obtained from each epoch by *Somnofy* and *PSGValidate*, and *PSGTrain* and *PSGValidate* were cross tabulated and the degree of agreement between them was quantified by means of the Cohen's kappa coefficient, as well as sensitivity (accuracy for detecting sleep), specificity (accuracy for detecting wake), and accuracy for classifying the individual sleep stages (N1/N2, N3, R and W). Cohen's kappa higher than 0.80 is considered to reflect almost perfect agreement, 0.80 to 0.61 substantial agreement, 0.60 to 0.41 moderate agreement, 0.40 to 0.21 fair agreement, 0.20 to 0.11 slight agreement, and values less than 0.10 are considered to reflect no agreement [12,15].

3. Results

3.1. Quantitative sleep parameters

Table 1 shows the mean, standard deviation and range of the quantitative sleep parameters for the 71 PSG recordings. Note that SOL, WASO and TWT were slightly higher than normal due to warm nights disrupting the sleep of some recordings.

The quantitative assessment of the agreement between PSG scorers (*PSGTrain* versus *PSGValidate*), and between Somnofy and one PSG scorer (*Somnofy* versus *PSGValidate*) for the different sleep

Table 1
Quantitative sleep parameters as calculated by PSG for the 71 nights used in the study.

	Mean (SD)	Range (min, max)
TST (min)	405 (55)	(190, 516)
SOL (min)	21 (17)	(3, 78)
R Latency (min)	96 (44)	(42, 248)
WASO (min)	37 (30)	(5, 142)
SE (%)	88 (8)	(58, 97)
TWT (min)	58 (36)	(12, 181)
N1/N2 (min)	233 (45)	(139, 326)
N3 (min)	85 (28)	(37, 159)
R (min)	87 (28)	(13, 141)
AHI (#/hour)	0.9 (1.1)	(0.0, 4.9)
PLMI (#/hour)	1.5 (2.4)	(0.0, 13.8)
ArI (#/hour)	7.3 (4.1)	(0.0, 23.5)

N1/N2, N3 and R represent the time in minutes spent in the corresponding sleep stages. TST = Total Sleep Time, SOL = Sleep Onset Latency, WASO = Wake After Sleep Onset, SE = Sleep Efficiency, TWT = Total Wake Time, AHI = apnoeas and hypopneas per hour of sleep, PLMI = periodic limb movements per hour of sleep, ArI = arousals per hour of sleep, SD = standard deviation.

parameters are illustrated in Table 2. The ICC coefficients indicate that the interrater agreement between the PSG scorers was moderate to excellent, while Somnofy varied from excellent to poor agreement compared to PSG defined sleep. For all the nine variables presented, the interrater agreement was higher for the PSG scorers than for Somnofy. The difference was smallest for stage N1/N2 and TST, for which Somnofy was about as reliable as PSG (ICC difference smaller than 0.10). Somnofy was slightly less reliable than PSG for SE and TWT (ICC difference between 0.10 and 0.20), and substantially lower than PSG for measures of SOL, R latency, WASO, N3 and R (ICC difference larger than 0.20).

The expected disagreement measured by MAD and the expected variance of the disagreement measured by SD showed that for N1/N2, Somnofy was almost equal to PSG. However, the expected disagreement for TST, R latency, SE, TWT, N3 and R was about twice as high for Somnofy compared to PSG. For SOL and WASO the expected disagreement of Somnofy was about three times higher than PSG.

The combined box and swarm plot in Fig. 2 displays the disagreement between Somnofy and PSG. For all sleep parameters, the pattern in terms of disagreement for Somnofy and PSG was similar. However, Somnofy had more outliers, which is consistent with the high standard deviations in Table 2. For R latency, the disagreements were either small if there were disagreements regarding when the first R cycle started, or they were large if the disagreements concerned whether the “first” R episode was present. Both PSG and Somnofy had both type of disagreements, but Somnofy had more of the large disagreements.

The Bland–Altman plots in Fig. 3 display the evaluation of the limits of agreement between Somnofy and the average of PSGValidate and PSGTrain per night. For SOL, Somnofy showed an average difference approximating zero and most of the points diverged little from this average. However, Somnofy may have difficulties in extreme cases, as suggested by the high mean value of SOL (>80 min).

