Sleep Medicine 96 (2022) 14–19

Contents lists available at ScienceDirect

Sleep Medicine

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

Performance of Somno-Art Software compared to polysomnography interscorer variability: A multi-center study

Laurie Thiesse [a], Luc Staner [b], Gil Fuchs [a], Débora Kirscher [a], Valentin Dehouck [a], Thomas Roth [c], Jean-Yves Schaffhauser [a], Jay B. Saoud [d, e], Antoine U. Viola [a, *]

[a] PPRS, Colmar, France
[b] Unité d'exploration des rythmes veille sommeil, Centre Hospitalier de Rouffach, France
[c] Sleep Disorders Center, Henry Ford Hospital, Detroit, MI, USA
[d] PPRS Research Inc., Groton, MA, USA
[e] PPDA, LLC, Groton, MA, USA

ARTICLE INFO

Article history:
Received 6 October 2021
Received in revised form
11 March 2022
Accepted 18 April 2022
Available online 27 April 2022

Keywords:
Inter-rater reliability
Intra-rater reliability
Automatic sleep scoring

ABSTRACT

The visual scoring of gold standard polysomnography (PSG) is known to present inter- and intra-scorer variability. Previously, Somno-Art Software, a cardiac based sleep scoring algorithm, has been validated in comparison to 2 expert visual PSG scorers. The goal of this research is to evaluate the performances of the algorithm against a pool of scorers.

Sixty PSG and actimetry recording nights, representative of clinical practice (healthy subjects and patients suffering from obstructive sleep apnea [OSA], insomnia or major depressive disorder), were scored by 5 different sleep scoring centers and by the Somno-Art Software. Intra-class correlation coefficient (ICC) and Wilcoxon Signed-Rank Test were calculated between each scorer and the average value of the 6 scorers, including Somno-Art Software. In addition, epoch-by-epoch agreement between scorers were analyzed.

Somno-Art Software estimation of sleep efficiency, wake, N1+N2, N3 and REM sleep fit within the interscorer range for the full dataset and the subgroups, except for underestimating N3 sleep in OSA patients. Additionally, Somno-Art Software overestimated sleep latency compared to the average scoring for insomniacs (+4.7 ± 1.6min). On the full dataset, Somno-Art Software had good ($0.75 < \text{ICC} < 0.90$) or excellent ($\text{ICC} > 0.90$) ICC scores for all sleep parameters except N3 sleep (moderate score, $0.50 < \text{ICC} < 0.75$). For the 4-stages epoch-by-epoch agreement, Somno-Art Software was slightly below that of the visual scorers except for the healthy sub-group where an overlap was demonstrated.

Somno-Art Software sleep scoring shows a good interscorer reliability in the range of the 5 visual polysomnography scorers.

1. Introduction

Polysomnography (PSG) visual scoring is the gold standard for measuring and characterizing sleep continuity and architecture. However, this task is known to suffer from inter- and intra-expert variability [1,2]. Although the scoring rules are revised and specified on a regular basis, they are still not specific enough to prevent variability in scoring, the mean agreement between 2 visual scorers being around 83% [3,4]. In parallel, automatic scoring algorithm do not suffer from intra-scorer variability as the scoring is highly reproducible. These new generation of scoring algorithms have drastically improved in the last years and gained interest as a possible option, even alternative to analyze sleep in specific settings [5]. Somno-Art Software is 1 of these algorithms, analyzing sleep from cardiac (electrocardiogram [ECG]) and body movement signal (actimetry), physiological signals that have been shown to vary with sleep states [6]. Somno-Art Software performances have been evaluated in healthy subjects and patients suffering from obstructive sleep apnea (OSA), insomnia (ISM) and major depressive disorder (MDD) with promising performances in sleep stages characterization against PSG [7,8]. However, Somno-Art Software was compared to a maximum of 2 visual PSG scorers and as

[*] Corresponding author. PPRS 4E Avenue du General de Gaulle, 68000, Colmar, France.
E-mail address: avi@pprs-research.com (A.U. Viola).

https://doi.org/10.1016/j.sleep.2022.04.013
1389-9457/

previously mentioned, the interscorer reliability between visual scorers may impact the observed performances.

