

Manual Sleep Stage and Arousal Scoring Agreement between a Reduced Montage and Conventional Polysomnography in Adults with Obstructive Sleep Apnea

Megan McPhee

megan@noxmedical.com

Nox Medical ehf

Kristófer Montazeri

Nox Medical ehf

Þóra Björg Sigmarsdóttir

Nox Medical ehf

Sigurður Ægir Jónsson

Nox Medical ehf

Eydís Arnardóttir

Nox Medical ehf

Eysteinn Finnsson

Nox Medical ehf

Natalie Morin

Nox Medical ehf

Rósa Hugosdóttir

Nox Medical ehf

Gerrit Weimann

BAYER Pharmaceutical AG

Frank Kramer

BAYER Pharmaceutical AG

Jón Skírnir Ágústsson

Nox Medical ehf

Article

Keywords: Polysomnography, Sleep staging, Obstructive sleep apnea, Electroencephalography, Arousal, Validation study

Posted Date: October 11th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10777912/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: Competing interest reported. M.M, K.M., P.B.S., S.Æ.J., E.A., E.F., N.M., R.H., & J.S.A. are full time employees of Nox Medical. At the time of the study G.W. and F.K. were full-time employees of BAYER Pharma AG, the sponsor of the SPRAY-SMART study.

Abstract

Conventional polysomnography (PSG) is resource-intensive and often uncomfortable for patients, whereas home sleep apnea testing lacks the electroencephalography, electrooculography, and electromyography signals required for sleep staging and arousal scoring. We conducted a retrospective method-comparison study evaluating a reduced PSG montage that can be self-applied while preserving these signals and compared manual epoch-level scoring with conventional PSG. Concurrent recordings were analyzed from 102 adults with moderate to severe obstructive sleep apnea (244 studies) using two Nox A1s recorders simultaneously: one configured as full PSG, the other a reduced frontal EEG montage. Recordings were manually scored according to AASM criteria. Sensitivity, specificity, and accuracy were calculated for sleep stages (Wake, REM, N1, N2, N3) and arousal presence; intraclass correlation coefficients (ICCs) were computed for derived variables. Epoch-level agreement showed the following sensitivity/specificity/accuracy: Wake 84/95/92%; REM 78/99/96%; N1 48/94/88%; N2 76/87/82%; and N3 94/92/92%. For arousal detection, sensitivity was 69%, specificity 89%, and accuracy 85%. ICCs for total sleep time, sleep efficiency, and arousal index demonstrated high agreement between montages. The reduced montage demonstrated substantial agreement with conventional PSG for manual sleep stage and arousal scoring, comparable to PSG interrater reliability, and may enable reliable sleep assessment where full PSG is impractical.

Introduction

Full in-lab polysomnography (PSG), the gold standard sleep study, has long been a complex and cumbersome process. As a result, there has been growing demand for simpler, more patient-friendly approaches to sleep testing[1]. Home Sleep Apnea Testing (HSAT) devices have emerged as a popular alternative, offering the convenience of at-home testing with a much less intrusive setup including only a few channels[2, 3]. However, these devices do not record electroencephalography (EEG), electrooculography (EOG), and electromyography (EMG) signals, and therefore do not provide any direct measurement of sleep and arousal states[2, 3]. As a result, HSAT devices tend to underestimate downstream clinical parameters, such as the apnea-hypopnea index (AHI)[4].

Here, a reduced PSG montage is introduced, that can be applied by a clinical professional or self-applied by the patient and used in their home. This reduced setup aims to strike a balance between the comprehensive data collection of full in-lab PSG and the patient comfort and convenience of HSAT, which may also offer a more realistic assessment of sleep in the patient's own environment. The reduced PSG montage deviates from the standard montage for EEG/EOG/EMG recordings in a PSG sleep study, as set out in the Manual for Scoring of Sleep and Associated Events by the American Academy of Sleep Medicine (AASM)[5], by using only frontal EEG and temporalis EMG signals while maintaining the EOG electrode placements. Instead of mastoid references, as the AASM manual recommends, this reduced PSG montage uses averaged signals from channels close to the cheekbones for reference. These deviations in the reduced PSG montage result in EEG/EOG/EMG signals that are of lower amplitude and less power than recommended AASM PSG montage, but with similar spectral

morphology[6]. Such differences can potentially be corrected by a filter that amplifies parts of the signal, or by altering the scoring rules to accommodate differences in amplitude[7], but scoring based on these signals still requires validation to ensure clinical appropriateness.

Preliminary studies using the reduced PSG montage have shown technical feasibility in obtaining recordings, with highly comparable physiological signal types and signal quality to conventional PSG.[6, 7] The feasibility of self-application and use in the home setting has also been established through a large cohort study[8] and recent clinical trial[9]. Manual scoring of reduced PSG montage signals alone shows agreement between scorers for sleep stages in children[10] consistent with typical interrater reliability shown in the literature[11]. However, direct comparison of manual sleep stage and arousal scoring between the reduced PSG montage and conventional PSG in adults has not yet been examined.

Sleep stage and arousal scoring can have significant impacts on downstream clinical parameters, both by altering total sleep time and by changing the quantity or classification of respiratory events that, for example, end in arousal or occur during specific sleep stages. Hence the accuracy of these parameters is critical to accurate diagnosis of sleep disordered breathing, among other sleep conditions. This article therefore aims to evaluate epoch-level agreement of sleep stage classification and arousal scoring between manual scoring of the reduced PSG montage versus the conventional PSG setup recorded concurrently.

Methods

Study Design and Setting

Data was obtained from a retrospective cohort of adult participants diagnosed with moderate to severe (AHI 15–60) obstructive sleep apnea (OSA), but otherwise healthy, that underwent sleep testing as part of a clinical trial (SPRAY-SMART Trial, NCT04713826) conducted at three clinical sites: Berlin and Essen, Germany, and Zurich Switzerland. The sleep tests were conducted in-lab using two sleep recorders simultaneously, with one hooked up as full PSG with all necessary channels according to the AASM guidelines, and the other recording only the reduced frontal EEG/EOG/EMG montage. Participants came into the sleep clinic in the late afternoon or evening and, after completing the standard safety questionnaires and physical examination, were hooked up for both the reduced PSG montage and full conventional PSG recordings by a sleep technician. The same participants underwent up to 3 sleep studies each, separated by a period of at least 7 days. These data were originally collected by Bayer AG, then de-identified and transferred to Nox Medical for further analysis. All participants included in this cohort initially provided written informed consent to for their data to be used in research. This project was approved by the institutional review boards at Ruhrlandklinik (Essen), Advanced Sleep Research GmbH (Berlin) and by Swissmedic (Zurich) and was conducted in accordance with Good Clinical Practice guidelines and the Declaration of Helsinki.