The slope of the regression line was almost flat for SOL and SE (absolute value of the slope times the range less than 0.1 SD). For TST, R latency and time spent in R the slope was positive, indicating that Somnofy tended to overestimate more the higher the value. On the other hand, for WASO, TWT, N1/N2 and N3 Somnofy tended to underestimate more the higher the value.

3.2. Epoch-by-epoch agreement analysis

Epoch-by-epoch (EBE) analysis was performed for both PSG and Somnofy. Fig. 4a and b shows the obtained confusion matrices of the EBE agreement for the sleep stages N1/N2, N3, R and W. The average Cohen's kappa coefficient was 0.63 (SD = 0.10) for Somnofy, indicating a substantial agreement with PSG, and 0.82 (SD = 0.10) for PSG indicating almost perfect agreement (Table 3). For Somnofy the agreement ranged between 0.72 and 0.78 whilst PSG scorers ranged between 0.83 and 0.96. The lowest agreement was obtained for N1/N2, both for Somnofy and for PSG scorers.

Finally, Table 3 shows the accuracy, sensitivity and specificity for PSG and Somnofy. Somnofy had substantial agreement on all three parameters; accuracy: 0.76 (SD = 0.07), sensitivity: 0.97 (SD = 0.03) and specificity: 0.72 (SD = 0.19). A representative hypnogram, with kappa = 0.62, accuracy = 0.74, sensitivity = 0.95, and specificity = 0.79 is presented in Fig. 5.

Table 2
Quantitative analysis of PSG vs Somnofy for the quantitative sleep parameters.

	PSG		Somnofy	
	ICC (95% CI)	MAD (SD)	ICC (95% CI)	MAD (SD)
TST	0.98 (0.97, 0.99)	7.91 (10.83)	0.94 (0.90, 0.96)	14.77 (20.01)
SOL	0.92 (0.87, 0.95)	3.32 (6.73)	0.38 (0.16, 0.56)	9.72 (21.67)
R Latency	0.59 (0.41, 0.72)	19.25 (45.10)	0.28 (0.05, 0.48)	39.96 (62.93)
WASO	0.94 (0.91, 0.96)	6.65 (10.13)	0.68 (0.54, 0.79)	15.96 (22.78)
SE (%)	0.95 (0.91, 0.97)	1.75 (2.43)	0.84 (0.75, 0.89)	3.14 (4.20)
TWT	0.95 (0.92, 0.97)	7.91 (10.83)	0.83 (0.74, 0.89)	14.77 (20.01)
N1/N2	0.62 (0.44, 0.74)	33.71 (33.08)	0.59 (0.42, 0.72)	35.09 (40.89)
N3	0.57 (0.38, 0.70)	21.96 (26.26)	0.08 (−0.15, 0.31)	34.11 (40.50)
R	0.78 (0.66, 0.85)	13.62 (16.39)	0.50 (0.30, 0.65)	24.73 (30.24)

Intraclass correlation coefficient (ICC) with 95% confidence interval (CI) and the mean absolute disagreement (MAD) in minutes (% for sleep efficiency) for PSG (PSGTrain vs PSGValidate) and Somnofy (Somnofy vs PSGValidate) with corresponding standard deviation (SD). N1/N2, N3 and R represent the time in minutes spent in the corresponding sleep stages. TST = Total Sleep Time, SOL = Sleep Onset Latency, WASO = Wake After Sleep Onset, SE = Sleep Efficiency, TWT = Total Wake Time.

3.3. Other analyses

The performance of Somnofy in relation to data collected in a home environment or in a sleep clinic did not differ as the Cohen's kappa was close to identical (home environment: 0.62, SD = 0.11, n = 34; and sleep clinic: 0.61, SD = 0.10, n = 37). Also, the position of Somnofy in the room did not differ in terms of sleep stage detection (nightstand: 0.61, SD = 0.10, n = 34; and mounted to the wall: 0.62, SD = 0.10, n = 37). The body position was measured by the PSG at the sleep clinic and showed only minor influence on Somnofy's accuracy (left side position: 0.79, n = 6173 epochs; right side position: 0.72, n = 5234; prone position: 0.75, n = 1362; and in supine position: 0.76, n = 13936). The results for the nightstand units (left side position: 0.76, n = 3182 epochs; right side position: 0.71, n = 3158; prone position: 0.79, n = 825; and in supine position: 0.75, n = 8167) and the wall units (left side position: 0.81, n = 2991; right side position: 0.75, n = 2076; prone position: 0.70, n = 537; and in supine position: 0.77, n = 5769) separately did not show significantly lower precision for the positions where the chest was facing away from the sensor, which was the right side position for the nightstand units and prone position for the wall units.

