Nature and Science of Sleep

Dovepress
open access to scientific and medical research

Open Access Full Text Article

CLINICAL TRIAL REPORT

Performance Investigation of Somfit Sleep Staging Algorithm

Marcus McMahon[1], Jeremy Goldin[2], Elizabeth Susan Kealy[3], Darrel Joseph Wicks[4], Eugene Zilberg[5], Warwick Freeman[5], Behzad Aliahmad[5]

[1]Department of Respiratory and Sleep Medicine, Epworth Hospital, Richmond, Victoria, Australia and Department of Respiratory and Sleep Medicine, Austin Health, Heidelberg, Victoria, Australia; [2]Department of Respiratory and Sleep Medicine, Royal Melbourne Hospital, Parkvile, Victoria, Australia; [3]Sleepmetrics Pty Ltd, Heidelberg, Victoria, Australia; [4]Sleep Disorders Unit, Epworth Hospital, Richmond, Victoria, Australia; [5]Medical Innovations, Compumedics Limited, Abbotsford, Victoria, Australia

Correspondence: Eugene Zilberg, Compumedics Limited, 30-40 Flockhart Street, Abbotsford, Victoria, 3067, Australia, Tel +61 412225842, Fax +61 3 84207399, Email ezilberg@compumedics.com.au

Purpose: To investigate accuracy of the sleep staging algorithm in a new miniaturized home sleep monitoring device – Compumedics® Somfit. Somfit is attached to patient's forehead and combines channels specified for a pulse arterial tonometry (PAT)-based home sleep apnea testing (HSAT) device with the neurological signals. Somfit sleep staging deep learning algorithm is based on convolutional neural network architecture.

Patients and Methods: One hundred and ten participants referred for sleep investigation with suspected or preexisting obstructive sleep apnea (OSA) in need of a review were enrolled into the study involving simultaneous recording of full overnight polysomnography (PSG) and Somfit data. The recordings were conducted at three centers in Australia. The reported statistics include standard measures of agreement between Somfit automatic hypnogram and consensus PSG hypnogram.

Results: Overall percent agreement across five sleep stages (N1, N2, N3, REM, and wake) between Somfit automatic and consensus PSG hypnograms was 76.14 (SE: 0.79). The percent agreements between different pairs of sleep technologists' PSG hypnograms varied from 74.36 (1.93) to 85.50 (0.64), with interscorer agreement being greater for scorers from the same sleep laboratory. The estimate of kappa between Somfit and consensus PSG was 0.672 (0.002). Percent agreement for sleep/wake discrimination was 89.30 (0.37). The accuracy of Somfit sleep staging algorithm varied with increasing OSA severity – percent agreement was 79.67 (1.87) for the normal subjects, 77.38 (1.06) for mild OSA, 74.83 (1.79) for moderate OSA and 72.93 (1.68) for severe OSA.

Conclusion: Agreement between Somfit and PSG hypnograms was non-inferior to PSG interscorer agreement for a number of scorers, thus confirming acceptability of electrode placement at the center of the forehead. The directions for algorithm improvement include additional arousal detection, integration of motion and oximetry signals and separate inference models for individual sleep stages.

Keywords: home sleep apnea testing, polysomnography, forehead electroencephalography, deep learning, interscorer agreement

Introduction

Polysomnography (PSG) is a standard investigation to analyze human sleep and diagnose sleep disorders. It involves full night recording of multiple scalp electroencephalographic (EEG) channels, electromyogram (EMG), electrooculogram (EOG), single channel electrocardiogram (ECG), limb movement signals, nasal airflow, chest and abdominal movement signals, oximetry, snoring and body position signals.[1] Subsequently, PSG data are reviewed and associated events are scored by a sleep technologist in a format appropriate for diagnosis of sleep disorders. As sleep staging is the major component of PSG data analysis, the guidelines of American Academy of Sleep Medicine (AASM) recommend continuous recording of the minimum of three EEG channels (frontal, central and occipital) referenced to the mastoid process as well as right and left EOG channels and chin EMG.[1] Given its complexity, PSG investigation has been recommended to be performed in the sleep laboratory in attendance of a sleep technologist.[2]

Received: 9 April 2024
Accepted: 1 July 2024
Published: 22 July 2024

© 2024 McMahon et al. This work is published and licensed by Dove Medical Press Limited. The full terms of this license are available at https://www.dovepress.com/terms.php and incorporate the Creative Commons Attribution – Non Commercial (unported, v3.0) License (http://creativecommons.org/licenses/by-nc/3.0/). By accessing the work you hereby accept the Terms. Non-commercial uses of the work are permitted without any further permission from Dove Medical Press Limited, provided the work is properly attributed. For permission for commercial use of this work, please see paragraphs 4.2 and 5 of our Terms (https://www.dovepress.com/terms.php).

Obstructive sleep apnea (OSA) is a condition of disordered breathing characterized by intermittent partial (hypopnea) and/or complete (apnea) upper airway obstruction during sleep. These apneic events lead to hypercapnia, hypoxemia, sympathetic stimulation and fragmentation of the sleep architecture by arousals, which often accompany these events. This subsequently leads to impairment of daytime function[3] and there is also evidence that repeated cortical arousals and hypoxemia contribute to increased mortality from cardiovascular disease in these patients.[4]

Given the high prevalence of OSA,[5] its heavy burden on population health and availability of effective treatments, as well as the excessive complexity and cost of PSG, the unattended home sleep apnea testing (HSAT) has emerged as a simpler and cheaper diagnostic option. The recommended HSAT configurations[1] typically do not include the channels required for sleep staging and constitute the type 3 diagnostic sleep studies, recommended for the initial diagnosis and longitudinal management of OSA.[5–8] The type 3 devices include those based on pulse arterial tonometry (PAT),[9,10] that may incorporate non-EEG based sleep staging algorithms. Misclassification of OSA was reported for these devices,[11,12] which can be at least partially attributed to the limitations of their sleep staging algorithms.[13,14] It was shown that type 3 studies may underestimate the presence and/or severity of OSA due to inability to correctly measure total sleep time (TST) and record arousals[15,16] with the latter factor also responsible for not reporting the respiratory event related arousals (RERA) and subsequently further underestimating respiratory disturbance index (RDI). Lack or inferior accuracy of sleep staging in type 3 HSAT does not allow to perform proper phenotyping of OSA specifically of the REM phenotype.[17]

Type 2 diagnostic studies are unattended home recordings that include at least one channel of EEG, EOG and EMG[5,7] and are therefore capable of reporting sleep stages and overcoming the deficiencies of type 3 HSAT devices. The Australasian recommendation[8] for the limited channel studies states that the central derivation C4-M1 should be used if only one EEG channel is available. This recommendation is based solely on the findings[18] of only a small shift in the sleep stages and no significant difference in the arousals when a single C4-M1 derivation was used instead of all three EEG channels. The major limiting factors in type 2 studies compared to type 3 studies are the difficulty of "wiring up" the patient which requires the sleep technician for the most reliable recording,[7] and a number of wires attached to the head that may prevent the patient from sleeping in their natural position and therefore compromise the correct assessment and phenotyping of OSA.[19] It would be important to investigate feasibility of accurate sleep staging based on the minimum number of electrodes attached to a small area on the forehead which would ensure simple and reliable self-application by the patient and the least invasiveness of overnight sleep.

While similarly to the approach used a prior study,[18] the early commercial single channel sleep staging algorithms utilized the central EEG derivations,[20] with the advent of HSAT the focus shifted to investigation of the role of the frontal EEG in sleep staging and evaluation of the automatic algorithms based on the frontal EEG. It was demonstrated[21] that kappa for overall sleep staging agreement between manual scoring from the single frontal channel (Fpz-M1) and from the full PSG was 0.72 and 0.73 for the two scorers, and epoch by epoch sleep/wake percent agreement was 92% and 95% respectively. Investigation of Michele Sleep Scoring[22] found that its performance on single channel EEG was as accurate with the frontal as with the central EEG derivations. The performance of Sleep Profiler device utilizing Fp1-Fp2 for EEG was reported.[23–29] The Sleep Profiler algorithm uses a rule based combination of spectral analysis, feature extraction and decision tree. The reported values of kappa statistics for agreement between the Sleep Profiler algorithm and the parallel PSG manual scoring were 0.61, 0.62,[25] and 0.63[27] in the investigation groups including the OSA sufferers, and 0.76 and 0.74 for the healthy participants.[24] When the accuracy of manual staging from the Sleep Profiler frontal electrode and the expert adjusted algorithm accuracy were analyzed in the groups with suspected OSA,[27,28] the values of kappa were 0.67 in both studies, with the mean overall percent agreement at 73.9% for the latter case.[27] It was also shown[24] that the power values of submental and frontal EMG were similar during wakefulness and sleep, thus further supporting suitability of the frontal montages for sleep staging. Another automatic algorithm for F4-M1 derivation utilizing a combination of feature extraction based on spectral analysis and a decision tree based on neural network reported kappa of 0.62 and percent agreement of 75.5% for the four categories (with N1 and N2 merged into light sleep) in the investigation group comprising 60% of participants with OSA.[30]

