Received: 20 May 2021 | Accepted: 30 May 2021

DOI: 10.1111/jsr.13424

Journal of Sleep Research

RESEARCH ARTICLE

Variability of sleep stage scoring in late midlife and early old age

Daphne Chylinski[1] | Christian Berthomier[2] | Eric Lambot[1] | Sonia Frenette[3,4] | Marie Brandewinder[2] | Julie Carrier[3,4] | Gilles Vandewalle[1] | Vincenzo Muto[1,5]

[1]GIGA-Cyclotron Research Centre-In Vivo Imaging (CRC-IVI), University of Liège, Liège, Belgium

[2]PHYSIP, Paris, France

[3]Centre for Advanced Research in Sleep Medicine (CARSM), CIUSSS of Nord-de l'Île-de-Montréal, Montreal, QC, Canada

[4]Department of Psychology, University of Montreal, Montreal, QC, Canada

[5]Psychology and Cognitive Neuroscience Research Unit, University of Liège, Liège, Belgium

Correspondence
Vincenzo Muto and Gilles Vandewalle, GIGA-Cyclotron Research Centre-In Vivo Imaging, Bâtiment B30, Allée du Six Août, 8, 4000- Liège, Belgium.
Email: vincenzo.muto@uliege.be (V.M.); gilles.vandewalle@uliege.be (G.V.)

Funding information
H2020 European Research Council, Grant/Award Number: GA 757763; Canadian Institutes of Health Research, Grant/Award Number: 190750; Fonds National de la Recherche Scientifique, Grant/Award Number: T.0242.19; Actions de Recherche Concertées, Grant/Award Number: 17/27-09

Abstract

Sleep stage scoring can lead to important inter-expert variability. Although likely, whether this issue is amplified in older populations, which show alterations of sleep electrophysiology, has not been thoroughly assessed. Algorithms for automatic sleep stage scoring may appear ideal to eliminate inter-expert variability. Yet, variability between human experts and algorithm sleep stage scoring in healthy older individuals has not been investigated. Here, we aimed to compare stage scoring of older individuals and hypothesized that variability, whether between experts or considering the algorithm, would be higher than usually reported in the literature. Twenty cognitively normal and healthy late midlife individuals' (61 ± 5 years; 10 women) night-time sleep recordings were scored by two experts from different research centres and one algorithm. We computed agreements for the entire night (percentage and Cohen's κ) and each sleep stage. Whole-night pairwise agreements were relatively low and ranged from 67% to 78% (κ, 0.54–0.67). Sensitivity across pairs of scorers proved lowest for stages N1 (8.2%–63.4%) and N3 (44.8%–99.3%). Significant differences between experts and/or algorithm were found for total sleep time, sleep efficiency, time spent in N1/N2/N3 and wake after sleep onset ($p \leq 0.005$), but not for sleep onset latency, rapid eye movement (REM) and slow-wave sleep (SWS) duration (N2 + N3). Our results confirm high inter-expert variability in healthy aging. Consensus appears good for REM and SWS, considered as a whole. It seems more difficult for N3, potentially because human raters adapt their interpretation according to overall changes in sleep characteristics. Although the algorithm does not substantially reduce variability, it would favour time-efficient standardization.

KEYWORDS
aging, automatic scoring, electroencephalography, inter-expert variability

1 | INTRODUCTION

Sleep stage scoring is a first, crucial step in analysing a sleep electroencephalogram (EEG) recording, whether in research settings or clinical practice. As it is partly subject to expert interpretation, it can carry substantial inter-expert variability. Although the algorithm for automated scoring methods is becoming increasingly available, time-consuming human expert sleep stage scoring remains the reference method.