To ensure generalizability of this dataset, the study population was intended to include both males and females, as well as individuals with a broad range of ages and BMIs, and varying AHI and PLMSI

categorizations. Sleep studies were excluded from analysis when the PSG recording time was < 4 hours, or if no manually scored epochs were present.

Polysomnography

Polysomnography was conducted according to the standard montage for a Type I Sleep Study as set out by the AASM[5]. A Nox A1s Recorder (Nox Medical, Iceland) was used to record this conventional PSG sleep data. This hookup included the Nox A1 EEG head cable and Nox A1 EEG 5 Lead Gold Cup Electrode Cable to collect EEG at F3/F4, C3/C4, O1/O2, and M1/M2, and EOG at E1/E2; 3 unipolar electrodes to collect submental EMG; as well as bi-polar snap on electrodes to collect two-lead ECG, and leg EMG; thoracic and abdominal RIP belts to monitor respiratory effort; a nasal cannula for monitoring airflow; the built-in microphone and triaxial accelerometer from the Nox A1s Recorder; and pulse oximetry using a Nonin WristOx2 Model 3150.

Reduced PSG Montage – the Nox SAS

A second Nox A1s Recorder was used to record signals from the reduced PSG montage. The Nox SAS consists of a set of electrodes and cables that connect to a Nox A1s Recorder and allow the recording of EEG, EOG, ECG, and EMG signals during sleep studies with a reduced frontal montage. As shown in Fig. 1, the Nox SAS head cable connects to 10 unipolar leads with 5 double snap-on electrodes placed over E2/E4, E1/E3, and AF3/AF7, AF4/AF8, with an AFz EEG channel with ground pair in the mid-forehead placed according to the standard 10–10 system for EEG recording[12]. The EOG channels (E1 and E2) are placed approximately 1cm above each eye and the E3 and E4 channels placed approximately 1cm below each eye, over the zygomatic bone to act as a reference instead of the typical mastoid channel. The full Nox SAS solution also includes two body cables (one right and one left). The Nox SAS body cables include 2 bipolar ECG channels, for recording 4-lead ECG, and 2 bipolar limb EMG channels, but these body cables were not used in the present study. As per conventional PSG, this reduced PSG montage setup can also be used with a nasal cannula, thoracic and abdominal respiratory inductance plethysmography (RIP) belts and pulse oximeter. However, in this study these accessory signals were only recorded by the primary device hooked up as full PSG. Therefore, the signals were synchronized with the reduced frontal montage recordings from the secondary device recording the Nox SAS for analysis.

Only the Zurich site used the Nox SAS Concept Cable, while the German sites used a collection of unipolar snap-on cables connected to the same electrode array. The solution is designed to be self-applied by the patient at home, with minimal assistance and guidance from video instructions[13]. However, in this dataset a double hook-up was performed in-lab by a trained sleep technician, to allow simultaneous recording of sleep studies with the reduced and conventional Nox A1s PSG montages.

Manual Sleep Stage and Arousal Scoring

Recordings from the reduced PSG montage and conventional PSG setup were anonymized and scored manually by one experienced sleep technologist with Noxturnal Software (v 6.1.0.29752, Nox Medical, Iceland) according to the 2012 AASM criteria, as detailed in the AASM Technical Manual[5]. In the case of reduced PSG montage studies, three adjustments were made to the sleep stage scoring criteria: 1. Reduction in EMG frontalis was not necessary to score REM; 2. Due to the lack of chin EMG tone (defined by the AASM as evidence of REM sleep), the process of scoring tonic REM differs from the AASM criteria, in that the first epoch of phasic REM, with low amplitude mixed frequency EEG and rapid eye movements should be scored as REM, and all previous epochs should also be identified as REM up until the epoch that presents any evidence requiring it to be scored as any other stage (N1, N2, N3 or Wake); 3. Stage N3 is scored when 20% or more of the epoch consists of delta waves which are < 2 Hz (duration at least 0.5 sec but no more than 2 sec) and have an amplitude greater than $60 \mu\text{V}$ (unlike the traditional $75 \mu\text{V}$ from the AASM scoring manual).

Sample Size Calculation

This retrospective study was a secondary analysis of sleep data collected during the SPRAY-SMART clinical trial, for which sample size was estimated and reported separately. Here we have conducted a theoretical sample size calculation to determine the minimum quantity of sleep studies needed to obtain appropriately powered diagnostic accuracy metrics. Therefore, this sample size was estimated based on the worst-case scenario with sensitivity and specificity both being 50%, and by assuming an $\alpha = 0.05$, a 95% confidence interval ($z = 1.96$) and a maximum 5% width for the confidence interval ($w = 0.05$). As sensitivity and specificity can be influenced by unbalanced data, in this case meaning the naturally differing proportions of sleep stages present in each sleep study, we used sample size formulas that take this imbalance into account[14]. The least prevalent sleep stage is wake, which is reported to be around 12% of total sleep time[15], hence a conservative worst-case percentage estimate for the output stage under examination of 10% was used ($p = 0.10$). Based on this calculation, the number of epochs required to reach the suggested sensitivity and specificity were 3,842 and 427 epochs, respectively. Assuming the minimum allowable study duration of 4 hours, this gives a minimum required sample of 81 sleep studies, with the actual number of studies included being 244.

Statistical Analysis

Statistical analysis of epoch-level agreement for sleep stages and arousals was conducted in Python (v3.11), while analysis of other variables was conducted in R studio (v2023.12.1 + 402). Demographic information for the sample, including sex, age and BMI are presented. Summary statistics of the clinically relevant parameters, AHI, AHI during REM sleep (REM-AHI), and Oxygen Desaturation Index (ODI) were analyzed per recording.

Sleep stage and arousal agreement were initially assessed using confusion matrices. The primary outcomes of sensitivity, specificity, and accuracy were then calculated for each sleep stage (Wake, REM, N1, N2 and N3) and for the presence of arousals (binary classification given for whether any arousal was

scored in the epoch: arousal, no arousal) in the manual scoring of the reduced montage recordings compared to PSG recordings. Overall observed agreement in sleep stage scoring was also calculated. As the placement of sleep stages and arousals impacts clinical indices, like AHI, derived from the sleep study, both by changing total sleep time and by determining which respiratory events are counted, epoch-level agreement was used. Results are presented as means with 95% confidence intervals, calculated using bootstrapping on the sleep study level with a random replacement iteration method and enough iterations to obtain consistent results (no less than 1000). For each iteration, N sleep studies were randomly selected from the sample of N studies and used to calculate sensitivity, specificity, and accuracy. The results therefore represent the means and confidence intervals of sensitivity, specificity, and accuracy across all iterations.