In recent years further significant improvement in the accuracy of sleep staging algorithms was achieved with application of Deep Learning (DL) models and software tools.[31] Several publications[32–43] investigated application of

DL to development of single EEG channel sleep staging algorithms, typically using multilayer convolutional neural networks (CNN)[34,35,38,40,42] or different implementations of recurrent neural networks (RNN)[32,37] or combinations of CNN and RNN.[33,36,39,41] While most of these algorithms demonstrate very high percent agreement with manual PSG that significantly exceeds 80% and even 85% for some configurations[39] it must be taken into account that these results were mostly achieved for the public sleep databases which often only include the healthy participants and the artifact free signals. Importantly it is demonstrated[39] that for the preferred DL model (bidirectional LSTM with attention[31]) sleep staging accuracy is highest at 81.72% for the frontal derivation (Fp1-A1) compared to 81.62% and 77.09% for the central (Cz-A1) and occipital (O1-A1) derivations respectively. A combined CNN-RNN (with bidirectional LSTM) DL model[36] was trained on a clinical data set of 717 diagnostic patients with suspected OSA, and achieved the respective accuracies of 82.9% on the single F4-M1 EEG channel and 83.8% on a combination of F1-M1 EEG channel and E1-M2 EOG channel when tested on 87 recordings from the same data set.

The accuracy of any sleep staging algorithm when it is compared with the manual scoring should be interpreted only in the context of interscorer agreement. The level of interscorer agreement varies significantly depending on the study population and affiliation of the scorers.[25,43] Typically interscorer agreement decreases for OSA population compared to the healthy participants and also with increase in the apnea-hypopnea index (AHI), and increases for scorers from the same facility following common training procedures. This was demonstrated by the scorers from SHHS study[44] who achieved the values of percent agreement between 86.8 and 88.1% and kappa statistics in the range of 0.81–0.83 despite a significant AHI between 14 and 15 in the study group. For all other reported interrater agreement studies involving the OSA populations, the levels of agreement were significantly lower. The reported percent agreement for 38 subjects with sleep disordered breathing was 71%.[45] The percent agreement and kappa for 31 subjects with mean AHI of 22 were reported at 82% and 0.73 respectively.[46] The mean percent agreement in a study[47] with a large number of scorers was reported at 82.6% on a data set with a small proportion of OSA patients. Mean percent agreement for the three scorers from different institutions and the data set with mean AHI of 15.8 was 77%.[18] Mean percent agreement for the two scorers on a data set of 56 subjects with 50% OSA sufferers was 78.9%[48] A study of eight different sleep centres[49] reported mean percent agreement and kappa of 76.8% and 0.68 respectively on a data set of 198 recordings with a mixture of sleep disorders. Another study with five sleep technologists and a mixed data set of 63 participants (52% OSA) revealed the mean interscorer percent agreement of 75.9%.[27]

Compumedics has developed a new miniaturized sleep monitoring device – Somfit (Figure 1), which is attached to the patient's forehead and combines the channels specified for a PAT-based HSAT device (oximetry and PAT, plus oximetry derived pulse rate, snoring and accelerometry derived head positions, angles and motion) with the neurological signals representing EEG, EOG and EMG derivations. The first evaluation of Somfit is presented for the early rule based sleep staging algorithm.[50] Recently a new DL based sleep staging algorithm was developed for Somfit. The objective of this study was to evaluate the latest Somfit sleep staging algorithm in a target clinical population, which is a combination of the patients with suspected OSA and those with known OSA requiring ongoing management.

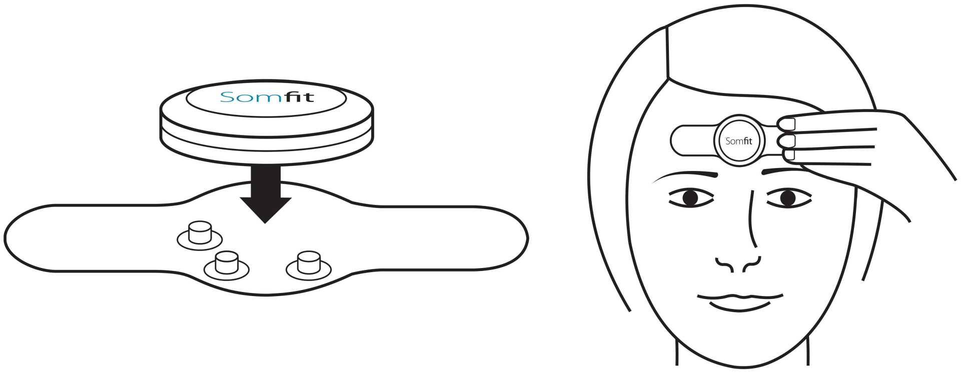

Figure 1 Somfit device.

Methods

Study Design

A total of 110 participants were enrolled in the study, which involved simultaneous recording of full overnight PSG and Somfit data. Eligible participants were patients who were referred for a sleep diagnostic study and satisfied at least one of the following criteria:

- Referral includes suspected sleep disordered breathing based on the judgement of a referring sleep physician;
- Patients have already been diagnosed with sleep disordered breathing and require a diagnostic study as part of their ongoing OSA management. Significant weight change of more than 10% was the main reason for a referral for a diagnostic study for such patients,[6] and ongoing fatigue despite good adherence to treatment being an additional reason.

Offers to participate in the study were made to the patients as they presented to the center in the order they presented. To confirm the high pretest probability of sleep apnea, the participants were asked to complete Epworth Sleepiness Scale (ESS) questionnaire.

The recordings were conducted at the three centers in Australia:

- Sleep Unit, Epworth Hospital, Camberwell Victoria – 40 subjects;
- SleepMetrics, Heidelberg Victoria – 35 subjects;
- Appleton Institute Central Queensland University, Wayville South Australia – 35 subjects.

This multicenter clinical trial was approved by HREC (Bellberry Limited Eastwood South Australia Australia), Protocol No. 2022-10-1133. The approval included the participant information sheet and consent form. According to this form informed consent was obtained from the study participants. The study was conducted in accordance with the guidelines outlined in the Declaration of Helsinki. The trial is also listed on ClinicalTrials.gov (ID NCT05647746).

Out of the 110 studies, four (two each in Centers 2 and 3) were excluded from analysis. The flowchart for the exclusion scenarios is presented in Figure 2.

Sleep Studies

The PSG recordings were conducted with Compumedics Grael system at Centers 1 and 2, and with Compumedics E-Series System at Center 3.

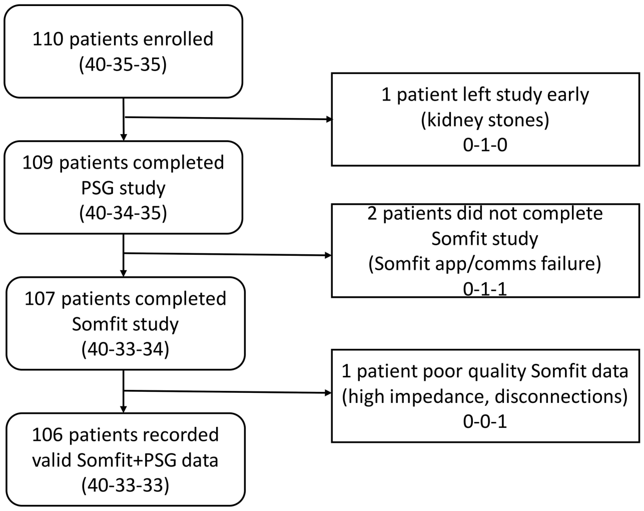

Figure 2 Study flow chart.

All PSG recordings were independently scored according to the AASM rules[1] by three qualified, experienced sleep scientists blinded to the subjects' identity and to the fact that the sleep studies were utilized in the trial. Each study at a given center was scored by the same three sleep scientists. The scorers at Center 1 were the sleep scientists continuously employed by the Epworth Sleep Units for a number of years, while the scorers from the two other centers were all accredited by Australasian Sleep Technologists Association (ASTA) while not necessarily working together for the same facility.

The consensus hypnogram generated from the three manual PSG hypnograms was used as the final hypnogram for each PSG recording for all subsequent comparisons with the Somfit automatic hypnogram.