The goal of this publication is to evaluate the performances of Somno-Art Software in scoring sleep against a panel of visual scorers. To do so, 60 PSG recordings representative of clinical practice (healthy subjects and patients suffering from OSA, ISM or MDD) have been scored by 5 different international sleep scoring centers and by Somno-Art Software.

2. Methods

2.1. Study sample

To obtain a dataset representative of clinical practice, PSG recordings from 5 studies were randomly selected to be used as the dataset. All study protocols have been approved by institutional review boards in accordance with the Declaration of Helsinki and the guidelines on Good Clinical Practice. Written consent was obtained from all participants according to local requirements.

In total, 60 PSG recordings from 15 healthy, 15 OSA, 15 ISM and 15 MDD were included in the analysis.

2.2. Study design

2.2.1. Polysomnography

Multiple PSG recording systems were used in the various studies (Compumedics ProFusion PSG 3; Compumedics Siesta 802a [Compumedics, Abbotsford, Australia]) but all used the American Academy of Sleep Medicine (AASM) recommended electroencephalogram derivations (C4-M1, F4-M1, O2-M1), 2 electrooculogram electrodes, 2 chin electromyogram and 2 ECG electrodes. Five expert visual scorers (VS1-VS5) from 5 international sleep scoring centers (St Luke's Hospital, Chesterfield, USA; Concordia University, Montréal, Canada; The Siesta Group, Vienna, Austria; CIRCS research center, Strasbourg, France; PPRS, Colmar, France) scored the full dataset. All the scorers received identical information on the scoring nights, consisting only in the group characteristic (healthy, OSA, ISM, MDD). The expert scorer from the Siesta group was assisted by the Somnolyzer software [9]. Sleep scoring was performed according to the AASM rules [4]. The resulting reference classes were obtained by combining N1 and N2 into a single "N1+N2" class while the remaining classes (wake (W), N3, and REM) were used unchanged.

2.3. Somno-Art Software

The Somno-Art Software v.2.7.0 [3.2.0] analysis was performed on precisely synchronized actimetry and ECG signals. Using heart rate at a beat-to-beat resolution and actimetry data at a 1 Hz resolution, sleep stage classification (W, N1+N2, N3, REM) was performed at a 1-s epoch resolution. The latter 1-s epoch classification was merged into 30-s epochs in order to be compared to visual scoring. To do so, the dominant stage (or the first occurring stage, if they were equally represented) was selected. The sleep classification algorithm is based on expert rules associated to Support Vector Machine (SVM) detectors. Precise data processing methodology is described in Ref. [7].

2.4. Statistical analysis

Agreements between scorers on sleep parameters (sleep efficiency (SE), sleep latency (SL), W, N1+N2, N3 and REM sleep duration) were assessed with intra-class correlation coefficient absolute agreement, average measures, 2-way random model (ICC_{AAAvg}): The degree of absolute agreement for measurements that are averages based on k independent measurements on randomly selected objects [10]. Similarly to another multi-scorer study, the average sleep parameters of the 6 scorers, the 5 visual scorers and Somno-Art Software (Av6), were considered to represent the true values of the parameters for each recording [11]. ICC and difference of the means were calculated between each scorer and Av6. Koo et al. provides commonly cited cut-offs for qualitative ratings of agreement based on ICC values < 0.50: "poor" agreement; 0.50–0.75: "moderate" agreement; 0.75–0.90: "good" agreement; >0.90: "excellent" agreement [12]. Comparison of the means was achieved with non-parametric paired Wilcoxon Signed-Rank Test between each scorer and Av6. Significance was set at $p < 0.05$.

In addition, overall epoch-by-epoch accuracy (percentage of epochs labeled with the same sleep stage W, N1+N2, N3 or REM) was assessed for each scoring pair on the full data set and the subgroups (Healthy, OSA, ISM, MDD). As an illustration, 2 hypnograms have been selected for each subgroup (Healthy, OSA, ISM, MDD), in taking the recording nights with the most scoring pairs having the best and worst percentage agreement.