Additional sleep variables were calculated from sleep staging as follows: total sleep time (TST) as the sum of non-wake sleep epochs divided by two (time in minutes), sleep efficiency (SE) as the percentage of total recording time spent asleep (e.g. in REM, N1, N2 or N3), sleep latency (SL) as number of wake epochs to the first sleep epoch divided by two (time in minutes), REM latency (REM-L) as number of sleep epochs to the first REM epoch divided by two (time in minutes), wakefulness after sleep onset (WASO) as sum of all wake epochs between first and last sleep epoch divided by two (time in minutes), and time spent in each sleep stage (REM, N1, N2, N3) as sum of all epochs within each sleep stage divided by two (time in minutes). As there were several variables with outliers, two-tailed Wilcoxon-signed rank tests were used to compare median values between reduced montage and PSG scoring, with an alpha level of $P < 0.05$ considered significant. To investigate agreement between reduced montage and PSG scoring, intraclass correlation coefficients (ICC (2,1)) using a mixed-effects, single-rater design with absolute agreement definition were calculated using bootstrapping with replacement, and Bland-Altman plots showing the mean bias and 95% confidence intervals were generated for each variable.

In the supplementary material, agreement has also been calculated for studies collected with each of the different cable types (Nox SAS Concept Cables and unipolar cables). As well, for ease of comparison with previously reported agreement metrics in human scorers of sleep stages[11, 16, 17] and arousals[18], Cohen's kappa values for each sleep stage, and F1 score for the binary classification of arousal or no arousal, were calculated and presented in the supplementary material or discussion.

Results

Sample Characteristics

A total of 244 double hook-up sleep studies (N = 244 reduced PSG montage and N = 244 conventional PSG recordings), collected in 2021, were included. The total number of participants was N = 102, N = 34 at the Essen site (73 recordings), N = 33 at the Berlin site (93 recordings) and N = 35 at the Zurich site (78 recordings). It should be noted that multiple recordings (maximum 3) were performed on some individuals. As shown in Table 1, the sample varied in terms of age, gender, BMI, and OSA severity.

Table 1
Demographic and clinical characteristics of sample per patient and clinical indices per recording

Age (years)	Category	N (%)	Mean $\pm$ SD (Range)
	Missing	8 (7.8)	59.2 $\pm$ 12.3 (30–79)
	18–30	0 (0.0)	
	30–45	13 (12.7)	
	45–60	28 (27.5)	
	> 60	53 (52.0)	
Gender	Missing	6 (5.9)	-
	Male	73 (71.6)	
	Female	23 (22.5)	
BMI (kg/m ²)	Missing	8 (7.8)	31.1 $\pm$ 5.1 (21.1–49.0)
	Normal (< 25)	9 (8.8)	
	Overweight (25–30)	36 (35.3)	
	Obese ($\geq$ 30)	49 (48.0)	
AHI (events/hour)	Missing	0 (0.0)	33.4 $\pm$ 18.1 (3.5–87.9)
	None (< 5)	1 (0.4)	
	Mild (5–15)	36 (14.8)	
	Moderate (15–30)	82 (33.6)	
	Severe ($\geq$ 30)	125 (51.2)	
REM-AHI (events/hour)	Missing	2 (0.8)	40.3 $\pm$ 23.8 (0.0–97.1)
	None (< 5)	15 (6.1)	
	Mild (5–15)	29 (11.9)	
	Moderate (15–30)	45 (18.4)	
	Severe ($\geq$ 30)	153 (62.7)	
ODI ₃ (events/hour)	Missing	0 (0.0)	29.3 $\pm$ 16.7 (0.2–77.9)

Sleep Stage Scoring Agreement

The confusion matrix for epoch level sleep stage classification in the reduced PSG montage versus conventional PSG recordings is shown in Table 2. Overall observed agreement across all epochs was 75.3%. Bootstrapped mean sensitivity, specificity, and accuracy with 95% confidence intervals were

calculated for epoch level agreement of each sleep stage classification and are shown in Table 3. Sensitivity, specificity, accuracy, and Cohen's Kappa mean values were also calculated separately for sites using each of the different cable types and these results are shown in the Supplementary Material(see Supplementary Table S1).

Table 2

Confusion matrix of epoch-level sleep stage classifications between manual scoring of the reduced PSG montage and conventional PSG presented as N epochs (observed positive percentage agreement of reduced PSG montage scoring with conventional PSG).

Reduced PSG Montage						
PSG	Wake	REM	N1	N2	N3	Total N Epochs
Wake	42,797 (83.71%)	902 (1.76%)	3,444 (6.74%)	3,761 (7.36%)	222 (0.43%)	51,126
REM	2,792 (10.26%)	21,192 (77.89%)	1,666 (6.12%)	1,530 (5.62%)	27 (0.10%)	27,207
N1	4,492 (15.14%)	878 (2.96%)	14,202 (47.88%)	9,790 (33.00%)	302 (1.02%)	29,664
N2	1,145 (1.22%)	535 (0.57%)	5,032 (5.36%)	71,728 (76.38%)	15,470 (16.47%)	93,910
N3	7 (0.06%)	0 (0.00%)	4 (0.04%)	673 (6.15%)	10,252 (93.75%)	10,936

Table 3

Epoch level agreement of sleep stage classification and arousals between manual reduced PSG montage and conventional PSG scoring.

Class	No. of epochs	Sensitivity [95%CI]	Specificity [95%CI]	Accuracy [95%CI]
<i>Sleep Stage Classification</i>				
Wake	51,126	0.84 [0.82, 0.86]	0.95 [0.94, 0.95]	0.92 [0.91, 0.93]
REM	27,207	0.78 [0.75, 0.80]	0.99 [0.99, 0.99]	0.96 [0.96, 0.96]
N1	29,664	0.48 [0.46, 0.50]	0.94 [0.94, 0.95]	0.88 [0.87, 0.89]
N2	93,910	0.76 [0.75, 0.78]	0.87 [0.86, 0.88]	0.82 [0.81, 0.83]
N3	10,936	0.94 [0.92, 0.95]	0.92 [0.91, 0.93]	0.92 [0.91, 0.93]
<i>Arousal Scoring</i>				
Arousals	41,172/179,550	0.69 [0.67, 0.70]	0.89 [0.88, 0.89]	0.85 [0.85, 0.86]

Arousal Scoring Agreement

The confusion matrix for epoch level binary classification of arousal presence in the reduced PSG montage versus conventional PSG recordings is shown in Table 4.

Table 4
Confusion matrix of epoch-level manual scoring of arousals between the reduced PSG montage and conventional PSG presented as N epochs (observed positive percentage agreement of reduced PSG montage scoring with conventional PSG).

Reduced PSG Montage				
PSG		Arousal	No Arousal	Total N Epochs
	Arousal	28,399 (68.98%)	12,773 (31.02%)	41,172
	No Arousal	20,300 (11.31%)	159,250 (88.69%)	179,550

The bootstrapped mean sensitivity, specificity, and accuracy with 95% confidence intervals represent the epoch level agreement for epochs containing an arousal or not containing an arousal and are shown in Table 3. Sensitivity, specificity, and accuracy were also calculated separately for sites using each of the different cables, as shown in the Supplementary Material (see Supplementary Table S1).