Somfit Sleep Staging Algorithm

The Somfit sleep staging algorithm is based on DL U-sleep CNN architecture.[36] It comprises 12 combinations of encoder and decoder units that take 35 consecutive 30 second epochs of EEG and EOG data as a two-dimensional input, map the input signals into a sequence of up to 302-dimensional feature spaces and then determine the probabilities of the five sleep stages for each one of the 35 epochs. As required by AASM,[1] Somfit records two EOG channels – right and left, therefore the inference model estimates the probabilities of sleep stages for the two input pairs – EEG+EOG (right) and EEG+EOG (left), then adds the respective probabilities and selects the stage with the highest combined probability as the output of the algorithm. This architecture was initially trained on 783 Somfit studies with the manual Somfit hypnograms used as the ground truth. The final model comprises more than three million optimized weight coefficients. To improve detection of stage N1, another similar DL architecture was trained on 216 parallel Somfit and PSG recordings with the long continuous stretches of stage N3 excluded from training and the manual PSG hypnograms used as the ground truth. If the second model detects N1, it overrules the output of the first model, if that output is Wake or N2 within 10 min of the nearest Wake epoch.

Statistical Analysis

The reported statistics include the following measures of agreement between the Somfit automatic hypnogram and consensus PSG hypnogram:

- Percent agreement across the five sleep stages – N1-N2-N3-REM-Wake;
- Kappa statistic across the five sleep stages;
- Percent agreement across the four sleep stages (Light NREM-N3-REM-Wake) with N1 and N2 merged into one category – light NREM sleep. This metric allows to remove the effect of errors in detection of N1 which is known for poor level of agreement[37] in all settings;
- Percent agreement across three categories – NREM-REM-Wake. This metric is important for OSA phenotyping which includes identification of the REM phenotype;[11]
- Percent agreement across two categories – Sleep-Wake. This statistic demonstrates the impact of sleep staging accuracy or agreement on the estimation of AHI;[1]
- Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and accuracy for each sleep stage.

The same statistics were generated to characterize the level of agreement between different manual PSG hypnograms.

Linear Mixed Models (LMM) with random intercepts were used to examine the significant differences ($P<0.05$) between the measures of Somfit/PSG agreement and interscorer agreement as well as between the estimates of agreement for different combinations of sleep stages. In addition to examination of superiority (that required non-negativity of 95% confidence interval of a respective regression coefficient) we explored noninferiority of the Somfit algorithm with respect to interscorer agreement. The margin of noninferiority [M_{NI}] was set at 5% for percent agreement, 0.05 for kappa statistic and 5% for the individual sleep stage performance measures based on the range of reported interscorer variability in various settings.[43–49] To establish noninferiority with LMM, 95% confidence interval of a respective regression coefficient had to fall within [$-M_{NI};\infty$]. ANOVA was used to test for the different mean agreement measures between the centers and sample categories (OSA presence/severity, gender, age, BMI). Repeated measures ANOVA was used to test for the different mean agreement measures between the pairs of scorers form the same center. Unpaired *t*-test was used to examine the significant differences between the sample categories with the hypothesized direction of change in the level of agreement. Stata 15 (StataCorp) statistical software was used for statistical analysis.

Results

Participants Characteristics

Characteristics of 106 patients who completed both Somfit and PSG recordings and were included into analysis are presented in Table 1. The participants' information includes ESS statistics and key comorbidities.

Integrative Estimates of Agreement Between Somfit Algorithm and PSG Consensus Hypnogram

The estimates of percent agreements between the Somfit hypnogram and consensus PSG hypnogram for the three centers separately and for the combined data set are presented in Table 2 and Figure 3. The percent agreement estimates for the individual studies are presented in the Tables S1–S3 for the respective centers. LMM analysis shows that for each center separately and for the combined data the percent agreement progressively increases when the number of categories is reduced from five sleep stages (N1-N2-N3-REM-Wake) to two (Sleep-Wake). ANOVA *F* values demonstrate that there are no significant differences between the estimates of percent agreement for the study centers for all sets of categories.

The respective kappa statistics for all five stages are shown in Table 3. For kappa statistic the two sets of results are presented – derived from the separate confusion matrices for each sleep study (left column) and from the pooled confusion matrix (right column). There is a noticeable difference between the two methods because of their impact on the

Table 1 Characteristics of the Study Cohort

Participants Characteristics		**Center 1, N=40**	**Center 2, N=33**	**Center 3, N=33**	**Total, N=106**
OSA status at enrolment	**Suspected**	37	4	33	74
	Known	3	29	0	32
ESS – Mean (SD)*		9.45 (4.71)	9.09 (5.39)	9.19 (4.45)	9.23 (4.82)
Cardiovascular Comorbidities		13	14	9	36
Respiratory Comorbidities		9	8	11	28
Diabetes		2	2	5	9
OSA severity	**No OSA**	11	6	4	21
	Mild	15	5	16	36
	Moderate	6	9	5	20
	Severe	8	13	8	29
Gender	**Male**	28	14	17	59
	Female	12	19	16	47
Age group	**<65 years old**	34	26	25	85
	≥65 years old	6	7	8	21
BMI category	**<25**	5	6	10	21
	[25:30]	17	12	14	43
	≥30	18	15	9	42

Notes: Cardiovascular comorbidities included high blood pressure, heart failure, heart attack/angina, stroke/brain aneurysm. Respiratory comorbidities included asthma and chronic obstructive pulmonary disease (COPD). *ANOVA F = 0.06 (*P*=0.98) – no difference between mean ESS at three Centers.

Table 2 Percent Agreement Between Somfit Algorithm and Consensus PSG Hypnogram

Center (No. Studies)	Percent Agreement – Mean (SE)			
	N1-N2-N3- REM-Wake *F* (P) = 1.21 (0.30)*	Light NREM-N3-REM-Wake *F* (P) = 0.50 (0.61)*	NREM-REM-Wake *F* (P) = 0.36 (0.70)*	Sleep-Wake *F* (P) = 0.42 (0.66)*
1 (40)	75.91 (1.15)**	79.87 (1.14)**	87.38 (1.15)**	89.89 (1.03)
2 (33)	74.72 (1.26)**	80.75 (0.98)**	86.15 (1.11)**	88.41 (0.98)
3 (33)	77.82 (1.72)**	81.66 (1.66)**	85.86 (1.82)**	89.45 (1.49)
Total (106)	76.14 (0.79)**	80.70 (0.74)**	86.52 (0.79)**	89.30 (0.37)

Notes: **F* and *P*-values for ANOVA tests between the three centers. **Inferior to the percent agreement with the reduced number of sleep stages, positioned to the right (*P*<0.05).

values of chance agreement. We decided to present both sets of results as the former method appears to be more adequate for characterization of single night studies and appropriate for statistical comparisons, yet the latter method is often used in the literature on staging algorithms.[22–30,32–42] ANOVA *F* value shows that there are no significant differences between the estimates of kappa for the study centers.

PSG Interscorer Agreement

The percent agreement estimates between all pairs of manual PSG hypnograms are shown in Table 4. It also demonstrates which interscorer agreement estimates are superior to the respective measures of agreement between Somfit and PSG consensus hypnograms, and for which study centers and comparison categories the Somfit agreement is noninferior. It is also evident form the repeated measures ANOVA tests that for the Centers 2 and 3 the hypotheses of no difference between mean pairwise comparisons cannot be supported. The ANOVA tests between the mean pairwise percent agreements at the three centers indicate the difference between the centers for all comparison categories.

Figure 4 shows all 95% confidence intervals for the interscorer percent agreement, together with the 95% confidence interval for the percent agreement between Somfit algorithm and consensus PSG hypnogram for the total data set.

Table 5 presents kappa statistics characterizing agreement between all pairs of manual PSG hypnograms for the five sleep stages. Similar to Table 4, all instances of superior interscorer PSG agreement and noninferior agreement between Somfit algorithm and PSG consensus hypnogram with respect to the interscorer agreement are marked based on the results of LMM regression. The findings of difference in the mean pairwise agreements at the Centers 2 and 3 (repeated measures ANOVA) and between the three Centers (ANOVA) are confirmed.

Somfit Algorithm Performance for OSA Severity and Other Study Subgroups

The percent agreement estimates for the Somfit algorithm against the consensus PSG hypnogram stratified for different OSA severity levels are presented in Table 6. With the expectation of reduction in the level of agreement as the severity of OSA increases, unpaired *t*-test demonstrates that only some comparisons reveal significant ($P<0.05$) decreases in the level of agreement – normals/mild versus severe for N1-N2-N3-REM-Wake and normals versus moderate/severe for NREM-REM-Wake and Sleep-Wake. ANOVA results also show that the mean percent agreements for the OSA categories are different only for N1-N2-N3-REM-Wake and Sleep-Wake comparisons.

The results for kappa between the Somfit algorithm and consensus PSG hypnogram are presented in Table 7. Unpaired *t*-test only shows decreases between normals/mild vs severe, and ANOVA confirms different means across the OSA categories.

The percent agreement and kappa statistics for the Somfit algorithm against the consensus PSG hypnogram stratified for the genders, age groups (younger and older than 65 years old) and BMI categories (normal, overweight and obese) are presented in the Tables S4–S9. The only detected change in the level agreement is the decrease in kappa between normal weight (BMI <25) and obese patients (BMI ≥30).

The percent agreement statistics across OSA severity, gender, age and BMI categories are graphed in Figure 5 – only showing the results for all five sleep stages (N1-N2-N3-REM-Wake).