Previous studies on inter-expert agreement have reported average kappa values showing almost perfect agreement (κ ~ 0.82) after prior

Gilles Vandewalle and Vincenzo Muto contributed equally to this study.

training aiming to standardize the interpretation of the scoring rules in the same research centre (Whitney et al., 1998). However, lower agreement is reached when comparing scorers across different centres, with mean kappa values ranging from moderate (κ = 0.57; Zhang et al., 2015) to good agreement (κ = 0.72–0.76; Basner, Griefahn, & Penzel, 2008; Danker-Hopfe, Anderer, & Zeitlhofer, 2009). Several factors may lead to lower inter-expert agreement, including pathologies affecting sleep quality, such as periodic limb movements syndrome and obstructive sleep apnoea, and also depression or neurodegenerative diseases such as Parkinson's disease (Danker-Hopfe et al., 2004). Agreement has also been reported to vary between sleep stages, with lower agreement in non-rapid eye movement (NREM) 1 sleep (N1; Basner et al., 2008; Rosenberg & Van Hout, 2013) and slow-wave sleep (SWS; Do Kim, Kurachi, Horita, Matsuura, & Kamikawa, 2007; Monroe, 1969). Lastly, increasing the number of scorers being compared decreases the number of epochs upon which they all agree.

It is now well established that sleep changes across the lifespan, with regards to its architecture, but also its fine microstructure. These changes are detectable starting at age 40 (Carrier et al., 2011). Studies have reported that advancing in age is associated with an increased prevalence of N1 and N2 sleep stages, as well as a diminution of time spent in SWS and rapid eye movement (REM) sleep (Březinová, 1975; Ohayon, Carskadon, Guilleminault, & Vitiello, 2004; Redline et al., 2004). In addition, frequent awakenings may lead to more frequent N1 transitions, which can be strenuous to score. At a finer level, aging has been related to a drop of EEG power and amplitude, particularly for the delta band (0.5–4 Hz) in anterior derivations (Gaudreau, Carrier, & Montplaisir, 2001; Landolt & Borbély, 2001; Silber et al., 2007), and of the density (number per minutes of NREM sleep) and amplitudes of detected slow waves (SW), (Carrier et al., 2011). These changes are considered to arise from a decrease in the build-up of sleep need during wakefulness (Cajochen, Münch, Knoblauch, Blatter, & Wirz-Justice, 2006; Gaudreau, Morettini, Lavoie, & Carrier, 2001), associated with a decrease in the need for sleep (Klerman & Dijk, 2008), as well as a reduction in neurite density (Pannese, 2011) and/or cortical thickness (Dubé et al., 2015). A recent investigation showed that taking into account the reduced amplitude of sleep slow waves does not abolish age-related difference in SW density, and allows inclusion of SWs of lower amplitude that are seen in increased amounts in older individuals (Rosinvil, Bouvier, & Dubé, 2020). The impact of age on inter-expert agreement remains scarcely investigated (Danker-Hopfe et al., 2009). One study that compared visual scoring in older individuals without using a fixed amplitude threshold for considering an oscillation as a slow wave showed good agreement in scoring for deep SWS (N3 and N4 sleep stages) (Webb & Dreblow, 1982). Yet, the American Academy of Sleep Medicine (AASM) recommends that scoring the N3 stage should be based on the percentage of slow oscillations reaching at least a 75 μV amplitude (Iber et al., 2007). Whether scoring agreement is good when applying (or trying to apply) AASM rules is not established (Berry, Brooks, & Gamaldo, 2017).

The aim of this study was to assess inter-expert variability of visual and automatic sleep stage scoring in a population composed of healthy individuals in late midlife/early old age (i.e., between 50 and 70 years when sleep is significantly altered by age, although potentially moderately). We compared the scorings of three raters, two experts from different research centres and an automatic sleep scoring algorithm that was previously validated in healthy young individuals and across several sleep disorders (Berthomier, Drouot, & Herman-stoïca, 2007; Peter-Derex et al., 2020). We hypothesised that given the changes in EEG density/amplitude associated with aging, agreement would prove lower than what is usually reported in the literature for healthy young individuals, whether between experts or between experts and algorithm scoring, particularly over the N3 stage.