3. Results

Fig. 1 illustrates the difference of the means of each scorer (VS1-VS5 and Somno-Art) compared to Av6 for the main sleep parameters.

On the full dataset, Somno-Art Software showed a mean SE comparable to Av6 (difference of the mean: 0.1 ± 0.9%, NS) and comprised within the interscorer range (visual scorer range: −3 ± 0.6 to 1.7 ± 0.4%). Similar results were observed for the 4 sub-groups (healthy, OSA, ISM and MDD). In addition, ICC scores of SE on the full dataset was judged as excellent agreement for each scorer compared to Av6 (Table 1).

With a difference of the means of 1.7 ± 2.2 min on the full dataset, Somno-Art Software was comparable to the visual scorer range in the estimation of SL (visual scorer range: −2.2 ± 1.2 to 1.4 ± 0.6 min). Somno-Art Software estimation of SL was not significantly different from Av6 except for the ISM group with a difference of the means of 4.7 ± 1.6 min (visual scorer range: −3.7 ± 3.5 to 0.6 ± 0.8 min). Additionally, all scorers showed excellent ICC scores compared to Av6 on the full dataset.

For the full dataset, W results for the visual scorers showed a difference of means ranging from −8.2 ± 1.8 min to 14.1 ± 2.7 min compared to Av6. Somno-Art Software results overlapped with the range with a not significantly different mean (−0.6 ± 4.6 min). Somno-Art Software also overlapped with the interscorer range in the sub-groups analysis. For the full dataset, all visual scorers and Somno-Art Software showed excellent ICC scores compared to Av6.

For all analyzed groups, Somno-Art Software results were consistent with the visual scorer range in the estimation of N1+N2 sleep. For the full dataset where the visual scorer range of N1+N2 sleep was of −17.8 ± 4.8 to 36.0 ± 3.0 min, Somno-Art Software results showed an underestimate of 6.3 ± 4.7 min. The Wilcoxon test results were not significantly different between Somno-Art Software and Av6. ICC between each scorer and Av6 was judged as excellent or good for all scorers on the full dataset.

Differences in N3 sleep visual scoring for the full dataset ranged between −29.9 ± 1.9 to 19.6 ± 3.1 min. For this parameter, the Somno-Art Software mean was 1.6 ± 3.9 min and was within the interscorer range and was not significantly different from Av6. Similar results were observed for healthy, ISM and MDD subgroups. In the case of OSA patients, Somno-Art Software had the highest, but not significantly different mean score (12.8 ± 1.5 min; visual scorer range: −23.1 ± 3.5 min to 9.9 ± 6.2 min). The ICC score showed a moderate agreement between Somno-Art Software and Av6 on the full dataset, while the other scorers had good or