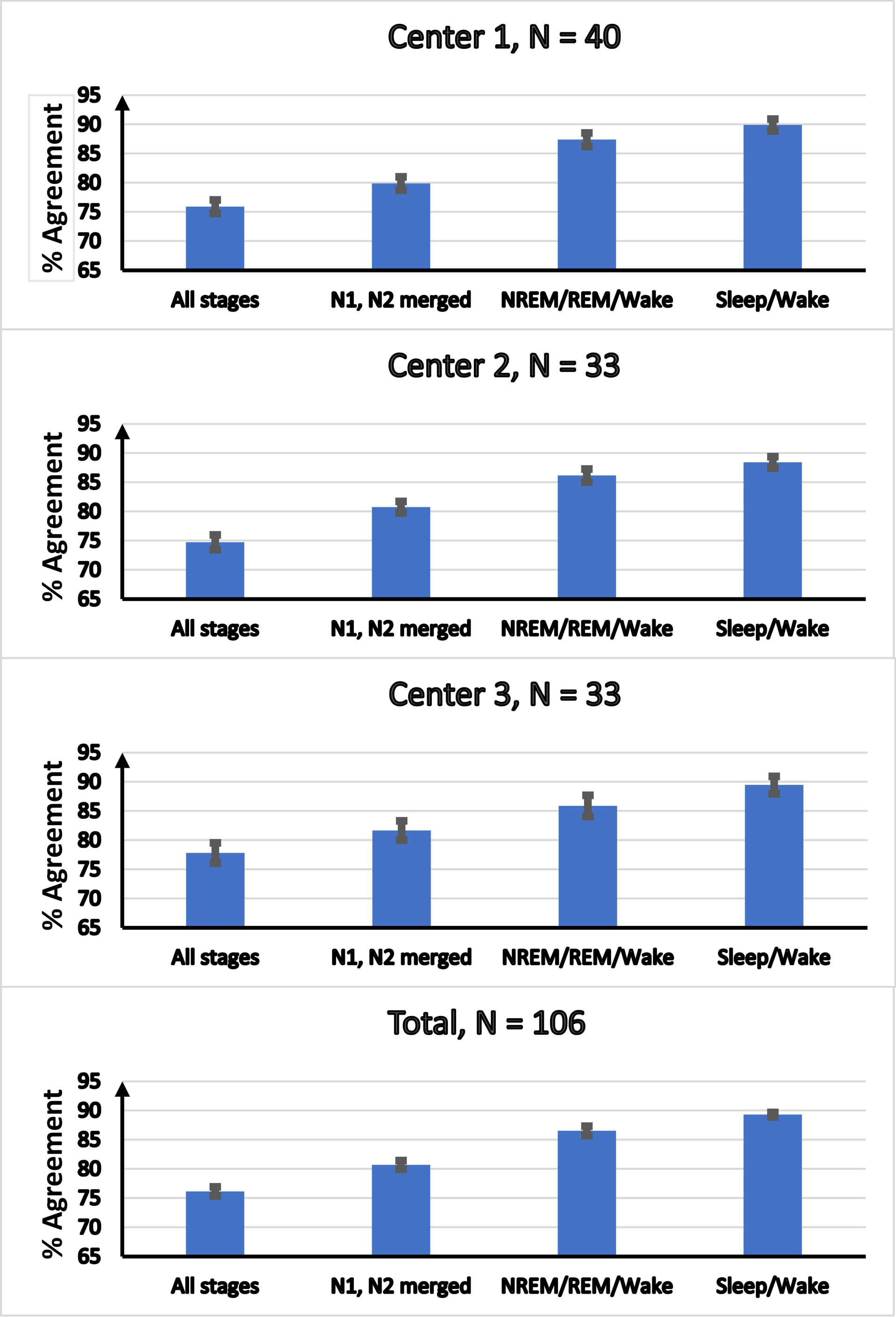

Figure 3 Percent agreement between Somfit algorithm and consensus PSG hypnogram.

Performance Measures for Individual Sleep Stages

Table 8 presents the estimates for five performance measures (sensitivity, specificity, PPV, NPV and accuracy), for different sleep stages. As these results are to be compared with the respective PSG interscorer performance measures (these estimates are presented in the Table S10), the results for the combined Centers 2 and 3, that employed the scorers from different facilities in contrast to the Centre 1, are shown separately in addition to the results for the full data set. On the basis of LMM regression

Table 3 Kappa Statistics for Agreement Between Somfit Algorithm and Consensus PSG Hypnogram – Five Sleep Stages (N1-N2-N3-REM-Wake)

Center (No. Studies)	**Kappa – Mean (SE)**	
	Sleep Study Based ***F* (P) = 0.85(0.43) ***	**Pooled Confusion Matrix**
1 (40)	0.650 (0.017)	0.672 (0.003)
2 (33)	0.637 (0.019)	0.655 (0.003)
3 (33)	0.676 (0.020)	0.689 (0.003)
Total (106)	0.654 (0.013)	0.672 (0.002)

Notes: **F* and *P*-values for ANOVA tests between the three centers.

Table 4 Interscorer Percent Agreement Between All Pairs of Manual PSG Hypnograms

Center (No. Studies)	**Scorers**	**Percent Agreement - Mean (SE)**			
		N1-N2-N3-REM-Wake[a,b,c]	**Light NREM-N3-REM-Wake**[a,b,c]	**NREM-REM-Wake**[a,b,c]	**Sleep-Wake**[a,b,c]
1 (40)	**A vs B**	84.45 (0.66)*	89.20 (0.56)*	95.23 (0.34)*	96.60 (0.30)*
	A vs C	85.50 (0.64)*	89.69 (0.51)*	95.14 (0.59)*	96.69 (0.24)*
	B vs C	84.46 (0.77)*	89.71 (0.67)*	95.28 (0.29)*	96.85 (0.27)*
	Average	84.81 (0.63)*	89.53 (0.51)*	95.22 (0.27)*	96.71 (0.25)*
2 (33)	**A vs B**	74.36 (1.93)**	81.85 (1.53)**	87.59 (1.47)**	89.59 (1.51)**
	A vs C	79.02 (1.27)*	86.60 (0.81)*	91.85 (0.66)*	93.20 (0.71)*
	B vs C	77.83 (1.37)*	83.57 (1.05)*	89.98 (1.10)*	92.12 (1.08)*
	Average	77.07 (1.43)**	84.01 (1.03)*	89.80 (1.02)*	91.63 (1.06)*
3 (33)	**A vs B**	76.67 (1.91)**	82.77 (1.95)**	86.95 (2.10)**	90.54 (2.12)**
	A vs C	83.90 (1.17)*	90.38 (0.71)*	94.00 (0.69)*	95.78 (0.71)*
	B vs C	78.41 (2.03)**	82.72 (2.07)**	86.67 (2.19)**	90.04 (2.21)**
	Average	79.66 (1.63)**	85.29 (1.51)*	89.21 (1.61)*	92.12 (1.65)**

Notes: *Superior to the respective percent agreement between Somfit algorithm and PSG consensus hypnogram (*p*<0.05). **Respective percent agreement between Somfit algorithm and PSG consensus hypnogram is noninferior (*P*<0.05). [a]Mean interscorer percent agreements at center 2 are different (*P*<0.05). [b]Mean interscorer percent agreements at center 3 are different (*P*<0.05). [c]Mean average interscorer percent agreements between the three centers are different (*P*<0.05).

tests, Table 8 highlights the performance measures where PSG interscorer estimates are superior to those of the Somfit algorithm, and where the Somfit algorithm estimates are superior and noninferior with respect to the interscorer estimates.