2 | METHODS

2.1 | Dataset

Twenty sleep EEG recordings from healthy individuals in late midlife/early older age (61 ± 5 years; 10 women) were analysed in this study. The sample size was determined based on a prior sensitivity analysis, which indicated that with a sample of N = 20, with a power of 0.8 and alpha of 0.05 in a multiple regression scheme, we were in a position to detect medium effect (r > 0.28; R^2 > 0.078), which we considered as likely prior to starting the study (as assessed using Gpower 3.1.9.4) (Faul, Erdfelder, Buchner, & Lang, 2009). Exclusion criteria were: self-reported sleep or cognitive complaints, smoking, intake of medication affecting the sleep–wake cycle or the central nervous system, a chronic medical condition that may affect sleep (e.g. pain) as assessed via an in-house questionnaire and semi-structured interview, and shift work or transmeridian travel in the 3 months preceding participation in the study. An extensive expert neuropsychological assessment allowed to rule out any potential cognitive impairment. Before proper data acquisition, an adaptation and screening night under polysomnography was performed, and volunteers with apnoea/hypopnoea index >10/h or periodic limb movements of sleep associated with an arousal >10/h were further excluded. All participants signed a written consent form and the study was approved by the Hospital of Sacré-Coeur de Montréal ethical committee.

Sleep recordings were acquired using a Grass Model 15A54 amplifier system. The sampling rate was set at 256 Hz, gain at 10.000, and hardware filters were set at 0.3 Hz and 100 Hz for EEG, with a notch filter at 60 Hz. Twenty EEG derivations were placed according to the 10–20 system, referenced against the mean of the two mastoids, as well as submental electromyogram (EMG) and electrooculogram (EOG) bipolar channels.

2.2 | Algorithm scoring (ALGO)

Aseega (PHYSIP, Paris, France) is an algorithm for automatic sleep scoring that relies on the analysis of a single EEG bipolar channel, either Cz-Pz or C4-O2, without further input from either EMG or

TABLE 1 Contingency matrix between the consensus of experts and algorithmscoring for the whole dataset

		Expert consensus (EXPERT_CON)						
Number of epochs (whole dataset)		Artefact or disagreement	W	REM	N1	N2	N3	Total
Algorithm (ALGO)	Artefact	77	17	151	39	86	1	371
	W	655	1757	105	360	170	2	3049
	REM	651	45	2244	192	232	0	3364
	N1	160	35	96	142	113	1	547
	N2	2188	102	136	443	6162	516	9547
	N3	789	3	0	0	516	792	2100
	Tot	4520	1959	2732	1176	7279	1312	18978

Abbreviation: N1/N2/N3, non-REM 1/2/3; REM, rapid eye movement sleep; W, wakefulness.

EOG channels. The detailed procedure can be found in Berthomier et al. (2007). Sleep recordings were analysed twice independently, first using Cz-Pz and then C4-O2, thus providing two sets of automatic hypnograms. Cz-Pz results can be found in the results section, whereas C4-O2 analysis can be found in the supplementary materials (boxplot of kappa and percentage of agreement in Figure S1, contingency matrix in Table S1, and agreement coefficient for comparisons with the algorithm [sensitivity, specificity, positive and negative predictive values; see Statistical analysis] in Table S2).

2.3 | Expert scoring

Visual scoring was performed independently by two human sleep experts, EXPERT_1 and EXPERT_2, one at the CARSM in Montréal, Canada, and one at the GIGA-CRC-IVI in Liège, Belgium, using AASM rules (Iber et al., 2007). No attempt at alignment was made prior to the execution of the study. The expert consensus, EXPERT_CON, was defined as the set of epochs for which EXPERT_1 and EXPERT_2 agreed.