Additional Sleep Variables

Analysis of additional sleep variables showed significant differences between manual scoring of the reduced montage compared to PSG for SL, REM latency and time spent in all 4 sleep stages ($P < 0.05$), but not for TST, WASO, or SE ($P > 0.42$, Table 5). The largest differences were seen for time spent in each non-REM sleep stage with especially N3 tending to be overscored in reduced PSG montage recordings compared to PSG.

Table 5

Mean $\pm$ SD values for the additional sleep variables from manual scoring of PSG and of the reduced PSG montage, along with mean differences (bias), Wilcoxon's V test statistic, P value and bootstrapped mean ICC and 95% confidence intervals.

TST (min)	N	PSG	Reduced Montage	Mean Bias	V	P	ICC [95%CI]
	244	331.4 $\pm$ 73.7	331.2 $\pm$ 69.8	0.22	14,527	0.96	0.93 [0.89, 0.95]
SE (%)	244	75.8 $\pm$ 12.4	76.0 $\pm$ 14.4	-0.41	14,287.5	0.79	0.87 [0.83, 0.89]
SL (min)	244	13.9 $\pm$ 14.0	14.9 $\pm$ 14.8	-0.97	5,574	< 0.01	0.92 [0.78, 0.97]
REM-L (min)	239	100.2 $\pm$ 54.0	104.3 $\pm$ 57.4	-3.70	8,268	0.01	0.85 [0.73, 0.92]
WASO (min)	244	83.0 $\pm$ 48.3	82.0 $\pm$ 51.2	0.91	15,562	0.43	0.86 [0.80, 0.90]
REM Time (min)	244	55.8 $\pm$ 25.2	48.2 $\pm$ 23.9	7.58	22,231	< 0.01	0.79 [0.67, 0.85]
N1 Time (min)	244	60.8 $\pm$ 30.0	49.9 $\pm$ 27.5	10.89	23,173	< 0.01	0.72 [0.64, 0.79]
N2 Time (min)	244	192.4 $\pm$ 53.5	179.3 $\pm$ 46.5	13.17	21,516.5	< 0.01	0.73 [0.65, 0.79]
N3 Time (min)	244	22.4 $\pm$ 25.2	53.8 $\pm$ 44.2	-31.43	224	< 0.01	0.49 [0.43, 0.55]
Arousal Index	244	30.5 $\pm$ 11.3	37.3 $\pm$ 12.2	-6.8	1,991	< 0.01	0.83 [0.79, 0.86]

Bland-Altman plots were used to further examine agreement between the reduced PSG montage and conventional PSG for the additional sleep variables (Fig. 2). No clear trends were observed in the distribution of error relative to mean scores, except in N3 sleep time. Here the differences are almost all negative (higher N3 sleep time in reduced montage) and in cases where no or very little N3 sleep time was scored on PSG, the mean is obviously linearly related to the difference as both values depend only on the reduced montage scoring.

Discussion

This article aimed to compare manual scoring of sleep stages and arousals from the reduced frontal montage (Nox SAS) to manual scoring of conventional PSG. High levels of agreement were observed for most sleep stage classifications and for arousal scoring, consistent with levels of agreement typically seen in the literature.

Agreement in sleep stage classification

Overall, substantial agreement in sleep stage scoring was observed. The highest levels of agreement between manual scoring of the reduced PSG montage and conventional PSG were observed for REM sleep, N3 sleep and wake classifications. This is consistent with prior literature where high interscorer agreement has been seen when comparing manual scoring of conventional PSG alone[11, 16, 17] (see Table 6 for direct comparison to prior literature). On conventional PSG recordings, these stages have clearly defined signal characteristics that make it easier for scorers to visually differentiate them. For example, the rapid eye movements and atonia observable in REM sleep, delta waves during N3 sleep or the low-amplitude high-frequency EEG waves during wake[5]. However, in the reduced montage, the typical submental EMG signal used primarily to score REM is absent and instead it is necessary to use surrogate signals from the temporalis muscle. The high level of agreement observed here is thus encouraging and supports that the reduced PSG montage can gather the necessary information to allow for sleep stage scoring using EEG and EMG signals from the forehead alone. Accuracy metrics were lower for N2 sleep and especially N1 sleep. However, even for these sleep stages, the accuracy of manually scored reduced PSG montage versus conventional PSG recordings was within a similar range or superior to the interscorer agreement on PSG alone[11, 16, 17] (see Table 6).

Table 6

Comparison of Cohen's Kappa (with 95% CI where available) for each sleep stage in the present analysis compared to other inter-scorer reliability analyses in the literature.

Class	Reduced PSG montage vs conventional PSG Results	Lee, et al. 2022[11] (Meta-analysis, PSG)	Nikkonen, et al. 2024[16] (Nox PSG)	Rosenberg & Van Hout 2013[17]* (PSG)
Wake	$\kappa = 0.81$ [0.80, 0.83]	0.70 [0.63, 0.77]	0.76	0.82
REM	$\kappa = 0.83$ [0.81, 0.84]	0.69 [0.58, 0.81]	0.86	0.90
N1	$\kappa = 0.49$ [0.48, 0.51]	0.24 [0.15, 0.33]	0.41	0.52
N2	$\kappa = 0.73$ [0.72, 0.75]	0.57 [0.54, 0.60]	0.67	0.72
N3	$\kappa = 0.53$ [0.50, 0.57]	0.57 [0.42, 0.71]	0.74	0.53
<i>*Cohen's kappa values were calculated from the confusion matrix given in the Rosenberg & Van Hout (2013) article for more direct comparability.</i>				

The lowest accuracy was observed in N1 sleep stage classification. This could be because the reduced PSG montage has less EEG channels available than the conventional PSG, making it harder to

discriminate between different sleep stages. However, stage N1 sleep can be difficult to discriminate even when using only conventional PSG recordings. Stage N1 is the first stage of non-REM sleep and is a transition period from wakefulness to deeper sleep stages, that is characterized by low amplitude, mixed frequency EEG activity, and slow eye movements. These characteristics sometimes overlap with those of wakefulness or other sleep stages, making it challenging to score accurately. In addition, the transitional nature of N1 sleep means that it is often only a brief period of a few epochs. Often discrepancies in scoring occur at the boundary between sleep stages (e.g. when one stage stops and another starts); hence, given the brevity of N1 sleep periods in typical sleep recordings, these boundaries between stages represent a much greater proportion of all N1 sleep epochs than they would for other stages.