Discussion

There were significant discrepancies in some patient characteristics between different centers. While almost all patients enrolled at Centers 1 and 3 only had suspected OSA, the majority of the patients enrolled at Center 2 had a known history of OSA and required diagnostic review, primarily due to a significant recent weight change. This resulted in over-representation of the moderate and severe OSA populations at Center 2 (two thirds of the enrolled participants) compared to the Centers 1 and 3 at 35% and 39%, respectively. Despite that, the hypothesis of no difference in the mean ESS

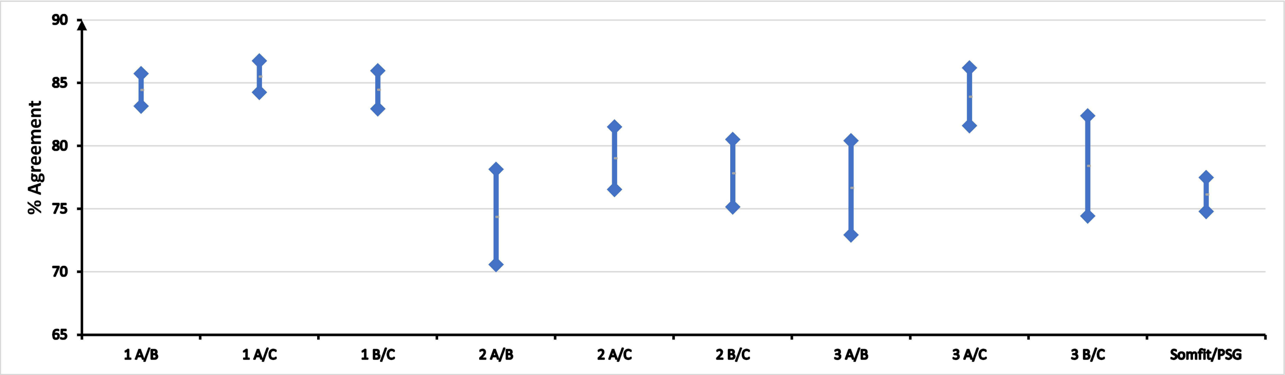

Figure 4 Ninety-five percent confidence intervals for percent agreement between all pairs of manual PSG hypnograms and percent agreement between Somfit algorithm and consensus PSG hypnogram.

between the centers could not be rejected, most likely because majority of the participants at Center 2 were on CPAP at home. Overall, the distribution of different OSA categories at 20%, 34%, 19%, and 27% for healthy, mild, moderate and severe OSA is similar to the respective values of 17%, 31%, 23%, and 29% reported for a data set of 891 consecutive diagnostic studies in a large Australian public sleep unit.[36]

Review of interscorer agreement (Table 4 and Figure 4) indicated that the mean overall percent agreement at Center 1 was consistently high at about 85% between all pairs of scorers, while for the other two centers it was under 80%, except the value of 83.90% between the scorers A and C at Center 3. The likely causes of such discrepancy are long-term affiliation of all scorers used by Center 1 with the same sleep laboratory and predominance of healthy and mild OSA

Table 5 Kappa Statistics for Agreement Between All Pairs of Manual PSG Hypnograms – Five Sleep Stages (N1-N2-N3-REM-Wake)

Center (No. Studies)	Scorers	Mean (SE)	
		Sleep Study Based[a,b,c]	Pooled Confusion Matrix
1 (40)	A vs B	0.781 (0.009)*	0.794 (0.002)
	A vs C	0.794 (0.010)*	0.807 (0.002)
	B vs C	0.780 (0.012)*	0.793 (0.002)
	Average	0.785 (0.009)*	
2 (33)	A vs B	0.649 (0.025)	0.661 (0.003)
	A vs C	0.711 (0.014)*	0.725 (0.003)
	B vs C	0.693 (0.019)*	0.705 (0.003)
	Average	0.684 (0.019)*	
3 (33)	A vs B	0.665 (0.028)**	0.676 (0.003)
	A vs C	0.775 (0.016)*	0.781 (0.003)
	B vs C	0.687 (0.029)**	0.696 (0.003)
	Average	0.709 (0.024)*	

Notes: *Superior to the respective kappa between Somfit algorithm and PSG consensus hypnogram (*P*<0.05). **Respective kappa between Somfit algorithm and PSG consensus hypnogram is noninferior (*P*<0.05). [a]Mean interscorer kappa at Center 2 are different (*P*<0.05). [b]Mean kappa at Center 3 are different (*P*<0.05). [c]Mean average interscorer kappa between the three centers are different (*P*<0.05).

Table 6 Percent Agreement Between Somfit Algorithm and Consensus PSG Hypnogram for Different OSA Categories

OSA Category (No. Studies)	Percent Agreement - Mean (SE)			
	N1-N2-N3-REM-Wake *F* (P) = 3.46 (0.02)*	Light NREM-N3-REM-Wake *F* (P) = 0.63 (0.60)	NREM-REM-Wake *F* (P) = 1.94 (0.13)	Sleep-Wake *F* (P) = 3.03 (0.03)*
No OSA (21)	79.67 (1.87)***	82.42 (1.73)	89.86 (1.31)*****	92.56 (1.15)*****
Mild (36)	77.38 (1.06)***	80.98 (0.99)	86.89 (1.08)	89.86 (0.84)
Moderate (20)	74.83 (1.79)	79.51 (1.72)	84.73 (2.02)	88.29 (1.64)
Severe (29)	72.93 (1.68)	79.94 (1.68)	84.87 (1.86)	86.97 (1.64)

Notes. *Mean percent agreements are different for OSA categories, ANOVA (*P*<0.05). **Greater than percent agreement for moderate OSA (*P*<0.05). ***Greater than percent agreement for severe OSA (*P*<0.05).

Table 7 Kappa Statistics for Agreement Between Somfit Algorithm and Consensus PSG Hypnogram – Five Stages (N1-N2-N3-REM-Wake), Different OSA Categories

OSA Category (No. Studies)	Kappa - Mean (SE)	
	Sleep Study Based *F* (P) = 4.90(<0.01)*	Pooled Confusion Matrix
No OSA (21)	0.711 (0.026)**	0.723 (0.004)
Mild (36)	0.678 (0.015)**	0.692 (0.003)
Moderate (20)	0.638 (0.026)	0.650 (0.004)
Severe (29)	0.595 (0.025)	0.618 (0.004)

Notes: *Mean kappa estimates are different for OSA categories, ANOVA (*P*<0.05). **Greater than kappa for severe OSA (*P*<0.05).

cases at this center compared to the other centers. The overall percent agreement levels of under 80% are widely reported for the data sets with significant OSA representation and the scorers affiliated with different sleep centers.[18,27,45,48,49] Table 5 shows that the respective mean kappa statistic estimates are above 0.78 for Center 1 and around 0.7 or lower for most comparisons at the other two centers. The estimates for the Center 1 are close to the range of 0.81–0.83 reported for the scorers following the same training procedure,[44] while the lower estimates for the Centers 2 and 3 are consistent with the values of 0.73 and 0.68 reported for the multicenter studies with significant OSA representation.[46,49] Our findings for kappa are also consistent with the overall meta-analysis estimate of 0.76[43] which summarizes the reported interrater agreement investigations. Considering the individual sleep stages the reported lowest interscorer agreement for Stage N1,[43,47] in comparison with the other stages, is confirmed by our findings of mean interscorer N1 sensitivity at 45.79 for all participants and 40.20 for the Centers 2 and 3 combined. As the scorers for Centers 2 and 3 were not trained at the same sleep laboratories, all measures of agreement indicated within center differences for these two Centers in contrast to those at Center 1. Particular prominent patterns of interscorer disagreement included the tendency of Scorer B from Center 2 to score sleep stages in a more "blocky" fashion in contrast to the other scorers more likely to present the fragmented hypnograms, and also a tendency of Scorer B from Center 3 to discriminate between sleep and wake more strictly according to AASM rules in contrast to the other scorers who appear to also take into account the breathing and desaturation patterns. Such differences in sleep scoring patterns lead to significant disagreements in the recordings of patients with moderate and severe OSA which is also confirmed by the outcomes of ANOVA tests between the centers.

Results of percent agreement between Somfit automatic hypnogram and consensus PSG hypnogram (Table 2 and Figure 3) at different centers demonstrate similar mean estimates at 75.91%, 74.72%, and 77.82%, with the lowest mean