2.4 | Statistical analysis

Scores were first compared on an epoch-by-epoch basis. Pairwise comparisons between EXPERT_1, EXPERT_2, EXPERT_CON and ALGO, for both Cz-Pz and C4-O2 analyses, were performed. The following agreement coefficients were computed: percentage of agreement, calculated as the number of epochs that were assigned the same sleep stage over the number of epochs for each recording; and Cohen's kappa (Cohen, 1960) (see Landis & Koch, 1977 for Cohen's kappa values' interpretation). Additionally, for ALGO versus EXPERT_CON, we computed the kappa value on pooled nights' scorings; that is, obtained by concatenating all the nights together, in order to counter a potential night-length effect (i.e., without attributing more/less weight to short/long nights, as is the case when computing the mean of kappa across separate nights) (Berthomier et al., 2007). For each sleep stage, we also computed the following: sensitivity (Se), defined as the number of epochs assigned a specific sleep stage by both scorers over the number of epochs scored as that stage by the rater used as reference; positive predictive value (PPV), calculated as the number of epochs assigned a specific sleep stage by both scorers over the number of epochs scored as that stage by the scorer being compared; specificity (Sp), defined as the number of epochs assigned as other than a specific stage by both scorers divided by the number of epochs assigned as other than a specific stage by the rater used as reference; and negative predictive value (NPV), defined as the number of epochs assigned as other than a specific stage by both scorers over the number of epochs assigned any other sleep stage by the rater used as reference (Altman, 1997).

Finally, potential rater effects on sleep variables (sleep onset latency to the first N2 epoch [SOL]; wake after sleep onset [WASO]; total sleep time [TST]; time spent in N1/N2/N3/REM [tN1/tN2/tN3/

tREM]; sleep efficiency [SE]; number of stage shifts [NSS]) were assessed through several generalized linear mixed models (GLMMs) with each sleep variable as a dependent variable in turn, using SAS version 9.4 (SAS Institute). The distribution of dependent variables was determined by fitting all parametric probability distributions to data, using the "allfitdist" function in Matlab (http://amir.eng.uci.edu/MvCAT.php; The Mathworks Inc.) and GLMMs were adapted accordingly. Subject was put as a random factor (intercept) and statistical significance was set at $p < .05$. Degrees of freedom were estimated using Kenward-Roger's correction and p-values in post-hoc contrasts (differences of least square means) were adjusted for multiple testing using Tukey's procedure.

3 | RESULTS

3.1 | EXPERT_1 versus EXPERT_2: whole-night agreements

We first considered the agreement between the two experts. Percentage of agreement ranged between 56% and 86%, with a mean of 76% (Figure 1A). Kappa values oscillated between 0.41 and 0.82, with 0.67 on average (Figure 1B), which constitutes a substantial agreement.

3.2 | ALGO versus EXPERTs: whole-night agreements

For the comparisons with ALGO, results were quite similar when using Cz-Pz or C4-O2. We detail here the results on Cz-Pz (see Supplementary Material for results using C4-O2).

Percentage of agreement between ALGO and EXPERT_1 ranged from 50% to 80%, with a mean of 67%, and κ ranged from 0.34 to 0.70 (mean 0.54; moderate agreement) (Figure 1A,B). For ALGO versus EXPERT_2, agreement values reached 64% to 81%, with a mean of 74%, and κ values went from 0.49 to 0.70, with a mean of 0.60 (substantial agreement). After discarding the epochs of expert disagreement, percentages of agreement between ALGO and EXPERT_CON were highest and ranged from 70% to 90%, with a mean of 78%. Kappa values oscillated between 0.54 and 0.84, with a mean of 0.66 (substantial agreement; mean = 0.68 following the concatenation procedure for potential night-length effect, see Methods). The contingency matrix (Table 1) provided the sensitivity, specificity, PPV and NPV for each sleep stage (Table 2).