The accuracy and Cohen's kappa were independent of gender with 0.76 (SD = 0.07) and 0.62 (SD = 0.10) for females and 0.75 (SD = 0.07) and 0.61 (SD = 0.11) for males, respectively. The present data set was too limited to conclude on significance in specific clinical population groups. However, our preliminary results showed that for the twenty-three nights excluded from the validation study due to indication of sleep disorders (PLMD, sleep apnoea or narcolepsy, n = 23), the Cohen's kappa was 0.53 (SD = 0.11), accuracy was 0.71 (SD = 0.08), sensitivity was 0.92 (SD = 0.10) and specificity was 0.69 (SD = 0.19). Compared to the results for the healthy population kappa, accuracy, sensitivity and specificity decreased with −0.10, −0.05, −0.05 and −0.03, respectively.

All nights were scored by two groups of sleep specialists, PSGTrain and PSGValidate. Table 4 displays epoch-by-epoch interscorer variability between all pairs of scorers present in the study. For our sample of scorers, USA_1 disagreed with Europe_3 more than any other pair. The difference stemmed mostly from scoring of N1/N2 and N3, where the American scorer tended to score less N3 (nightly average of 53 min vs. 87 min) and more N1/N2 (nightly average of 280 min vs. 226 min) than the European scorer. The ICC for PSG for N1/N2 and N3 in Table 2 showed moderate reliability, but if USA_1 was excluded from the calculations the ICC would be 0.85 and 0.83 respectively (ie good reliability). Somnofy was validated against USA_1 and Europe_4, and the Cohen's kappa agreement was 0.60 and 0.62, respectively.

4. Discussion

The present study demonstrates the ability of Somnofy to estimate sleep and wake in a healthy population. Compared to

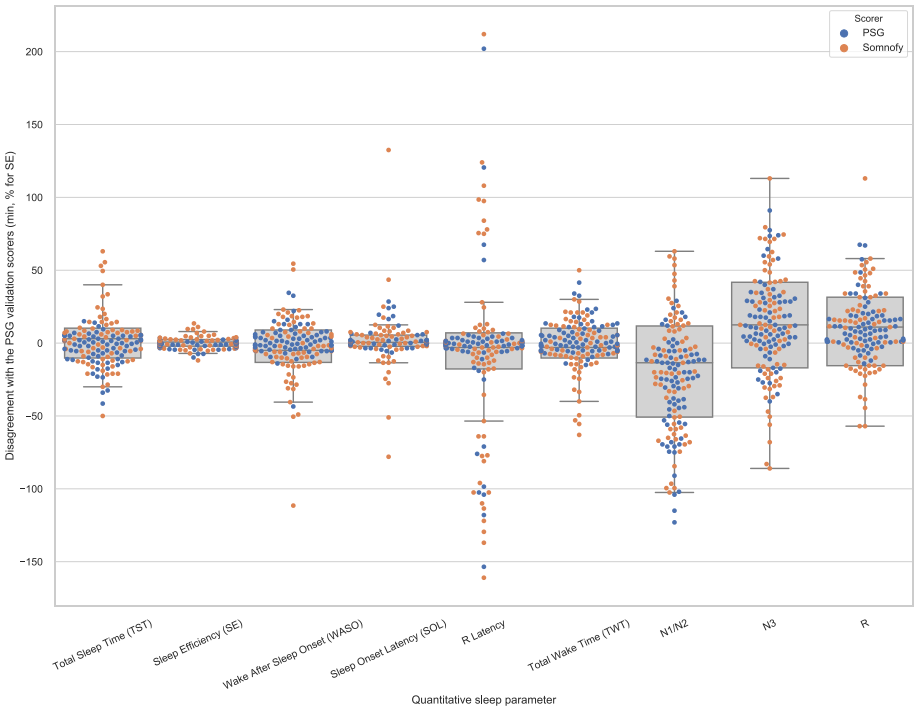

Fig. 2. The boxes refer to the first and third quartiles of the disagreement between Somnofy and PSGValidate. Each orange dot represents the agreement for one night between Somnofy and PSGValidate, while each blue dot represent the difference between PSGTrain and PSGValidate for the same nights. The y-axis indicates the difference in minutes (% for sleep efficiency). All numbers are in minutes, except for SE which is given in %.