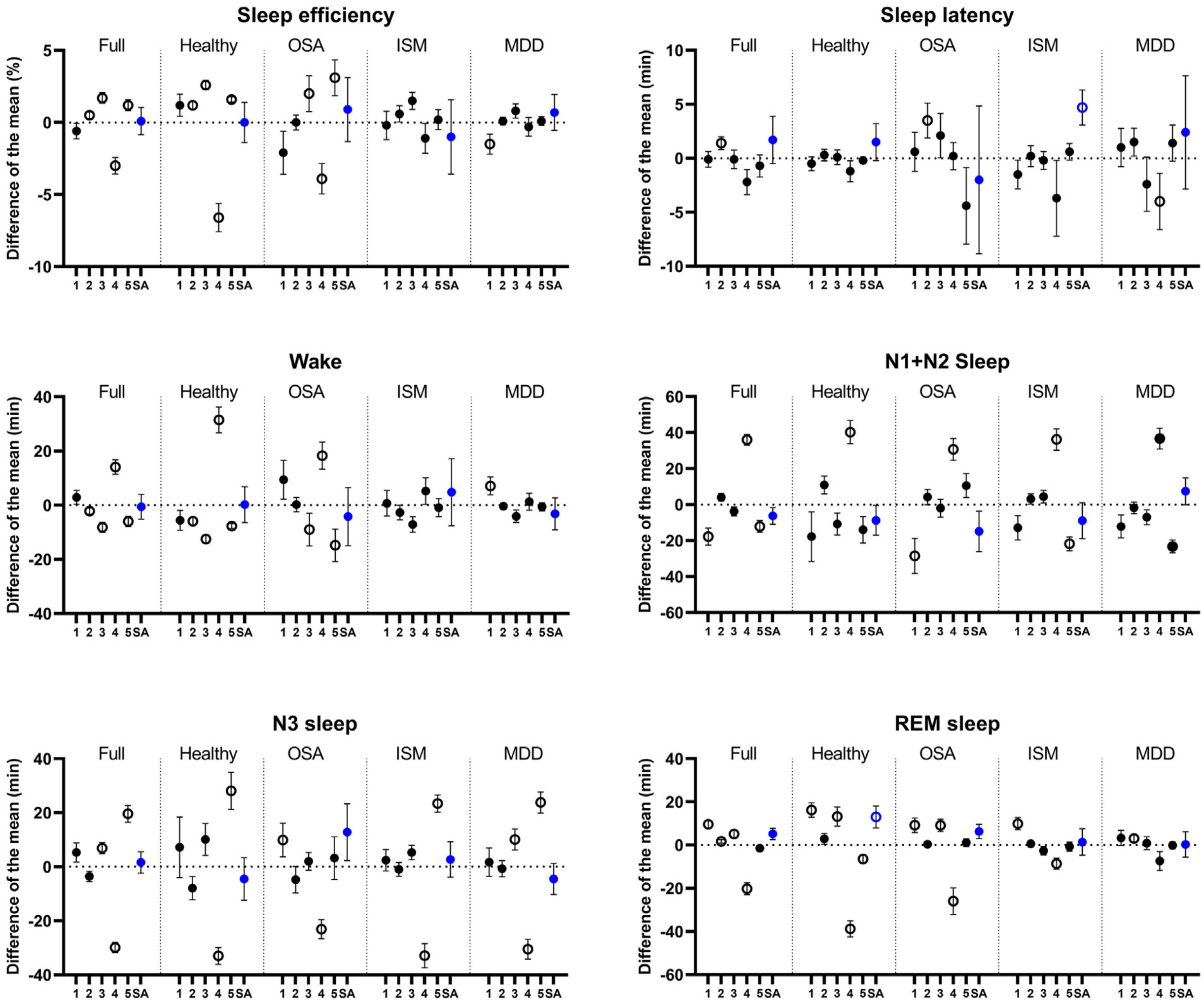

Fig. 1. Difference of the means (±SE) between each scorer (1–5: visual scorers 1–5; SA: Somno-Art Software) and the average of the 6 scorers for the full dataset and each subgroup (healthy, obstructive sleep apnea (OSA), insomnia (ISM), major depressive disorder (MDD)). Empty circles represent significant Wilcoxon Signed-Rank Test ($p < 0.05$); filled circles non-significant differences.

Table 1
Intra-class correlation coefficient of each scorer against the average of the 6 scorers (VS1-VS5: visual scorers 1–5; SA: Somno-Art Software) on the full dataset.

Scorer	Sleep efficiency (%)	Sleep latency (min)	Wake (min)	N1+N2 sleep (min)	N3 sleep (min)	REM sleep (min)
VS1	0.97	0.99	0.97	0.86	0.84	0.90
VS2	0.99	0.99	0.99	0.98	0.96	0.98
VS3	0.98	0.98	0.98	0.97	0.95	0.93
VS4	0.95	0.96	0.95	0.87	0.82	0.69
VS5	0.98	0.97	0.98	0.94	0.88	0.98
SA	0.90	0.90	0.90	0.86	0.70	0.80

excellent agreement.