3.3 | Rater effect on sleep variables

GLMMs showed significant rater effects for all sleep variables (Figure 2 A-F, H-I) except SOL and tREM (Figure 2G, Table 3). Post-hoc analyses showed that for tN1, the experts did not significantly differ between themselves, but both differed from ALGO, with higher tN1 values than ALGO. For WASO, tN2 and tN3, EXPERT_1 differed from both EXPERT_2 and ALGO, with lower WASO and tN2 values, and higher tN3. TST and SE showed lower values for EXPERT_2 than for EXPERT_1, and lower values for ALGO than for EXPERT_2. For the number of stage shifts, ALGO showed significantly lower values than the two experts, who showed no statistical difference for this variable. Interestingly, when pooling N2 and N3 duration together (i.e., SWS duration), no significant differences between scorers were detected (Table 3, Figure 3G). Figure 3 shows EXPERT_1, EXPERT_2 and ALGO scoring in a representative participant.

4 | DISCUSSION

Sleep stage scoring is of prime importance in sleep research, as well as in sleep medicine. It remains mostly performed through visual

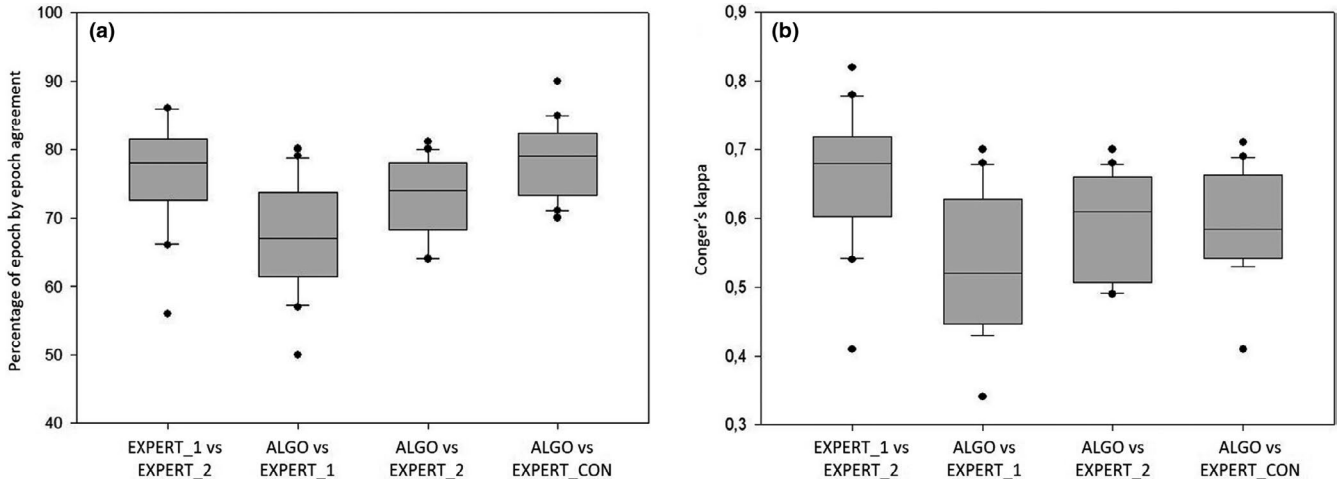

FIGURE 1 (a) Boxplot of percentage of agreement between visual scorers (EXPERT_1 and EXPERT_2), algorithm scoring (ALGO) and EXPERT_1, ALGO and EXPERT_2 and ALGO and consensus of expert rater scores (EXPERT_CON); (b) boxplot of Cohen's kappa for the same comparisons. The boxes' central line represents mean values, whereas one dot corresponds to one recording, and outliers were not removed from the plot

TABLE 2 Sensitivity (Se), Specificity (Sp), Negative Predictive Value (NPV) andPositive Predictive Value (PPV) for agreement between scorers for each sleep stage