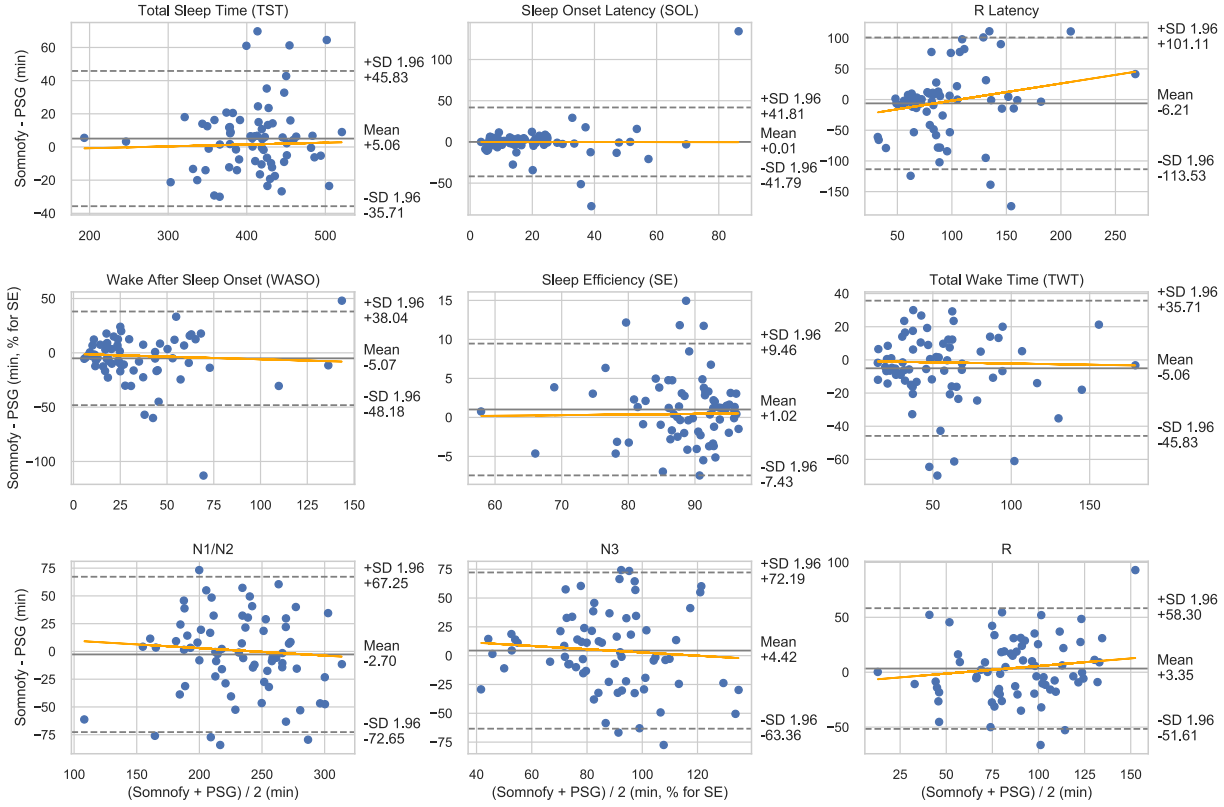

Fig. 3. Disagreement between Somnofy and PSG (average of PSGTrain and PSGValidate) for each night is plottet on the y-axis against the average of Somnofy and PSG on the x-axis. The solid black line indicates the bias for Somnofy and the dashed lines the Bland–Altman limits of disagreement (+/−1.96 SD). The orange regression line is added to show any trend in the distribution of errors and is calculated neglecting points outside the limits of disagreement.