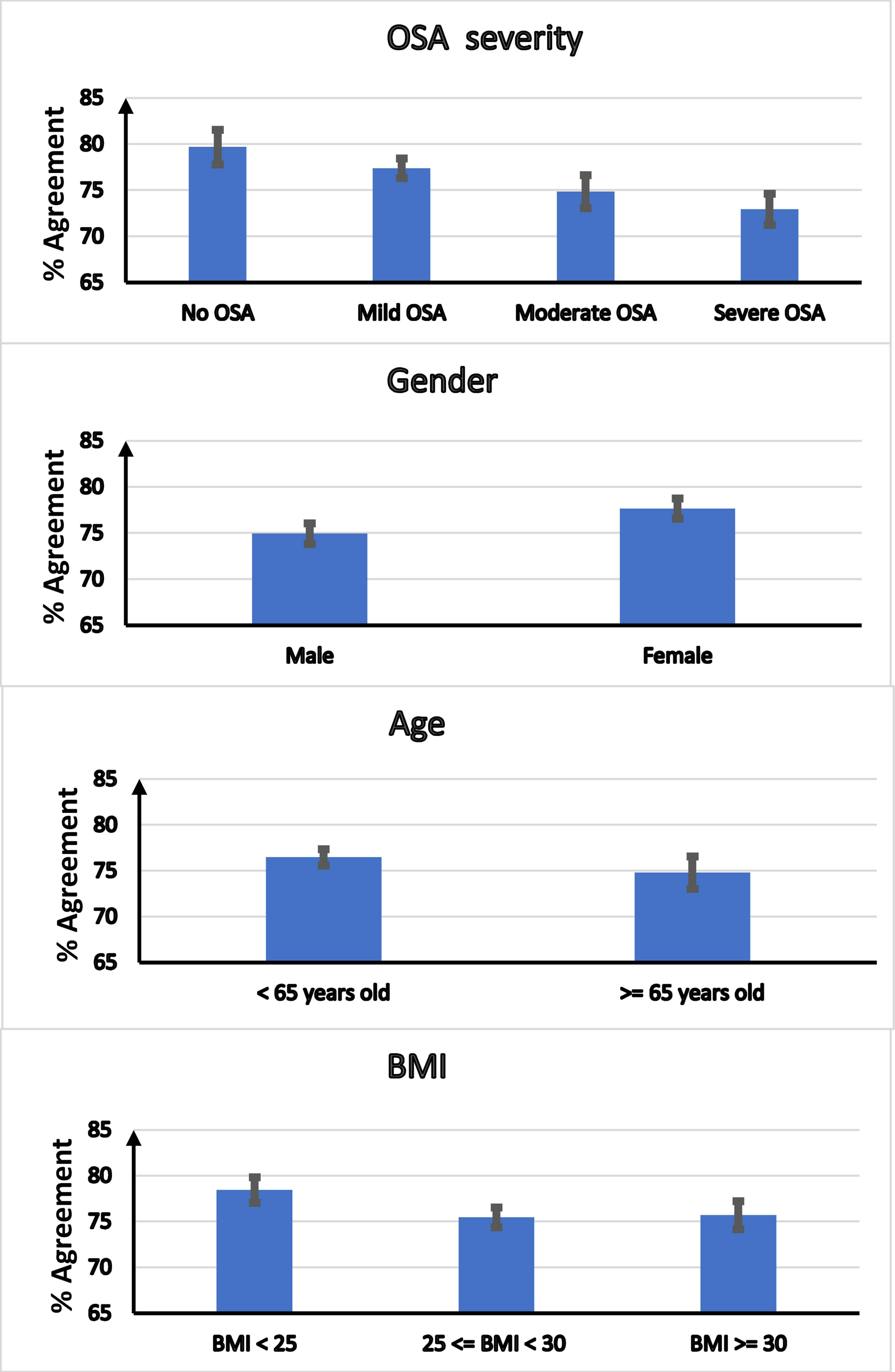

Figure 5 Percent agreement between Somfit algorithm and consensus PSG hypnogram (five sleep stages) across different patient characteristics.

agreement at Centre 2, which had the largest combined proportion of moderate and severe OSA at 67%. ANOVA confirms no difference in means at the three centers for all comparison categories, As demonstrated in Table 2, percent agreement estimates for all centers consistently increased when the number of categories was reduced from the five sleep stages to four (N1 and N2 merged to light NREM), then to three (NREM-REM-Wake), and finally to two (Sleep-Wake) with respective mean percentages at different centers being within two percent of each other. For the total data set, the

Table 8 Somfit Algorithm Performance Measures for Sleep Stages – Mean (SE)

Measure	Centers (No. Studies)	Sleep Stage				
		Wake	N1	N2	N3	REM
Sensitivity	**All (106)**	78.23 (1.53)*	22.34 (1.13)*	83.91 (1.40)***	58.40 (2.78)*	87.17 (1.77)**
	2+3 (66)	74.29 (2.06)*	21.74 (1.49)*	84.76 (1.76)***	62.38 (3.47)*	87.10 (2.23)**
Specificity	**All (106)**	91.45 (1.03)*	96.83 (0.24)***	82.74 (0.85)*	98.68 (0.21)***	95.85 (0.39)**
	2+3 (66)	92.53 (1.29)**	21.74 (1.49)***	82.59 (1.10)*	98.42 (0.30)***	95.57 (0.56)**
PPV	**All (106)**	76.30 (1.56)*	38.13 (1.57)*	75.98 (0.94)*	84.6 (2.33)***	75.74 (2.13)*
	2+3 (66)	79.99 (1.80)*	36.05 (1.93)**	76.31 (1.23)*	82.61 (2.80)***	75.71 (2.63)*
NPV	**All (106)**	93.19 (0.56)**	93.13 (0.45)**	89.96 (0.62)***	94.17 (0.48)**	98.18 (0.22)**
	2+3 (66)	91.61 (0.76)**	92.98 (0.62)**	90.35 (0.79)***	95.68 (0.41)**	98.11 (0.28)**
Accuracy	**All (106)**	89.30 (0.67)*	90.58 (0.47)**	83.73 (0.56)**	93.78 (0.39)**	94.88 (0.36)**
	2+3 (66)	88.93 (0.89)**	90.29 (0.64)***	83.92 (0.73)**	94.85 (0.36)**	94.55 (0.50)**

Notes: *Respective PSG interscorer estimate is superior (*P*<0.05). **Noninferior to respective PSG interscorer estimate (*P*<0.05). ***Superior to respective PSG interscorer estimate (*P*<0.05).

respective mean percent agreement estimates increased from 76% to 81% when N1 and N2 were merged, to above 86% for the three categories – NREM-REM-Wake, and then to above 89% for the sleep/wake discrimination. The high value of the latter number would contribute to reliability of the AHI values reported by Somfit, as they rely on an accurate TST. Table 4 demonstrates that the Somfit algorithm percent agreement across five categories is noninferior to PSG interscorer agreement for the three out of nine pairwise comparisons, and also for the average interscorer percent agreements at Centers 2 and 3. There are also a number of noninferior comparisons at Centers 2 and 3 with the reduced sets of categories, however none of the comparisons for Centre 1 reveal noninferiority of the Somfit algorithm. The relationships for the N1-N2-N3-REM-Wake comparisons are further highlighted in Figure 4 with 95% confidence intervals for all percent agreement estimates.

All estimates of kappa (Tables 3, 5, 7 and also S5, S7 and S9 in the Supplement) reveal the reduction of up to 0.02 with the change from the pooled confusion matrix (right column), that is typically used in the publications, to the estimates based on kappa for individual studies (left column), that appears more relevant to the research questions and amenable to the correct statistical analysis. Such difference can be explained by the fact that the individual studies with lower kappa impact the overall estimate to the greater extent when they are included into analysis directly rather via the pooled confusion matrix. Similar to the percent agreement, there is no difference between mean kappa estimates at the three centers, with the mean kappa for the full data set at 0.654, or 0.672 if the pooled confusion matrix method is used (Table 3). According to Table 5, kappa for the agreement between Somfit and PSG consensus is noninferior to PSG interscorer kappa only for the pairs A/B and B/C at Center 3.

The presented percent agreement and kappa estimates demonstrate that the accuracy of Somfit sleep staging algorithm most likely exceeds that reported by Sleep Profiler[25,27] and the frontal EEG algorithm[30] for the data sets including OSA sufferers. The range of kappa statistic for those studies was between 0.61 and 0.63. However, the DL algorithm based on a combination of EEG derivation F4-M1 and EOG derivation E1-M2 reported the higher reported accuracy of 83.8% and kappa of 0.78.[36] We will comment further on possible reasons for such discrepancy and outline the steps suggested to improve the accuracy of the Somfit algorithm.

Table 8 shows that the sensitivity of the stages N2 and REM is noninferior for Somfit with respective to the mutual PSG interscorer sensitivities, however the sensitivities of Wake, N1 and N3 are not noninferior and in fact the respective PSG interscorer sensitivities are superior. While the sensitivity of Wake is not noninferior, it is still relatively high at 78.23% for the full data set. However the sensitivities of N3 and especially N1 are significantly lower for the Somfit algorithm. The reduction in N3 sensitivity is exacerbated by the fact that our analysis did not use the pooled confusion matrix, so the moderate/severe OSA studies with small N3 durations and likely underestimation of N3 by Somfit would have significant impact on the mean estimate for the complete data set. It is well established that detection of stage N1 is the most difficult for both automatic algorithms and interscorer agreement[43] which is supported by the respective accuracies of 23% and 47% reported by Sleep Profiler[27] and DL algorithm.[36] The correct detection of N1 is more challenging for the frontal EEG derivations owing to the greater distance from the occipital placement preferred for detection of the alpha patterns.[1] Table 8 includes the estimates for the partial date set formed by Centers 2 and 3 to eliminate the effect of scorer training at the same center. This does not change the outcome of sensitivity analysis however for a number of performance measures (specificity and accuracy of Wake and PPV of N1) noninferiority can be established for the reduced data set in contrast to the full data set.

The hypothesized reduction in the level of agreement with the increase in OSA severity was only established for some comparisons – with transitions from no OSA or mild OSA to severe OSA for the five categories percent agreement and kappa, as well with transitions from no OSA to moderate or severe OSA for NREM-REM-Wake and Sleep-Wake percent agreement (Tables 6 and 7 and Figure 5), despite the consistent trends for the mean estimates across all OSA severity categories. These observations are supported by the findings for interscorer agreement[43] and for the accuracy of sleep staging algorithms.[27,30,36] Similar trends without the significant differences, except reduction in kappa between the normal and overweight BMI, were observed when we compared the genders (Tables S4 and S5 and Figure 5), age groups (Tables S6 and S7 and Figure 5) and BMI categories (Tables S8 and S9 and Figure 5). Presence and increasing severity of OSA and the subsequent increase in sleep fragmentation is most likely the main factor responsible for reduction in algorithm accuracy.