		W	REM	N1	N2	N3
EXPERT_1 vs EXPERT_2	Se	74.2	84.8	63.4	73.3	99.3
	Sp	98.7	98.2	93.2	92.5	87.7
	PPV	89.9	90.8	50.3	91.5	37.6
	NPV	95.9	96.9	95.9	76.0	99.9
ALGO vs EXPERT_1	Se	83.9	83.9	8.2	83.1	44.8
	Sp	92.6	93.8	97.8	71.9	96.4
	PPV	59.8	71.0	34.2	68.3	74.1
	NPV	97.8	97.0	88.5	85.4	88.4
ALGO vs EXPERT_2	Se	85.1	85.0	11.6	80.0	60.4
	Sp	94.9	95.0	98.0	80.7	92.5
	PPV	73.2	77.1	38.0	82.2	38.0
	NPV	97.5	97.0	91.2	78.3	96.8
ALGO vs EXPERT_CON	Se	90.5	86.9	12.5	85.7	60.4
	Sp	94.8	96.0	98.1	82.8	96.0
	PPV	73.4	82.7	36.7	83.7	60.4
	NPV	98.4	97.1	92.8	84.8	96.0

Abbreviations: ALGO, algorithm scoring; EXPERT, human experts from 2 different research centres; EXPERT_CON, expert consensus. REM, rapid eye movement sleep.

inspection of the EEG. This leads to notable intra- and inter-expert variability when no alignment process is performed between experts (Berthomier et al., 2020). Inter-expert variability is likely to be higher when scoring sleep in individuals aged 40 years or older, as their sleep, although normal and healthy, shows more frequent transitions between wake and sleep, and lower SW density and amplitude (Carrier et al., 2011; Gaudreau, Carrier, et al., 2001; Landolt & Borbély, 2001; Silber et al., 2007). Yet, investigation of this likely phenomenon remains insufficient. Hence, we investigated inter-expert variability at a relatively early stage of the aging process, by comparing three sleep stage scorings, two performed by visual experts from different centres (EXPERT_1 and EXPERT_2) without prior alignment between them, and automatic scoring (ALGO) using a previously validated stage scoring algorithm (Berthomier et al., 2007; Peter-Derex et al., 2020).

Our analyses show that agreement between the two experts was substantial (κ = 0.67 on average), while it was lower, at moderate levels (κ = 0.54 and 0.60), when comparing ALGO to the two experts, but attained substantial levels when ALGO was compared to the consensus of experts (κ = 0.66; κ = 0.68 for the concatenation of all recordings). These results fall into the lower range of previous Cohen's coefficient values for comparison between experts from different centres. Previous studies with prior expert alignment showed kappa values ranging from 0.57 to 0.78 in a mixed population (Basner et al., 2008; Danker-Hopfe et al., 2009; Zhang et al., 2015) or equal to 0.76 in a mainly healthy population (Danker-Hopfe et al., 2009) where the level of agreement was found to decrease with subjects' age. Without prior expert alignment, kappa values ranged from 0.54 to 0.58 in a population mainly composed of patients (Zhang et al., 2015) or from 0.61 to 0.82 in various patient populations (Danker-Hopfe et al., 2004).

The main output of this study is the anticipated low agreement between the experts for N3 sleep. Agreement values showed that if nearly all the epochs scored as N3 by EXPERT_2 got the same label from EXPERT_1 (Sensitivity= 99.3%), about only one out of three epochs of N3 epochs of EXPERT_1 was scored the same way by EXPERT_2 (PPV = 37.6%). Regarding algorithm analysis, N3 assignment seemed to be intermediate between both experts. N3 scoring differences are illustrated in the representative subject displayed in Figure 3. These differences may arise from the fact that N3 is a sleep stage heavily relying on an amplitude criterion (75 μV), when amplitude is known to be affected by several factors, such as skull or scalp thickness (Cuffin, 1993), or even the contact quality between the scalp and the EEG sensor. Furthermore, it is now well established that sleep undergoes a series of changes during aging, starting as early as 40 years, amongst which a reduction in time spent in SWS and SW density (Carrier et al., 2011; Landolt, Dijk, Achermann, & Borbély, 1996; Mander, Winer, & Walker, 2017). It was suggested that this reduction observed in aging was due to an overall decrease in sleep EEG amplitude not specific to SW generation (Webb & Dreblow, 1982). Recent findings propose, however, that adapting the detection threshold decreases the differences seen between younger and older individuals, as it allows picking up of lower amplitude slow waves that are still generated in aged populations, but does not abolish them (Rosinvil et al., 2020). Our findings and these established variations in EEG signal call the 75 μV criterion into question, particularly in the aged population.