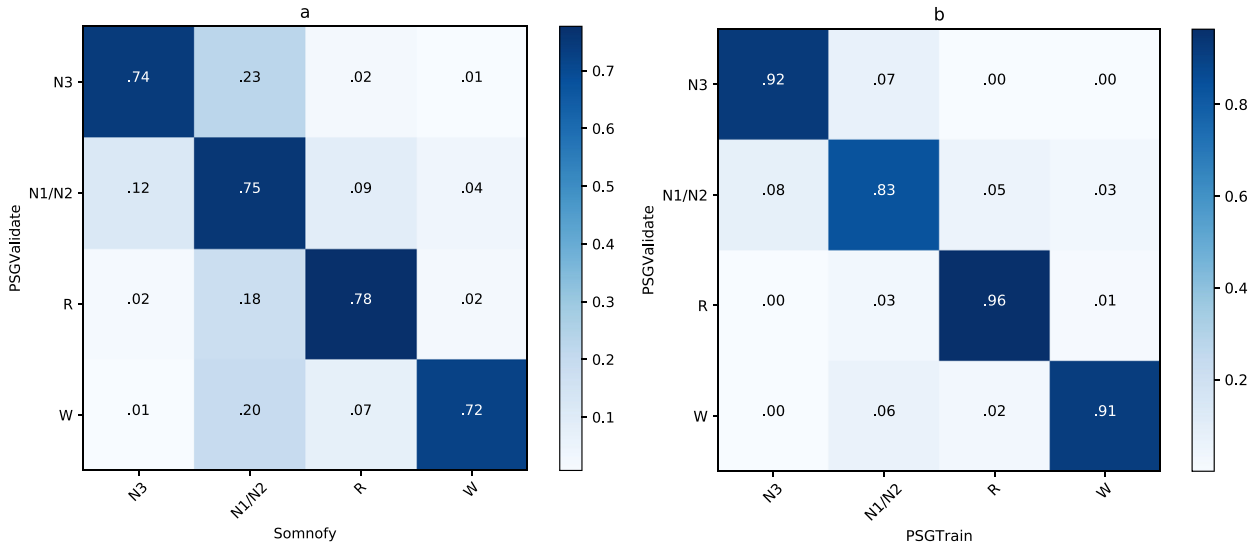

Fig. 4. Distribution of epoch-by-epoch agreement of sleep state classification between PSGValidate on the y-axis and Somnofy (a) or PSGTrain (b) on the x-axis. Numbers are normalized such that each row sums up to one. Epochs agreed on are found on the diagonal.

Table 3
Epoch-by-epoch analyses of PSG vs Somnofy.

	PSG	Somnofy
	Mean (SD)	Mean (SD)
Cohen's kappa	0.82 (0.10)	0.63 (0.10)
Accuracy	0.88 (0.06)	0.76 (0.07)
Sensitivity	0.99 (0.02)	0.97 (0.03)
Specificity	0.85 (0.11)	0.72 (0.19)

Comparison of Cohen's kappa (for W, N1/N2, N3, R classification), accuracy (for W, N1/N2, N3, R classification), sensitivity (accuracy for detecting sleep) and specificity (accuracy for detecting wake) for PSG (PSGTrain vs PSGValidate) and Somnofy (Somnofy vs PSGValidate). SD = standard deviation.

manually scored PSG, Somnofy scored sleep/wake robustly with 0.97 of true sleep epochs scored correctly, and 0.72 of true wake epochs scored correctly. The sensitivity and specificity are like that of simplified PSG solutions that only use frontopolar EEG [16]. The challenge for non-intrusive sleep tracking devices is to reliably detect wakefulness. The specificity found in the present study was higher than that reported for actigraphy (0.34–0.65) [17–19]. To our knowledge, Somnofy shows the highest specificity compared to other non-EEG systems, including other contactless monitoring devices that use technology based on passive infrared, sonography, or pressure sensation [17–23]. Validation of the radar based Resmed S+ showed similar specificity, but their data set contained almost twice the amount of wake as ours, most likely making it easier to correctly classify wake [24].

The utility of the radar and the temporal neural network sleep scoring introduced here are illustrated by the ability to detect time spent in the different sleep stages and sleep timing parameters. Epoch-by-epoch comparisons showed that Somnofy accurately detected N1/N2 in 0.75, N3 in 0.74 and R in 0.78 of the epochs compared to PSG, with an average absolute disagreement of 1, 12 and 11 min more than between manual PSG scores, respectively. Such disagreements should be tolerable considering the average total amount of N1/N2 (233 min), N3 (85 min) and R (87 min) detected by PSG. PSG N3 and PSG R were mostly misclassified by Somnofy as N1/N2 with 0.23 and 0.18 of the epochs, respectively. Cohen's kappa for PSG was significantly larger than for Somnofy (0.82 versus 0.63, respectively), for which much of the difference was due to more precise timing of state transitions for PSG. Distinguishing between N1/N2, N3 and R with non-EEG based systems has been and still is challenging. The sleep stage differentiation of