Two conference abstracts have presented preliminary results like the present work, comparing sleep stage classification from the reduced PSG montage to full PSG[19, 20]. One of these was conducted in a pediatric population, showing both high agreement between manual scoring of frontal EEG and conventional PSG ($\kappa = 0.70$) and between different scorers of frontal EEG alone ($\kappa = 0.70$)[20]. The recent article detailing the full analysis further shows comparable levels of agreement between scorers on the reduced montage to that of PSG, and presents agreement metrics per sleep stage and for additional sleep indices of similar magnitude to the present work[10]. In the other, overall accuracy was shown for each sleep stage, ranging from 77.8% for stage N1 to 92.0% for REM, which is slightly lower than the present results, but also based on only a small sample of sleep studies[19].

In the largest available investigation of interscorer reliability by the AASM[17], sensitivity in N1 classification was 63%, compared to 48% in the present work. However, accuracy in the present work may also have been different if additional scorers were used. A more recent meta-analysis of interscorer reliability studies also found very low agreement in N1 classification[11], with Cohen's κ ranging from just 0.12 to 0.40 in the individual articles and an overall pooled estimate of 0.24, compared to 0.49 in the present work. Taken together, the present results show similar accuracy between reduced PSG montage and conventional PSG recordings for N1 scoring as observed in other interscorer reliability studies on PSG alone, suggesting that the variation in scoring may be more related to variability in the scorers' decisions rather than the differences in available signals. Others have shown that inter-rater reliability of sleep stage scoring can be improved by using automated pre-scoring algorithms[21], perhaps mostly by increasing concordance in the scoring of ambiguous epochs. Such automated pre-scoring was not used in the present work but is available in the Noxturnal software used for analysis, suggesting that higher agreement may have been observed if the use of automated scoring assistance were allowed.

Agreement in arousal detection

Arousal events are one of two main components contributing to hypopnea scoring and thus AHI calculation, hence their precise detection and scoring is of great clinical importance. Arousal scoring agreement in the present work was relatively high compared to prior studies, with a sensitivity of 69% and specificity of 89%. Here it is important to consider the natural imbalance in the data and how this

impacts agreement. There are clearly many more epochs without arousals than those with arousals, which means that overall accuracy and specificity of scoring may artificially appear high even if no arousals are ever scored. Naturally, specificity was higher than sensitivity in this analysis, as there are many more possibilities to correctly classify epochs without arousals, but importantly the sensitivity was also fair, and ICCs of arousal index were good, suggesting that scorers were able to identify arousals using signals from the reduced montage.

Arousal scoring has long been known to show high variability, especially for arousals not associated with respiratory events or movements[22]. The AASM's minimum 3 second definition has likely improved reliability of arousal scoring since its introduction in 2007, and scorer reliability can be further improved by continual proficiency testing[23]. However, the scoring rules still leave room for differing subjective interpretation and debate continues around the optimal definition and classification of arousals[24]. In a recent multicenter interscorer reliability paper on arousal scoring of PSG recordings, comparing 10 scorers from 7 different international sleep labs, the overall ICC was only 0.41[25]. When looking at pairs of scorers, there was also substantial variation from near no agreement to fair agreement, with better consistency noted between scorers from the same sleep center[25]. This reinforces the notion that arousal scoring remains subjective and supports that the agreement shown in the present work is highly relative within the field.

Differences in signals and scoring

As the reduced PSG montage does not include submental EMG, nor the classic mastoid reference for EEG/EOG channels, there are differences in the presence and morphology of signals as discussed in prior literature[6, 7]. For this reason, the scoring rules were adjusted in the present study to adapt which signals should be used for scoring each stage, and to amend rules that are specific to EEG amplitudes or frequencies. In the present work, the amended scoring rules and surrogate signals were sufficient to achieve agreement metrics consistent with usual inter-rater reliability. However, since this data was collected and scored, a filter has been developed, which intends to amplify parts of the reduced montage signal to make it appear more like conventional PSG. In short, the filter provides a linear combination of the original signal and two band-pass filtered versions of the original signal, the slow wave band (scaled by 0.5, pass band 0–2 Hz) and the alpha band (scaled by 3.0, pass band 8–13 Hz). In theory, it should also be possible to derive a predicted submental EMG signal from those recorded with the reduced montage, though this is yet to be developed. Such a filter, and derived submental EMG signal, should both make it easier to adopt the reduced PSG setup into clinical practice and allow for scoring according to the AASM Manual's scoring rules for conventional PSG, but the use of this filter is yet to be examined.

Additional Sleep Variables

Significant differences were observed between manual scoring of the reduced montage and for time spent in non-REM sleep stages, with overscoring of N3 and underscoring of N1 and N2 on the reduced PSG montage. Such a large difference in, e.g. N3 sleep time, is likely related to the scoring instructions

used for the reduced PSG montage, which have since been improved through the addition of filters. It should be noted that epoch-level agreement for N3 sleep was still very high (> 0.92), likely due to the low proportion of true N3 sleep epochs available to score. When adjusting for this low proportion of N3 sleep epochs, as done when calculating Cohen's kappa, the accuracy of N3 sleep scoring was considerably lower ($k = 0.53$). Further investigation of differences in non-REM sleep time when following the revised scoring guidelines and filter are needed, though preliminary evidence suggests that this has improved the consistency of scoring[20].

Although these differences in non-REM sleep time are not ideal, most misclassifications seen in the confusion matrix for these sleep stages are with other non-REM stages (e.g. N2 sleep was often misclassified as N3 sleep) and thus should not greatly impact e.g. respiratory or other event counts or TST. Of importance, other additional sleep variables, such as TST, SE, SL, and REM-L showed only small differences, if any, and all demonstrated good to excellent reliability on ICCs. These variables are often used clinically to characterize sleep quantity and quality, and in the case of TST, directly impact the calculation of AHI and other clinical parameters.

Reduced montages – comparison to other frontal montage devices

Several alternative EEG montages to the AASM guidelines have been investigated in the literature, with the majority using EEG electrodes that are placed on the forehead, face or near the ear. Such electrode placement is more convenient, as it avoids needing adhesives in the hair, and allows for easier self-application by participants making home sleep testing even more viable[26, 27]. Articles comparing sleep stage scoring from a single EEG channel to a full EEG montage have shown high levels of agreement both when using automatic or manual scoring[28–32], and of the single EEG channels available, frontal channels seem to show the best classification performance[29]. Much like the present work, articles using self-applied frontal montages have shown this approach to be feasible, produce reliable signals, and show good accuracy of sleep stage scoring compared to conventional PSG[33–35]. However, these articles also report a small number of additional recording failures due to misplacement of electrodes[33–35].