A contributing factor into reduction in the level of agreement for the Somfit algorithm with the PSG hypnograms was that every full night Somfit study (except the one from Center 3 with almost continuous electrode disconnection) was included in the data analysis. Five studies (two from Center 1–10 and 35, one from Center 2–20 and two from Center 3–4 and 11) revealed the noisy EEG signals as a consequence of incorrect electrode application and/or excessive patient movement. The range of percent agreement with the consensus PSG hypnogram for these five studies was between 48% and 62% resulting in significant impact on the overall level of agreement. In two of these studies (4 and 11 from Center 3) the PSG signals also had excessive high frequency resulting in very poor interscorer percent agreement of as low as 38% for one pair of scorers. It was evident that a small percentage of severe OSA studies demonstrated very noisy appearance of EEG/EOG signals not only for Somfit, but also on a number of occasions for PSG recordings. In addition to these observations there were inconsistencies with the manual scoring of stage N3 at Center 1, resulting in apparent overdetection of N3 by Center 1 scorers. This factor resulted in relatively poor overall percent agreement for a number of studies at Center 1, for example 17, 23 and 33, and the pooled confusion matrix sensitivity for Stage N3 at this center being at 59%, while staying above 69% for the other two centers.

Regarding the comparison of our results with those reported for the combined frontal EEG and EOG DL algorithm[36] we believe that there were multiple factors responsible for the difference in performance:

- DL algorithm[36] excluded about 5% of consecutive diagnostic studies;
- DL algorithm[36] was trained and tested on studies recorded at the same center and scored by the same scorers;
- The average interscorer agreement in our study was 80.43% therefore it is unlikely that the agreement of Somfit algorithm with the PSG consensus can be higher than this level;
- The mean AHI for the study[36] was 15.8 while it was 21.1 for our study which makes it more challenging for the sleep staging algorithm performance;
- DL algorithm[36] utilized EEG and EOG derivations referenced to the mastoid process which makes the signal patterns cleaner but this diagnostic set-up problematic for self-application.

While the presented results demonstrate that the accuracy of the Somfit sleep staging algorithm approaches the level of interrater agreement for accredited sleep technologists, this study also allows to identify a number of steps for further improvement of the algorithm to realize the maximum potential of the DL approach:

- Integration of a DL arousal detection algorithm and retraining of a DL sub-model for stage N1 on a larger simultaneous Somfit/PSG data set (500 studies plus) to rectify "blocky" appearance of a hypnogram, represent sleep fragmentation more accurately and lift the stage N1 sensitivity to the level of at least 50%;
- Addition of a DL sub-model for stage N3 trained on a simultaneous Somfit/PSG data set to increase the stage N3 accuracy to the level of at least 80%;
- Exploration of a larger training data set and possibility of addition of the Somfit motion signal to the main DL model to rectify occasional misclassification of Wake as REM sleep – examples of such errors can be found in studies 22 and 28 from Center 3;
- Addition of single channel DL models or DL models based on non-EEG Somfit channels (motion, SpO2, PAT) to be activated under the conditions of excessive noise or high impedance.

The results of the study also imply that, with the possibility of manual editing, the Somfit data are likely to be immediately fully adequate for producing the hypnograms of equivalent accuracy to the full PSG. As also supported by the results[36] and a number of other single channel DL models,[39] the separate chin EMG signal may not necessarily be required to produce valid hypnograms for the correct diagnosis of OSA.

Conclusion

Our multicenter clinical trial in a population with suspected or known OSA investigated the performance of a sleep staging algorithm implemented in Compumedics Somfit, a miniaturized HSAT device attached to a patient's forehead. We established that the estimates of agreement between the Somfit hypnogram and simultaneous PSG hypnogram were noninferior to the level of PSG interscorer agreement for a number of indicators. It was possible to achieve a high level of algorithm accuracy owing to the application of deep learning methodology.

Data Sharing Statement

The data that support the findings of this study are available from the corresponding author, EZ, upon reasonable request. The authors intend to share the deidentified questionnaire, PSG and Somfit data of all study participants and the study protocol. The shared data will be available for a period of seven years after the date of publication.

Disclosure

Dr Jeremy Goldin is a medical consultant for Compumedics Pty Ltd. Ms Elizabeth Kealy reports personal fees from Compumedics, during the conduct of the study; personal fees from ResMed and Compumedics, outside the submitted work. Mr Darrel Wicks reports grants from Compumedics Ltd, during the conduct of the study. The authors report no other conflicts of interest in this work.

References

1. Troester MM, et al. The AASM Manual for the Scoring of Sleep and Associated Events. *Rules, Terminology and Technical Specifications*; 2023. Version 3 AASM.
2. Kushida CA, Littner MR, Morgenthaler T, et al. Practice parameters for the indications for polysomnography and related procedures: an update for 2005. *Sleep*. 2005;28(4):499–521. doi:10.1093/sleep/28.4.499
3. Barnes M, McEvoy R, Pierce R. Neurobehavioural impairment in mild sleep apnea patients compared to control subjects. *Sleep*. 2001;24S:A276.
4. Worsnop C, Pierce R, McEvoy R. Obstructive sleep apnoea. *Aust N Z J Med*. 1998;28(4):421–427. doi:10.1111/j.1445-5994.1998.tb02074.x
5. Kapur VK, Auckley DH, Chowdhuri S, et al. Clinical practice guideline for diagnostic testing for adult obstructive sleep apnea: an American Academy of Sleep Medicine clinical practice guideline. *J Clin Sleep Med*. 2017;13(3):479–504. doi:10.5664/jcsm.6506
6. Caples SM, Anderson WM, Calero K, Howell M, Hasimi SD. Use of polysomnography and home sleep apnea tests for the longitudinal management of obstructive sleep apnea in adults: an American Academy of Sleep Medicine clinical guidance statement. *J Clin Sleep Med*. 2021;17(6):1287–1293. doi:10.5664/jcsm.9240
7. Douglas JA, Chai-Coietzer CL, McEvoy D, et al. Guidelines for sleep studies in adults - a position statement of the Australasian sleep association. *Sleep Med*. 2017;36(Suppl 1):S2–S22. doi:10.1016/j.sleep.2017.03.019
8. Jorgensen G, Downey C, Goldin J, Melehan K, Rochford P, Ruehland W. An Australasian commentary on the AASM manual for the scoring of sleep and associated events. *Sleep Biol Rhythms*. 2020;18(3):163–185. doi:10.1007/s41105-020-00259-9
9. Penzel T, Kesper K, Pinnow I, Becker HF, Vogelmeier C. Peripheral arterial tonometry, oximetry and actigraphy for ambulatory recording of sleep apnea. *Physiol Meas*. 2004;25(4):1025–1036. doi:10.1088/0967-3334/25/4/019