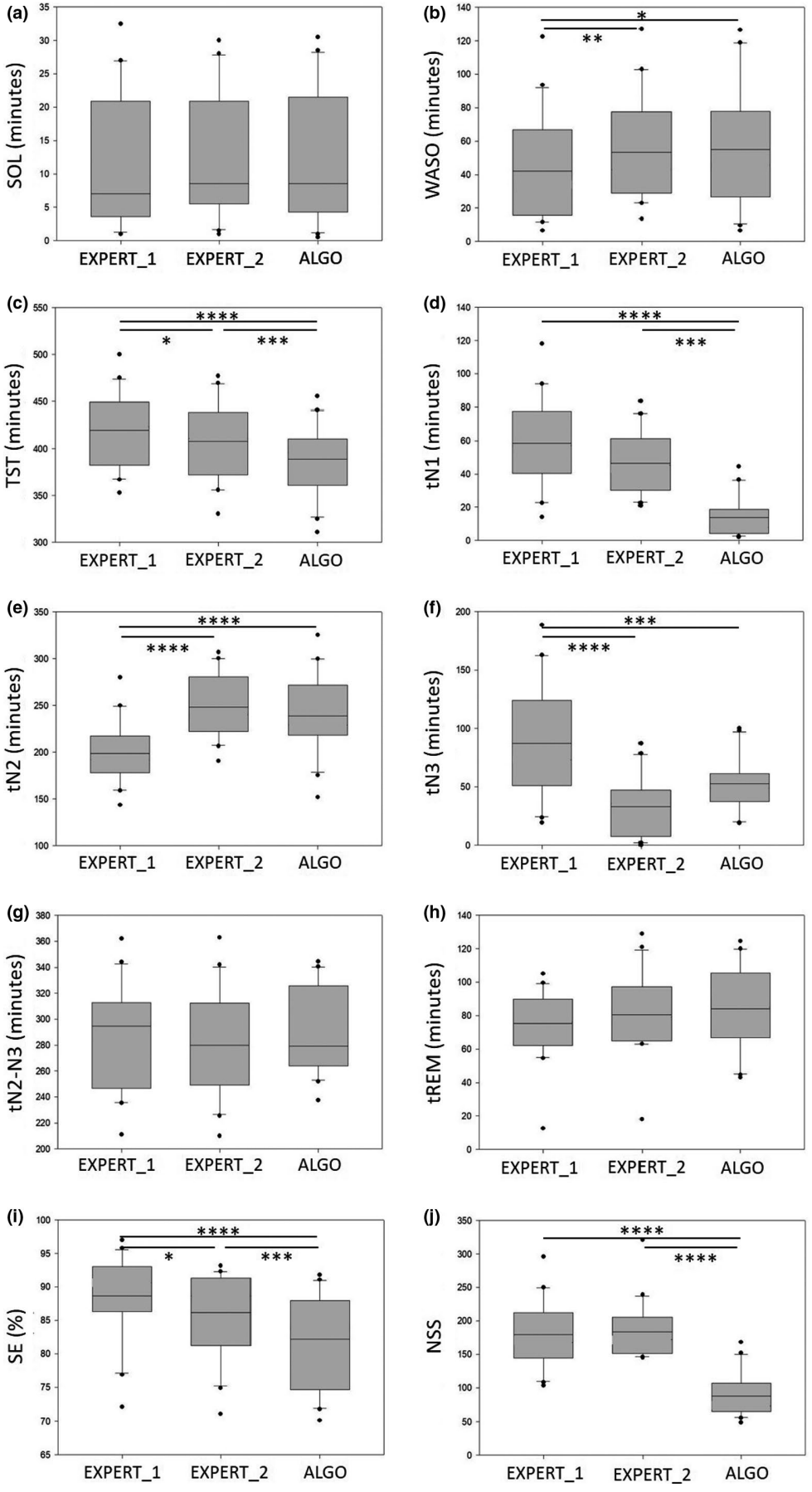
(a)
SOL (minutes)
EXPERT_1
EXPERT_2
ALGO
(b)
WASO (minutes)
*
**
(c)
TST (minutes)
****
*
***
(d)
tN1 (minutes)
(e)
tN2 (minutes)
(f)
tN3 (minutes)
(g)
tN2-N3 (minutes)
(h)
tREM (minutes)
(i)
SE (%)
(j)
NSS