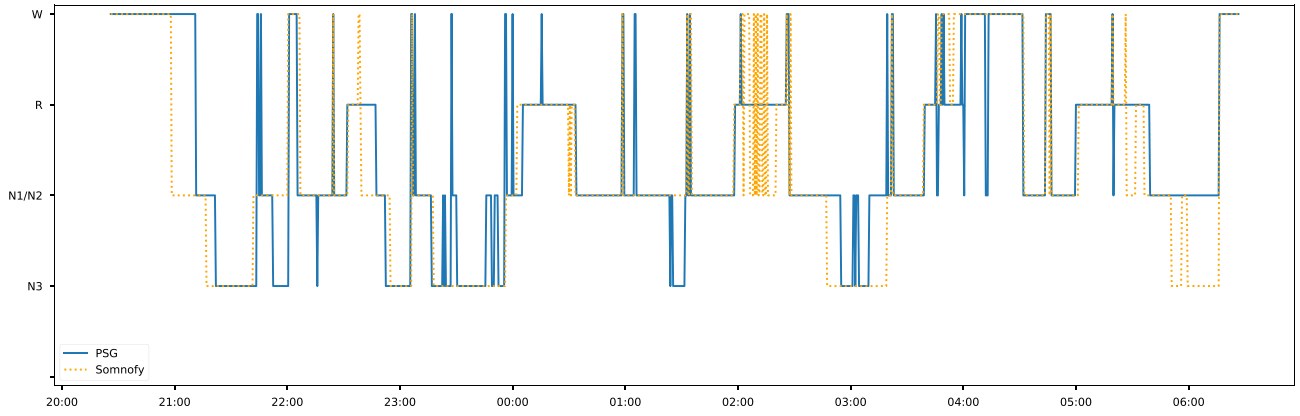

Fig. 5. Example of an hypnogram from Somnofy and PSG for a night with approximately average score on Cohen's kappa for W, N1/N2, N3, R classification (0.62), accuracy for W, N1/N2, N3, R classification (0.74), sensitivity (0.95) and specificity (0.79).

Table 4
PSG inter-scorer reliability.

	Europe_1–Europe_4	Europe_2–Europe_4	Europe_3–Europe_4	Europe_3–USA_1
Count (%)	19 (27%)	15 (21%)	7 (10%)	30 (42%)
Kappa	0.88	0.85	0.82	0.75
Accuracy	0.92	0.90	0.88	0.84
Kappa W	0.89	0.84	0.88	0.83
Kappa N1/N2	0.85	0.81	0.77	0.70
Kappa N3	0.90	0.87	0.77	0.68
Kappa R	0.91	0.89	0.87	0.82

Measures of PSG inter-scorer variability for the paired scorer combinations in the study. Kappa is short for Cohen's kappa for W, N1/N2, N3 and R classification. Europe_1, Europe_2 and Europe_3 was used to train the sleep algorithm, and Europe_4 and USA_1 was used to validate the algorithm.

Somnofy was more precise than that of other non-EEG technology providing sleep stages as Fitbit Charge 2 [22], Oura ring [21] and Resmed S+ [24]. Simplified PSG with only frontopolar EEG showed higher accuracy than Somnofy [16], but the less intrusive alternative, in ear-EEG, was worse [25].

An important factor evaluating sleep detecting technology is performance on quantitative sleep parameters. The Bland–Altman plots revealed that Somnofy was consistent with PSG on TST, TWT, WASO, SE and SOL, except from extreme cases with long SOL, short TST or low SE. Although the Bland–Altman intervals of agreement were quite wide in the present study, the mean differences between Somnofy and PSG were low, indicating little bias. The expected absolute disagreement per night (MAD) between Somnofy and PSG was 8 min more for TST, 9 min more for WASO, 8 more minutes for TWT and 7 min more for SOL than the expected disagreement between two manual PSG scorers. For TST, which averaged 405 min per night, 8 min is negligible, while for WASO (37 min per night), TWT (58 min per night) and SOL (21 min per night) the disagreements are substantial. Similar Bland–Altman plots have however been found for other technologies [21,22].