Somno-Art Software results overlapped with the visual scorer range for REM sleep on the full dataset and on the subgroups (full dataset Somno-Art Software difference of the means: 5.2 ± 2.6 min; visual scorer range: −20.2 ± 2.7 to 9.6 ± 1.7 min). The ICC agreement was judged as good between Somno-Art Software and Av6, while visual interscorers agreements were generally excellent except for VS4 that was moderate.

Table 2 represents epoch-by-epoch percentage agreement for each scoring pair. For the full dataset, the lowest and highest agreement between visual scorers were between 78.5% and 88.7%, respectively. Somno-Art Software had an interscorer range between 69.9% and 71.2%. Similar ranges were observed for OSA, ISM and MDD subgroups. In the healthy sub-group, agreement ranged between 68.9% and 86.9% for visual scorers and between 62.0 and 71.4% for Somno-Art Software.

4. Discussion

Inter- and intra-scorer variability between visual PSG scorers is a known bias in sleep analysis [1,2]. However, it remains the standard method for scoring sleep recordings. Automatic sleep scoring algorithms have the advantage of not having intra-scorer variability as they are totally reproducible. Most automatic algorithms are

Table 2
Overall percentage agreement (W/N1+N2/N3/REM) between each pair of scorer (VS1-VS5: visual scorers 1–5; SA: Somno-Art Software).

Ful (n=60)

	VS2	VS3	VS4	VS5	SA
VS1	85.3%	84.5%	78.6%	83.8%	69.9%
VS2		88.2%	83.4%	88.7%	71.2%
VS3			79.1%	87.0%	70.6%
VS4				78.5%	67.0%
VS5					70.0%

Healthy (n=15)

	VS2	VS3	VS4	VS5	SA
VS1	82.6%	81.4%	70.0%	80.4%	70.4%
VS2		86.9%	75.1%	86.7%	71.4%
VS3			69.5%	85.7%	69.6%
VS4				68.9%	62.0%
VS5					68.5%

OSA (n=15)

	VS2	VS3	VS4	VS5	SA
VS1	83.5%	83.5%	78.3%	82.6%	69.2%
VS2		87.7%	85.0%	88.5%	69.5%
VS3			80.2%	86.1%	70.4%
VS4				80.9%	65.7%
VS5					69.3%

ISM (n=15)

	VS2	VS3	VS4	VS5	SA
VS1	88.0%	87.1%	83.2%	86.7%	69.2%
VS2		89.8%	87.2%	89.9%	70.8%
VS3			84.4%	88.6%	70.9%
VS4				82.6%	69.0%
VS5					70.9%

MDD (n=15)

	VS2	VS3	VS4	VS5	SA
VS1	87.2%	86.1%	83.0%	85.6%	70.6%
VS2		88.1%	86.5%	89.7%	73.2%
VS3			82.4%	87.5%	71.6%
VS4				81.5%	71.0%
VS5					71.3%

validated against a single visual scorer, which may present a bias taking the interscorer variability into consideration. This paper aims at evaluating the Somno-Art Software, based on cardiac and body activity, against a pool of expert visual scorers. Somno-Art Software estimation of SE, W, N1+N2, N3 and REM sleep belonged to the interscorer range for the full dataset and the investigated subgroups (healthy, OSA, ISM and MDD). One exception concerned the OSA subgroup with a non-significant overestimation of N3 sleep of 12.8 ± 10.5 min, while the interscorer range was between −23.1 ± 3.5 to 9.9 ± 6.2 min. However, some visual scorers underestimated this sleep stage even more than Somno-Art Software overestimated it. Interestingly, full dataset N3 sleep showed an important interscorer variability (−29.9 ± 1.9 to 19.6 ± 3.1 min). Sleep stage N3 is known to present a high inter-rater variability between visual scorers. This is generally due to the complexity in characterizing slow waves (SW) duration and amplitude. In contrast, Somno-Art Software overestimated N3 sleep by less than 2 min on the full dataset (1.6 ± 3.9 min). Being an automatic algorithm, Somno-Art Software is consistent in its definition of SW and may therefore yield a more accurate and reproducible results.