Limitations

In our sample size estimation, we determined 81 sleep studies were required based on the assumption that epoch classifications are independent. It is likely, however, that adjacent epoch classifications are correlated and thus the sample size may have been underestimated. In this analysis, we include significantly more data than the minimum required and therefore expect that the observed 95% confidence intervals in sensitivities, specificities and accuracies are still adequately powered. It should also be noted that some of the sleep studies included in this validation are from repeated assessments on the same individual (up to a maximum of 3 per participant), so data may not be strictly independent. However, this is unlikely to be problematic in the current context as it is generally well known that there is

high between-night variability in sleep within an individual and with the assessments here typically being at least a week apart. As noted in the methods, for technical reasons the German sites in this study used a collection of unipolar snap-on cables for the reduced PSG montage setup rather than the research version of the Nox SAS head cable used at the Zurich site. However, this difference in cables would not be expected to impact the results, as the electrode placement and thus signals obtained remain the same. Overall, sensitivity, specificity and accuracy, as shown in the supplementary material, was even slightly higher for the German sites, though it is not clear if this is related to the technical aspects of the hookup, or rather to the sample size, scorers, or other site-related factors.

Finally, the Nox SAS solution used as the reduced PSG montage in this study is designed to be self-applied by the patient and used in the patient's home environment, which it was not in the present study due to the double hook-up required. As mentioned above, self-application of electrodes may introduce slightly higher rates of electrode misplacement and subsequent study failure, as seen in initial cohort studies using the research version of the Nox SAS[8].

Future directions

Further research is needed to explore the accuracy of both automatic and manual scoring for reduced montage sleep studies performed in the home setting and with the reduced hook-up self-applied by the patient. As well, as mentioned previously, it is now possible to filter EEG signals from the reduced montage hook-up such that they are selectively amplified to appear more like conventional PSG signals. While this filtering should further improve the accuracy of sleep scoring based on the reduced montage, further investigation is needed for empirical confirmation.

Conclusion

The present work has demonstrated high levels of agreement between manual scoring of sleep stages and arousals in recordings from a reduced PSG montage compared to the gold standard conventional PSG recording. Specificity was typically higher than sensitivity and sleep stages with more easily identifiable indicators, such as REM and Wake, showed higher classification accuracy consistent with prior literature. This suggests that manual scoring of reduced PSG montage studies according to the AASM Manual is feasible and can be used to obtain accurate information on sleep staging and arousals. Accurate sleep stage and arousal scoring should allow for more accurate estimates of downstream clinical parameters than HSAT, though further studies are needed to evaluate the accuracy of AHI, PLMSI and other clinically relevant parameters scored from this reduced PSG montage.

Declarations

Competing Interests

M.M., K.M., P.B.S., S.Æ.J., E.A., E.F., N.M., R.H., & J.S.A. are full time employees of Nox Medical. At the time of the study G.W. and F.K. were full-time employees of BAYER Pharma AG, the sponsor of the SPRAY-SMART study.

Ethics approval and consent to participate:

The project was approved by the local ethical committees (BASEC-Nr. 2020–02810). All included participants provided written informed consent for their data to be used in research.

Consent for publication:

All authors reviewed and approved the final manuscript and consent to submitting the manuscript for publication.

Funding:

No funding was received in relation to the secondary analysis in this publication.

Clinical trial registration:

This article is a secondary analysis of data collected for the trial registered as EudraCT no. 2020-000520-19 and NCT04713826.

Author Contribution

M.M.: Conceptualization; Methodology; Formal analysis; Data curation; Visualization; Writing – original draft; Writing – review & editing. K.M.: Methodology; Software; Formal analysis; Data curation; Writing – review & editing. P.B.S.: Conceptualization; Methodology; Supervision; Writing – review & editing. S.Æ.J.: Software; Formal analysis; Data curation; Writing – review & editing. E.A.: Conceptualization; Resources; Project administration; Supervision; Writing – review & editing. E.F.: Methodology; Software; Validation; Writing – review & editing. N.M.: Data curation; Methodology; Writing – review & editing. R.H.: Resources; Project administration; Writing – review & editing. G.W.: Investigation; Resources; Writing – review & editing. F.K.: Investigation; Resources; Writing – review & editing. J.S.A.: Conceptualization; Supervision; Project administration; Writing – review & editing. All authors reviewed and approved the final manuscript.

Acknowledgement

The authors thank the participants and clinical staff involved in the SPRAY-SMART study for their contributions to data collection.

Data Availability

The data used in this study is proprietary and subject to contractual obligations that do not permit the authors to publicly share the data. All data sharing requests should be directed to the corresponding author.

References

1. Markun, L. C. & Sampat, A. Clinician-Focused Overview and Developments in Polysomnography. *Curr. Sleep. Med. Rep.* **6**, 309–321 (2020).
2. Berry, R. B. Uses and Limitations of Portable Monitoring for Diagnosis and Management of Obstructive Sleep Apnea. *Sleep. Med. Clin.* **6**, 309–333 (2011).
3. Go, B. C. & Thaler, E. R. Home Sleep Testing versus Traditional Polysomnography: Pros and Cons. *Otolaryngol. Clin. North Am.* S0030-6665(23)00198–6 (2023). 10.1016/j.otc.2023.11.003
4. Bianchi, M. T. & Goparaju, B. Potential Underestimation of Sleep Apnea Severity by At-Home Kits: Rescoring In-Laboratory Polysomnography Without Sleep Staging. *J. Clin. Sleep. Med.* **13**, 551–555 (2017).
5. Berry, R. B. et al. The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications, Version 2.6. *Darien, Illinois: American Academy of Sleep Medicine*; (2020).
6. Kainulainen, S. et al. Comparison of EEG Signal Characteristics Between Polysomnography and Self Applied Somnography Setup in a Pediatric Cohort. *IEEE Access.* **9**, 110916–110926 (2021).
7. Rusanen, M. et al. Self-applied somnography: technical feasibility of electroencephalography and electro-oculography signal characteristics in sleep staging of suspected sleep-disordered adults. *J. Sleep Res.* e13977 (2023).
8. Punjabi, N. M. et al. Methods for home-based self-applied polysomnography: the Multicenter AIDS Cohort Study. *Sleep. Adv. J. Sleep. Res. Soc.* **3**, zpac011 (2022).
9. Schweitzer, P. K., Maynard, J. P., Wylie, P. E., Emsellem, H. A. & Sands, S. A. Efficacy of atomoxetine plus oxybutynin in the treatment of obstructive sleep apnea with moderate pharyngeal collapsibility. *Sleep. Breath. Schlaf Atm.* 1–9. 10.1007/s11325-022-02634-x (2022).
10. Sigurdardottir, S. et al. Validation of manually scored multichannel frontal electroencephalography against polysomnography in a paediatric cohort. *J. Sleep. Res.* **e70012** 10.1111/jsr.70012 (2025).
11. Lee, Y. J., Lee, J. Y., Cho, J. H. & Choi, J. H. Interrater reliability of sleep stage scoring: a meta-analysis. *J. Clin. Sleep. Med. JCSM Off Publ Am. Acad. Sleep. Med.* **18**, 193–202 (2022).
12. Acharya, J. N., Hani, A., Cheek, J., Thirumala, P. & Tsuchida, T. N. American Clinical Neurophysiology Society Guideline 2: Guidelines for Standard Electrode Position Nomenclature. *J. Clin. Neurophysiol.* **33**, 308–311 (2016).
13. Nox SAS Hookup Instruction Video.