Somnofy seems to handle different sensor locations and sleeping environments well. There was no significant difference in performance when the unit was placed in a home environment or in a sleep clinic, nor if the unit was placed on a nightstand or mounted to the wall. Neither the sleeping position seemed to matter in terms of validity. The accuracy was also consistent across genders.

The results show that while PSG remains the reference method for sleep scoring, Somnofy showed high precision in an automated and non-invasive way. Sleep analysis with Somnofy is less rich in content than that of PSG, as no brain wave morphology like spindles and K-complexes are detected. Despite this fact, the hypnograms from Somnofy provide good reliability of the night's sleep quality. This could make Somnofy an adequate alternative to PSG for longitudinal studies on healthy adults as the cost, scalability and user simplicity should be superior to PSG. Increased access to accurate longitudinal studies could enhance sleep research by uncovering new correlations and understandings about sleep and sleep dependent physical and mental performance.

4.1. Limitations

This study was limited to a healthy population of mostly young adults. Further studies are required in order to validate Somnofy for elderly people that move more during sleep, and for populations with different (sleep)disorders, including sleep related breathing disorders and movement disorders.

Moreover, lights-out/lights-on was indicated by the participants. Somnofy's own algorithm for detecting these markers were not investigated. Further, the study only investigated full nights of sleep; data on power naps were not investigated. The validated hypnograms were generated by Somnofy after final wake-up, in the same way as manual PSG is scored in hindsight. Somnofy can also do real-time sleep classification during the night, but this was not validated in this study. Furthermore, the participants in the present study slept alone. Somnofy can differentiate between two subjects sharing the bed by setting the distance parameter in Somnofy to a distance between them. In this study the distance parameter was set to 3 m.

4.2. Future research

Although, yet to be investigated, we reason that Somnofy has large potential for clinical utilization. While the present study mainly included healthy adults, twenty-three participants showed indications of PLMD, sleep apnoea or narcolepsy. For these nights, Cohen's kappa and specificity were only reduced by 0.10 and 0.03, respectively. These preliminary results must be further validated in much larger populations as sleep consolidation and movements will vary in accordance with age and clinical status. Nevertheless, the results are promising and if neural networks are trained on more cases of sleep disorders, we hypothesize a better performance. Furthermore, information on sleep stages combined with movement and respiration data from the radar, have the potential to be used for development of algorithms that can be validated as a screening tool for specific sleep disorders.

4.3. Conclusions

The present study shows that Somnofy, using radar technology and machine learning, can provide information not only about sleep and wakefulness, but also about sleep stages. The study demonstrated that Somnofy can classify sleep stages with substantial agreement against PSG for healthy young adults, making it promising for epidemiological sleep research on this population. Further validation studies are needed in order to conclude about the precision of this device in clinical settings, and across different age groups.

Disclosure statement

All authors have seen and approved the current version of the manuscript. The data collection was partly financed by VitalThings. Ståle Toften works for VitalThings. The other authors report no conflict of interest.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Statement of significance

Polysomnography (PSG) is the gold standard for objective sleep measurements. Unfortunately, PSG is both intrusive and costly

which make it impractical for longitudinal studies. The present study validates non-intrusive radar technology and machine learning (Somnofy®) against PSG. The results show that Somnofy can provide automatic sleep stage classification with a precision close to PSG in a sample of healthy, mostly young subjects. This type of technology can open a wide range of opportunities for epidemiological sleep research.

CRediT authorship contribution statement

Ståle Toften: Methodology, Formal analysis, Resources, Writing - original draft, Visualization, Funding acquisition. **Ståle Pallesen:** Conceptualization, Methodology, Formal analysis, Resources, Writing - review & editing, Funding acquisition. **Maria Hrozanova:** Writing - review & editing. **Frode Moen:** Writing - review & editing. **Janne Grønli:** Conceptualization, Methodology, Formal analysis, Resources, Writing - original draft, Supervision, Project administration, Funding acquisition.

Acknowledgements

We wish to thank Tony Wader and Nikita Zhitniy at the Colosseum Clinic in Oslo and Ingvild Saxvig, Torhild Pedersen, Irina Oltu, and Jelena Mrdalj at the Faculty of Psychology at University of Bergen for their contribution with PSG hook up and scoring.

We also want to thank Ask Krogstad, data scientist, for his help in developing the sleep algorithm, and Hanne Siri Heglum, PhD candidate, for her contribution with the radar apparatus.