Somno-Art Software tended to overestimate SL in healthy and MDD sub-groups and significantly overestimate SL in ISM (4.7 ± 1.6 min; interscorer range: −3.7 ± 3.5 to 0.6 ± 0.8 min). However, an overestimation of less than 5 min on a mean SL of 44 min is not considered as clinically relevant. To note, long periods of wake-after-sleep-onset, which are present in ISM, are well detected with Somno-Art Software, as illustrated on Fig. 2 (best agreement of the ISM sub-group).

On the full dataset, Somno-Art Software had good or excellent ICC scores for all sleep parameters analyzed except for N3 sleep (moderate score). However, ICC scores of SE, N3 and REM sleep were above the 1 reported in the literature for the validation of an automatic PSG scoring algorithm [13].

4-stages epoch-by-epoch agreement ranged between 78.5 and 88.7% within visual scorers and 69.9 and 71.2% for Somno-Art Software. On the healthy sub-group, visual scorers agreed between 68.9 and 86.9%, while Somno-Art Software overlapped this range with an interscorer range of 62.0–71.4%. Fig. 2 illustrates Somno-Art Software's tendency to fragment less sleep than the visual scorers, a characteristic which could explain the lower epoch-by-epoch agreement observed for Somno-Art Software. On the other hand, reduced fragmentation increases intra-scorer reliability which is of interest for longitudinal studies.

It is to note that Fig. 2 displays some misclassifications between N3 and REM sleep with Somno-Art Software (Fig. 2, worst agreement of ISM and MDD patients). However, as reported in previous validation studies of Somno-Art Software, this kind of misclassification is rare: REM sleep to N3 sleep misclassification was lower than 1.5%, while N3 sleep to REM sleep was lower than 2% on a dataset of 458 recording nights with healthy and pathological sleep profiles [7,8].

This paper shows good inter-scorer reliability of Somno-Art Software compared to PSG visual scorers despite the use of different raw signals (cardiac and wrist activity vs PSG leads). Unlike visual scoring, Somno-Art Software has the advantage to have perfect intra-scorer reliability and can therefore be of interest in studies with repetitive measures, where a drift in the scoring can become critical (i.e. pharmacology, research, treatment follow-up) [14]. Moreover, the analysis of sleep through Somno-Art Software takes less than 1 min while a visual scorer needs several hours depending on the complexity of the night.

In conclusion, Somno-Art Software performance demonstrated reliable scoring over various scoring centers and is in the PSG inter-