10. Tondo P, Drigo R, Scioscia G, et al. Usefulness of sleep events detection using a wrist worn peripheral arterial tone signal device (WatchPAT™) in a population at low risk of obstructive sleep apnea. *J Sleep Res*. 2021;30(6):e13352. doi:10.1111/jsr.13352
11. Ioachimescu OC, Allam JS, Samarghandi A, et al. Performance of peripheral arterial tonometry–based testing for the diagnosis of obstructive sleep apnea in a large sleep clinic cohort. *J Clin Sleep Med*. 2020;16(10):1663–1674. doi:10.5664/jcsm.8620
12. Iftikhar IH, Finch CE, Shah AS, Augunstein CA, Ioachimescu OC. Ioachimescu OC A meta-analysis of diagnostic test performance of peripheral arterial tonometry studies. *J Clin Sleep Med*. 2022;18(4):1093–1102. doi:10.5664/jcsm.9808
13. Hedner J, White DP, Malhotra A, et al. Sleep staging based on autonomic signals: a multi-center validation study. *J Clin Sleep Med*. 2011;7 (3):301–306. doi:10.5664/JCSM.1078
14. Holmedahl NH, Fjeldstad OM, Engan H, Saxvig IW, Grønli J. Validation of peripheral arterial tonometry as tool for sleep assessment in chronic obstructive pulmonary disease. *Sci Rep*. 2019;9(1):19392. doi:10.1038/s41598-019-55958-2
15. Bianchi M, Goparaju B. Potential underestimation of sleep apnea severity by at-home kits: rescoring in-laboratory polysomnography without sleep staging. *J Clin Sleep Med*. 2017;13(4):551–555. doi:10.5664/jcsm.6540
16. Setty AR. Underestimation of sleep apnea with home sleep apnea testing compared to in-laboratory sleep testing. *J Clin Sleep Med*. 2017;13 (4):531–532. doi:10.5664/jcsm.6534
17. Gabryelska A, Bialasiewicz P. Association between excessive daytime sleepiness, REM phenotype and severity of obstructive sleep apnea. *Sci Rep*. 2020;10(1):34. doi:10.1038/s41598-019-56478-9
18. Ruehland WR, O'Donoghue FJ, Pierce PJ, et al. The 2007 AASM recommendations for EEG electrode placement in polysomnography impact on sleep and cortical arousal scoring. *Sleep*. 2011;34(1):73–81. doi:10.1093/sleep/34.1.73
19. Yo SW, Joosten SA, Wimaleswaran H, et al. Body position during laboratory and home polysomnography compared to habitual sleeping position at home. *J Clin Sleep Med*. 2022;18(9):2103–2111. doi:10.5664/jcsm.9990
20. Berthomier C, Drouot X, Herman-Stoïca M, et al. Automatic analysis of single-channel sleep EEG validation in healthy individuals. *Sleep*. 2007;30 (11):1585–1587. doi:10.1093/sleep/30.11.1587
21. Light MP, Casimire TN, Chua C, et al. Addition of frontal EEG to adult home sleep apnea testing: does a more accurate determination of sleep time make a difference? *Sleep Breath*. 2018;22(4):1179–1188. doi:10.1007/s11325-018-1735-2
22. Younes M, Younes M, Giannouli E. Accuracy of automatic polysomnography scoring using frontal electrodes. *J Clin Sleep Med*. 2016;12 (5):735–746. doi:10.5664/jcsm.5808
23. Levendowski DJ, Popovic D, Berka C, Westbrook PR. Retrospective cross-validation of automated sleep staging using electroocular recording in patients with and without sleep disordered breathing. *Int Arch Med*. 2012;5(1):21. doi:10.1186/1755-7682-5-21
24. Popovic D, Khoo M, Westbrook PR. Automatic scoring of sleep stages and cortical arousals using two electrodes on the forehead: validation in healthy adults. *J Sleep Res*. 2014;23(2):211–221. doi:10.1111/jsr.12105
25. Stepnowsky C, Levendowski DJ, Popovic D, Ayappa I, Rapoport DM. Scoring accuracy of automated sleep staging from a bipolar electroocular recording compared to manual scoring by multiple raters. *Sleep Med*. 2013;14(11):1199–1207. doi:10.1016/j.sleep.2013.04.022
26. Finan PH, Richards JM, Gamaldo CE, et al. Validation of a wireless, self-application, ambulatory electroencephalographic sleep monitoring device in healthy volunteers. *J Clin Sleep Med*. 2016;12(11):1443–1451. doi:10.5664/jcsm.6262
27. Levendowski DJ, Ferini-Strambi L, Gamaldo C, Cetel M, Rosenberg R, Westbrook PR. The accuracy, night-to-night variability, and stability of frontopolar sleep electroencephalography biomarkers. *J Clin Sleep Med*. 2017;13(6):791–803. doi:10.5664/jcsm.6618
28. Lucey BP, McLeland JS, Toedebusch CD, et al. Comparison of a single-channel EEG sleep study to polysomnography. *J Sleep Res*. 2016;25 (6):625–635. doi:10.1111/jsr.12417
29. Levendowski DJ, St Louis EK, Ferini-Strambi L, Galbiati A, Westbrook PR, Berka C. Comparison of EMG power during sleep from the submental and frontalis muscles. *Nat Sci Sleep*. 2018;10:431–437. doi:10.2147/NSS.S189167
30. Lee PL, Huang YH, Lin PC, et al. Automatic sleep staging in patients with obstructive sleep apnea using single-channel frontal EEG. *J Clin Sleep Med*. 2019;15(10):1411–1420. doi:10.5664/jcsm.7964
31. Kapoor A, Gulli A, Pal S, Chollet F. *Deep Learning with TensorFlow and Keras*. Packt Publishing. 3rd Ed; 2022.
32. Tsinalis O, Matthews PM, Guo Y. Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders. *Ann Biomed Eng*. 2016;44(5):1587–1597. doi:10.1007/s10439-015-1444-y
33. Supratak A, Dong H, Wu C, Guo Y. DeepSleepNet: a model for automatic sleep stage scoring based on raw single-channel EEG. *IEEE Trans Neural Syst Rehabil Eng*. 2017;25(11):1998–2008. doi:10.1109/TNSRE.2017.2721116
34. Phan H, Andreotti F, Cooray N, Chen OY, De Vos M. Joint classification and prediction CNN framework for automatic sleep stage classification. *IEEE Trans Biomed Eng*. 2019;66(5):1285–1296. doi:10.1109/TBME.2018.2872652
35. Mousavi S, Afghah F, Acharya UR. SleepEEGNet: automated sleep stage scoring with sequence to sequence deep learning approach. *PLoS One*. 2019;14(5):e0216456. doi:10.1371/journal.pone.0216456
36. Korkalainen H, Aakko J, Nikkonen S, et al. Accurate deep learning-based sleep staging in a clinical population with suspected obstructive sleep apnea. *IEEE J Biomed Health Inform*. 2020l;24(7):2073–2081. doi:10.1109/JBHI.2019.2951346
37. Qu W, Wang Z, Hong H, et al. A residual based attention model for eeg based sleep staging. *IEEE J Biomed Health Inform*. 2020;24 (10):2833–2843. doi:10.1109/JBHI.2020.2978004
38. Zhang J, Yao R, Ge W, Gao J. Orthogonal convolutional neural networks for automatic sleep stage classification based on single-channel EEG. *Comput Methods Programs Biomed*. 2020;183:105089. doi:10.1016/j.cmpb.2019.105089
39. Fu M, Wang Y, Chen Z, et al. Deep learning in automatic sleep staging with a single channel electroencephalography. *Front Physiol*. 2021;12:628502. doi:10.3389/fphys.2021.628502
40. Khalili E, Asl BM. Automatic sleep stage classification using temporal convolutional neural network and new data augmentation technique from raw single-channel EEG. *Comput Methods Programs Biomed*. 2021;204:106063. doi:10.1016/j.cmpb.2021.106063
41. Eldele E, Chen Z, Liu C, et al. An attention-based deep learning approach for sleep stage classification with single-channel EEG. *IEEE Trans Neural Syst Rehabil Eng*. 2021;29:809–818. doi:10.1109/TNSRE.2021.3076234
42. Perslev M, Darkner S, Kempfner L, Nikolic M, Jennum PJ, Igel C. U-Sleep: resilient high-frequency sleep staging. *NPJ Digit Med*. 2021;4(1):72. doi:10.1038/s41746-021-00440-5

43. Lee YJ, Lee JY, Cho JH, Choi JH. Interrater reliability of sleep stage scoring: a meta-analysis. *J Clin Sleep Med*. 2022;18(1):193–202. doi:10.5664/jcsm.9538
44. Whitney CW, Gottlieb DJ, Redline S, et al. Reliability of scoring respiratory disturbance indices and sleep staging. *Sleep*. 1998;21(7):749–757. doi:10.1093/sleep/21.7.749
45. Norman RG, Pal I, Stewart C, Walsleben JA, Rapoport DM. Interobserver agreement among sleep scorers from different centres in a large dataset. *Sleep*. 2000;23(7):901–908. doi:10.1093/sleep/23.7.1e
46. Pittman SD, MacDonald MM, Fogel RB, et al. Assessment of automated scoring of polysomnographic recordings in a population with suspected sleep-disordered breathing. *Sleep*. 2004;27(7):1394–1400. doi:10.1093/sleep/27.7.1394
47. Rosenberg RS, Van Hout S. The American Academy of sleep medicine inter-scorer reliability program: sleep stage scoring. *J Clin Sleep Med*. 2013;9(1):81–87. doi:10.5664/jcsm.2350
48. Younes M, Raneri J, Hanly P. Staging sleep in polysomnograms: analysis of inter-scorer variability. *J Clin Sleep Med*. 2016;12(6):885–894. doi:10.5664/jcsm.5894
49. Danker-Hopfe H, Kunz D, Gruber G, et al. Interrater reliability between scorers from eight European sleep laboratories in subjects with different sleep disorders. *J Sleep Res*. 2004;13(1):63–69. doi:10.1046/j.1365-2869.2003.00375.x
50. Miller DJ, Sargent C, Roach GD. A validation of six wearable devices for estimating sleep, heart rate and heart rate variability in healthy adults. *Sensors*. 2022;22(16):6317. doi:10.3390/s22166317

Nature and Science of Sleep Dovepress

Publish your work in this journal

Nature and Science of Sleep is an international, peer-reviewed, open access journal covering all aspects of sleep science and sleep medicine, including the neurophysiology and functions of sleep, the genetics of sleep, sleep and society, biological rhythms, dreaming, sleep disorders and therapy, and strategies to optimize healthy sleep. The manuscript management system is completely online and includes a very quick and fair peer-review system, which is all easy to use. Visit http://www.dovepress.com/testimonials.php to read real quotes from published authors.

Submit your manuscript here: https://www.dovepress.com/nature-and-science-of-sleep-journal