Last, we want to acknowledge VitalThings' contribution and support throughout the study.

Conflicts of interest

The ICMJE Uniform Disclosure Form for Potential Conflicts of Interest associated with this article can be viewed by clicking on the following link: https://doi.org/10.1016/j.sleep.2020.02.022.

References

[1] Mostov K, Liptsen E, Boutchko R. Medical applications of shortwave FM radar: remote monitoring of cardiac and respiratory motion. Med Phys 2010;37(3):1332–8.
[2] Staderini EM. UWB radars in medicine. IEEE Aero El Sys Mag 2002;17(1):13–8.
[3] Yim D, Lee WH, Kim JI, et al. Quantified activity measurement for medical use in movement disorders through IR-UWB radar sensor. Sensors-Basel 2019;19(3).
[4] Sun Kang YL, Young-Hyo Lim, Hyun-Kyung Park, et al. Validation of noncontact cardiorespiratory monitoring using impulse-radio ultra-wideband radar against nocturnal polysomnography. Sleep Breath 2019.
[5] Stefani A, Gabelia D, Mitterling T, et al. A prospective video-polysomnographic analysis of movements during physiological sleep in 100 healthy sleepers. Sleep 2015;38(9):1479–87.
[6] Chung GS, Choi BH, Lee J-S, et al. REM sleep estimation only using respiratory dynamics. Physiol Meas 2009;30(12):1327–40.
[7] Pallesen S, Grønli J, Myhre K, et al. A pilot study of impulse radio ultra wideband radar technology as a new tool for sleep assessment. J Clin Sleep Med 2018;14(7):1249–54.
[8] Nakazaki K, Kitamura S, Motomura Y, et al. Validity of an algorithm for determining sleep/wake states using a new actigraph. J Physiol Anthropol 2014;33(1):31.
[9] Berry RB. The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications. Am Acad Sleep Med 2018.
[10] Collop NA. Scoring variability between polysomnography technologists in different sleep laboratories. Sleep Med 2002;3(1):43–7.
[11] Drinnan MJ, Murray A, Griffiths CJ, et al. Interobserver variability in recognizing arousal in respiratory sleep disorders. Am J Resp Crit Care 1998;158(2):358–62.
[12] Ranganathan P, Pramesh CS, Aggarwal R. Common pitfalls in statistical analysis: measures of agreement. Perspect Clin Res 2017;8(4):187–91.
[13] Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016;15(2):155–63.
[14] Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;327(8476):307–10.
[15] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159–74.
[16] Levendowski DJ, Ferini-Strambi L, Gamaldo C, et al. The accuracy, night-to-night variability, and stability of frontopolar sleep electroencephalography biomarkers. J Clin Sleep Med 2017;13(6):791–803.
[17] Paquet J, Kawinska A, Carrier J. Wake detection capacity of actigraphy during sleep. Sleep 2007;30(10):1362–9.
[18] Marino M, Li Y, Rueschman MN, et al. Measuring sleep: accuracy, sensitivity, and specificity of wrist actigraphy compared to polysomnography. Sleep 2013;36(11):1747–55.
[19] de Souza L, Benedito-Silva AA, Pires MLN, et al. Further validation of actigraphy for sleep studies. Sleep 2003;26(1):81–5.
[20] Bhat S, Ferraris A, Gupta D, et al. Is there a clinical role for smartphone sleep apps? Comparison of sleep cycle detection by a smartphone application to polysomnography. J Clin Sleep Med 2015;11(7):709–15.
[21] de Zambotti M, Rosas L, Colrain IM, et al. The sleep of the ring: comparison of the OURA sleep tracker against polysomnography. Behav Sleep Med 2019;17(2):124–36.
[22] de Zambotti M, Goldstone A, Claudatos S, et al. A validation study of Fitbit charge 2 (TM) compared with polysomnography in adults. Chronobiol Int 2018;35(4):465–76.
[23] Walch O, Huang Y, Forger D, et al. Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device. Sleep 2019;42(12).
[24] Schade MM, Bauer CE, Murray BR, et al. Sleep validity of a non-contact bedside movement and respiration-sensing device. J Clin Sleep Med 2019;15(7):1051–61.
[25] Mikkelsen KB, Villadsen DB, Otto M, et al. Automatic sleep staging using ear-EEG. Biomed Eng Online 2017;16(1):111.