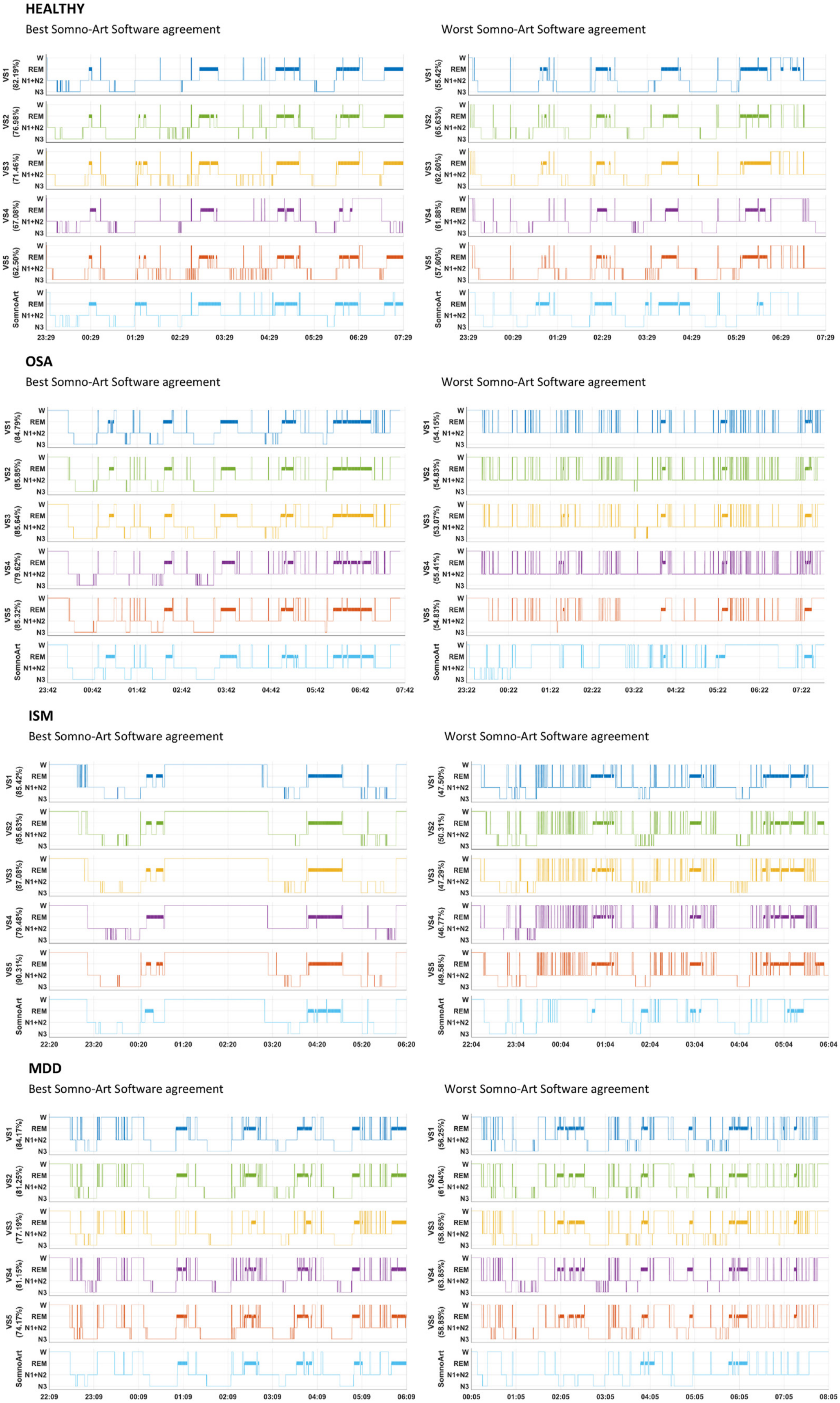

Fig. 2. Hypnograms obtained from the 6 different scorers. For each subgroup the best (left column) and the worst (right column) agreement between Somno-Art Software and the majority of visual scorers are represented.

scorer range of agreement for most of the sleep parameters investigated.

Declaration of competing interest

This study was supported by PPRS SAS. Jean-Yves Schaffhauser and Jay B. Saoud are shareholders of PPRS SAS. Laurie Thiesse, Gil Fuchs, Valentin Dehouck, Débora Kirscher and Antoine U. Viola, are employees of PPRS SAS. Jay B. Saoud is a payed consultant of PPRS Research lnc. Jean-Yves Schaffhauser is shareholder of V-WATCH SA.

Acknowledgments

We would like to thank the sleep experts of the 5 sleep centers (St Luke's Hospital, Chesterfield, USA; Concordia University, Montréal, Canada; The Siesta Group, Vienna, Austria; CIRCS research center, Strasbourg, France; PPRS, Colmar, France) for having scored the full dataset. We would like to thank Alain Muzet, Bruno Muller, Julien Barascud and Régis Lengellé for their help in developing the Somno-Art Software algorithm.

Some studies used data obtained from clinical trials sponsored by Janssen and Minerva Neurosciences, Inc. The interpretation and reporting of research using these data are solely the responsibility of the authors and does not necessarily represent the official views of Janssen and Minerva.

References

[1] Penzel T, Zhang X, Fietze I. Inter-scorer reliability between sleep centers can teach us what to improve in the scoring rules. J Clin Sleep Med 2013;9(1): 89–91. https://doi.org/10.5664/jcsm.2352.