14. Buderer, N. M. F. & Statistical Methodology I. Incorporating the Prevalence of Disease into the Sample Size Calculation for Sensitivity and Specificity. *Acad. Emerg. Med.* **3**, 895–900 (1996).
15. Hertenstein, E. et al. Reference Data for Polysomnography-Measured and Subjective Sleep in Healthy Adults. *J. Clin. Sleep. Med.* **14**, 523–532 (2018).
16. Nikkonen, S. et al. Multicentre sleep-stage scoring agreement in the Sleep Revolution project. *J. Sleep. Res.* **33**, e13956 (2024).
17. Rosenberg, R. S. & Van Hout, S. The American Academy of Sleep Medicine Inter-scorer Reliability Program: Sleep Stage Scoring. *J. Clin. Sleep. Med.* **09**, 81–87 (2013).
18. Brink-Kjaer, A. et al. Automatic detection of cortical arousals in sleep and their contribution to daytime sleepiness. *Clin. Neurophysiol.* **131**, 1187–1203 (2020).
19. Punjabi, N. M., Kaplan, P. W., Margolick, J. & Aurora, R. N. 0319 A Simplified Bipolar Frontal Montage for Recording and Staging Sleep. *Sleep* **41**, A122–A122 (2018).
20. Sigurdardottir, S. et al. Multi-channel frontal EEG – validation on manual sleep staging in a pediatric cohort - Record details - Embase. in (Brazil. *Rio de Janeiro*. 10.1016/j.sleep.2023.11.749 (2023).
21. Alvarez-Estevéz, D. & Rijsman, R. M. Computer-assisted analysis of polysomnographic recordings improves interscorer associated agreement and scoring times. *PLoS ONE*. **17**, e0275530 (2022).
22. Loredó, J. S., Clausen, J. L., Ancoli-Israel, S. & Dimsdale, J. E. Night-to-Night Arousal Variability and Interscorer Reliability of Arousal Measurements. *Sleep* **22**, 916–920 (1999).
23. Ruehland, W. R., Rochford, P. D., Pierce, R. J., Singh, P. & Thornton, A. T. External proficiency testing improves inter-scorer reliability of polysomnography scoring. *Sleep. Breath.* **27**, 923–932 (2023).
24. Berry, R. B. et al. Should the Arousal Scoring Rule Be Changed? *J. Clin. Sleep. Med.* **11**, 497–499 (2015).
25. Pitkänen, H. et al. Multi-centre arousal scoring agreement in the Sleep Revolution. *J. Sleep. Res.* **e14127** 10.1111/jsr.14127 (2023).
26. Korkalainen, H. et al. Self-Applied Home Sleep Recordings. *Sleep. Med. Clin.* **16**, 545–556 (2021).
27. Myllymaa, S. et al. Assessment of the suitability of using a forehead EEG electrode set and chin EMG electrodes for sleep staging in polysomnography. *J. Sleep. Res.* **25**, 636–645 (2016).
28. Korkalainen, H. et al. Deep learning enables accurate sleep staging based on a single frontal EEG channel. *Sleep. Med.* **64**, S202–S203 (2019).
29. Radha, M., Garcia-Molina, G., Poel, M. & Tononi, G. Comparison of feature and classifier algorithms for online automatic sleep staging based on a single EEG signal. in *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* 1876–1880 (IEEE, Chicago, IL, 2014). 1876–1880 (IEEE, Chicago, IL, 2014). (2014). 10.1109/EMBC.2014.6943976
30. Lee, P. L. et al. Automatic Sleep Staging in Patients With Obstructive Sleep Apnea Using Single-Channel Frontal EEG. *J. Clin. Sleep. Med. JCSM Off Publ Am. Acad. Sleep. Med.* **15**, 1411–1420 (2019).

31. Light, M. P. 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. Schlaf Atm.* **22**, 1179–1188 (2018).
32. Lucey, B. P. et al. Comparison of a single-channel EEG sleep study to polysomnography. *J. Sleep. Res.* **25**, 625–635 (2016).
33. Hof, Z. et al. Portable PSG for sleep stage monitoring in sports: Assessment of SOMNOwatch plus EEG. *Eur. J. Sport Sci.* **20**, 713–721 (2020).
34. Miettinen, T. et al. Success Rate and Technical Quality of Home Polysomnography With Self-Applicable Electrode Set in Subjects With Possible Sleep Bruxism. *IEEE J. Biomed. Health Inf.* **22**, 1124–1132 (2018).
35. Levendowski, D. J. et al. The Accuracy, Night-to-Night Variability, and Stability of Frontopolar Sleep Electroencephalography Biomarkers. *J. Clin. Sleep. Med. JCSM Off Publ Am. Acad. Sleep. Med.* **13**, 791–803 (2017).