[2] Zhang X, Dong X, Kantelhardt JW, Li J, Zhao L, Garcia C, Glos M, Penzel T, Han F. Process and outcome for international reliability in sleep scoring. Sleep Breath 2015;19(1):191–5. https://doi.org/10.1007/s11325-014-0990-0.

[3] Rosenberg RS, Van Hout S. The American Academy of Sleep Medicine interscorer reliability program: sleep stage scoring. J Clin Sleep Med 2013;9(1): 81–7. https://doi.org/10.5664/jcsm.2350.

[4] Iber C, Ancoli-Israel S, Chesson AL, et al. The AASM manual for the scoring of sleep and associated events: rules, terminology, and technical specifications. 1st ed. Westchester, Illinois: American Academy of Sleep Medicine; 2007.

[5] Depner CM, Cheng PC, Devine JK, Khosla S, de Zambotti M, Robillard R, Vakulin A, Drummond SPA. Wearable technologies for developing sleep and circadian biomarkers: a summary of workshop discussions. Sleep 2020;43(2). https://doi.org/10.1093/sleep/zsz254.

[6] Viola AU, Simon C, Ehrhart J, Geny B, Piquard F, Muzet A, Brandenberger G. Sleep processes exert a predominant influence on the 24-h profile of heart rate variability. J Biol Rhythm 2002;17(6):539–47. https://doi.org/10.1177/0748730402238236.

[7] Muzet A, Werner S, Fuchs G, Roth T, Saoud JB, Viola AU, Schaffhauser JY, Luthringer R. Assessing sleep architecture and continuity measures through the analysis of heart rate and wrist movement recordings in healthy subjects: comparison with results based on polysomnography. Sleep Med 2016;21: 47–56. https://doi.org/10.1016/j.sleep.2016.01.015.

[8] Thiesse L, Staner L, Bourgin P, Roth T, Fuchs G, Kirscher D, Schaffhauser J-Y, Saoud JB, Viola AU. Validation of Somno-Art Software, a novel approach of sleep staging, compared with polysomnography in disturbed sleep profiles. SLEEP Advances 2021;3(1). https://doi.org/10.1093/sleepadvances/zpab019.

[9] Anderer P, Moreau A, Woertz M, Ross M, Gruber G, Parapatics S, Loretz E, Heller E, Schmidt A, Boeck M, Moser D, Kloesch G, Saletu B, Saletu-Zyhlarz GM, Danker-Hopfe H, Zeitlhofer J, Dorffner G. Computer-assisted sleep classification according to the standard of the American Academy of Sleep Medicine: validation study of the AASM version of the Somnolyzer 24 x 7. Neuropsychobiology 2010;62(4):250–64. https://doi.org/10.1159/000320864.

[10] McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods 1996;1(1):30–46. https://doi.org/10.1037/1082-989X.1.1.30.

[11] Younes M, Kuna ST, Pack AI, Walsh JK, Kushida CA, Staley B, Pien GW. Reliability of the American Academy of sleep medicine rules for assessing sleep depth in clinical practice. J Clin Sleep Med 2018;14(2):205–13. https://doi.org/10.5664/jcsm.6934.

[12] Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016;15(2):155–63. https://doi.org/10.1016/j.jcm.2016.02.012.

[13] Malhotra A, Younes M, Kuna ST, Benca R, Kushida CA, Walsh J, Hanlon A, Staley B, Pack AI, Pien GW. Performance of an automated polysomnography scoring system versus computer-assisted manual scoring. Sleep 2013;36(4): 573–82. https://doi.org/10.5665/sleep.2548.

[14] Berthomier C, Muto V, Schmidt C, Vandewalle G, Jaspar M, Devillers J, Gaggioni G, Chellappa SL, Meyer C, Phillips C, Salmon E, Berthomier P, Prado J, Benoit O, Bouet R, Brandewinder M, Mattout J, Maquet P. Exploring scoring methods for research studies: accuracy and variability of visual and automated sleep scoring. J Sleep Res 2020;29(5):e12994. https://doi.org/10.1111/jsr.12994.