Figures

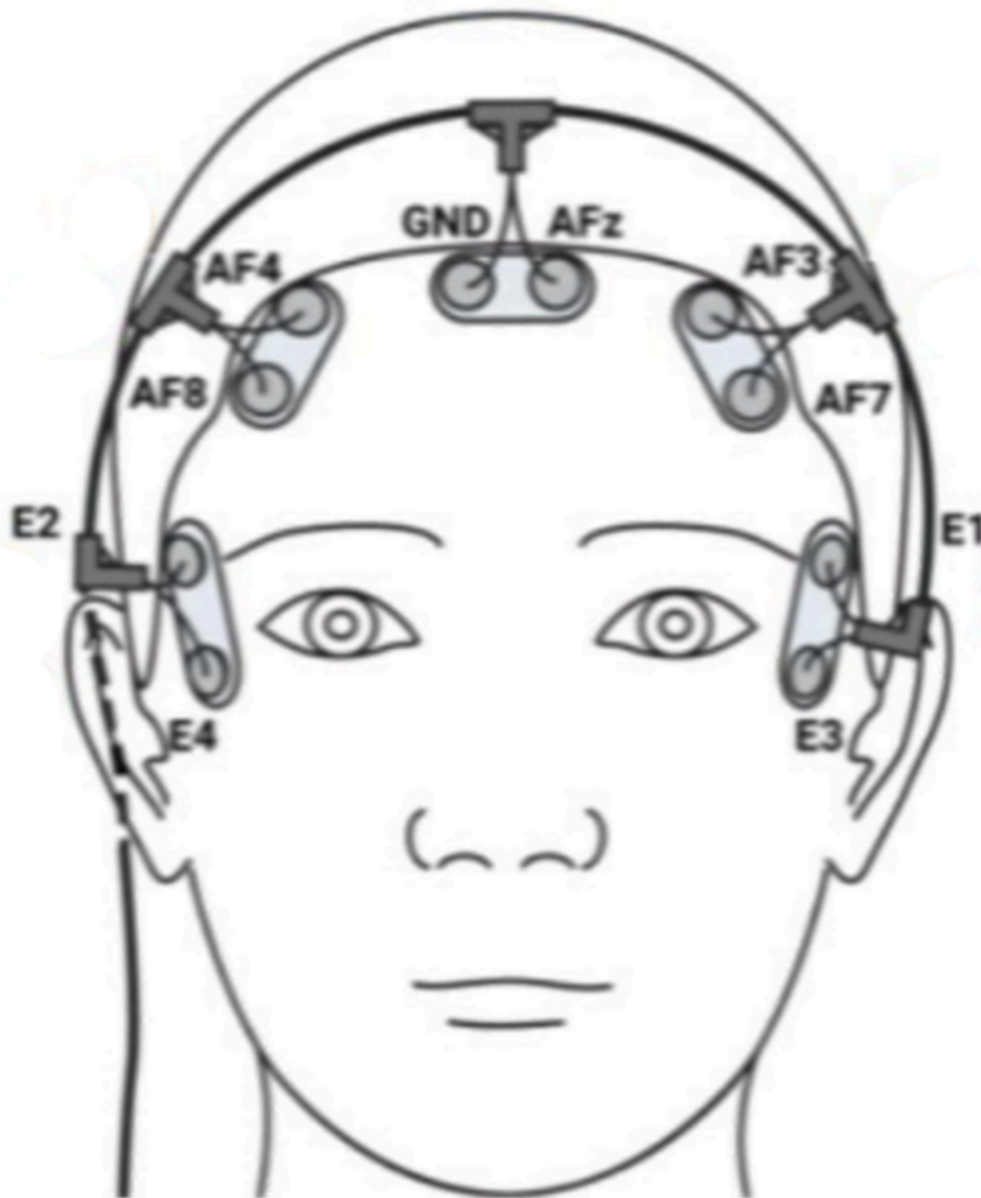


Figure 1

Schematic diagram of the reduced PSG montage (Nox SAS) hook-up showing the Nox SAS Concept Cable with electrode positions labelled.

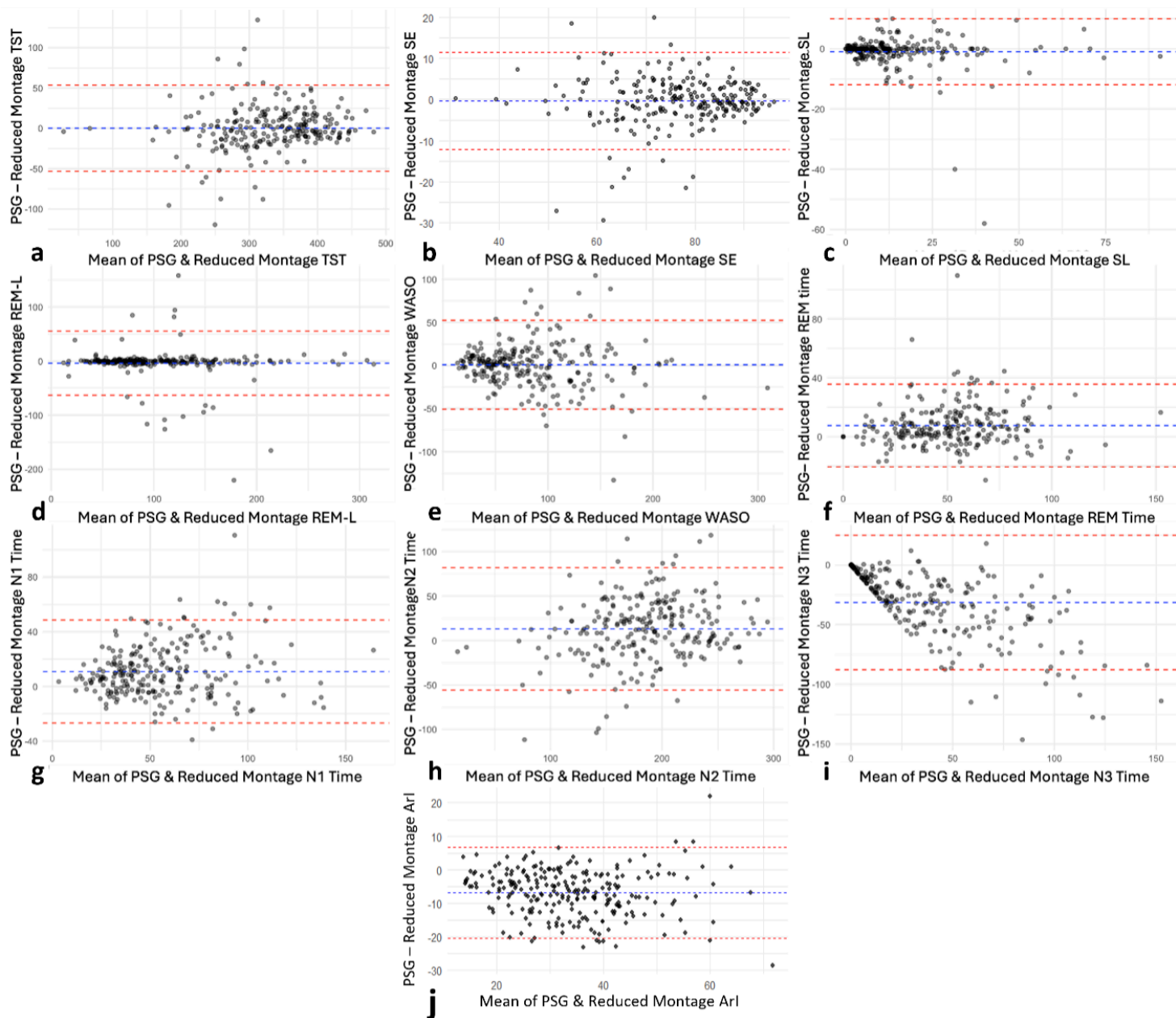


Figure 2

Bland-Altman plots of the difference between PSG and reduced montage scoring (y-axis) against the mean of PSG and reduced montage scoring (x-axis) for additional sleep variables, showing mean bias (blue) and 95% limits of agreement. The additional sleep variables include: (a) total sleep time, (b) sleep efficiency, (c) sleep latency, (d) REM latency, (e) wakefulness after sleep onset, (f) REM sleep time, (g) N1 sleep time, (h) N2 sleep time, (i) N3 sleep time, and (j) Arousal Index.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SASvsPSGValidationArticleSuppMaterial.docx](#)