FIGURE 2 Boxplot of sleep variable values for each rating method (EXPERT_1, EXPERT_2 and ALGO). Panel (a) sleep onset latency (SOL); (b) wake after sleep onset (WASO); (c) total sleep time (TST); (d) time spent in non-rapid eye movement (NREM) 1 sleep (tN1); (e) time spent in NREM2 (tN2); (f) time spent in NREM3 (tN3); (g) time spent in NREM2 and NREM3 without distinction; (h) time spent in rapid eye movement (REM) sleep (tREM); (i) sleep efficiency (SE); (j) number of stage shifts (NSS). The boxes' central line represents mean values, whiskers extend to the most extreme data points not considered outliers, and outliers were not removed from the plot.* means $p \leq 0.05$; ** means $p \leq 0.01$; *** means $p \leq 0.001$; **** means $p \leq 0.0001$

TABLE 3 Outcomes of GLMMs with each sleep variable assessing differences between rating methods

	Main effect		Post-hoc					
			EXPERT_1 vs. EXPERT_2		EXPERT_1 vs. ALGO		EXPERT_2 vs. ALGO	
	F(2,38)	*p*	*t*	*p*	*t*	*p*	*t*	*p*
SOL	0.97	0.39	NA					
WASO	6.24	**0.0045**	−3.32	**0.0055**	−2.70	**0.0272**	0.63	0.8073
TST	21.36	**<0.0001**	2.44	**0.0495**	6.47	**<0.0001**	4.03	**0.0007**
tN1	40.28	**<0.0001**	1.30	0.4064	.34	**<0.0001**	7.04	**<0.0001**
tN2	28.21	**<0.0001**	−7.06	**<0.0001**	−5.75	**<0.0001**	1.31	0.3977
tN3	20.27	**<0.0001**	6.29	**<0.0001**	4.02	**0.0008**	−2.26	0.0738
tN2-N3	2.44	0.10	NA					
tREM	1.97	0.1540						
SE	21.54	**<0.0001**	2.51	**0.0426**	6.50	**<0.0001**	3.99	**0.0008**
NSS	93.88	**<0.0001**	−0.52	0.8618	11.60	**<0.0001**	12.12	**<0.0001**

Abbreviations: ALGO, algorithm scoring; EXPERT, human experts from two different research centres; NSS, number of stage shifts; REM, rapid eye movement sleep; SE, sleep efficiency; SOL, sleep onset latency; tN1/N2/N3/N2-N3/REM, time spent in N1/N2/N3/N2-N3 (without distinction)/REM; TST, total sleep time; WASO, wake after sleep onset.

The automatic scoring algorithm used here (Aseega) does not use absolute amplitude criteria, but rather relies on data-driven individual thresholds for scoring. Yet, it does not appear to assign more epochs to N3 than EXPERT_1. Although visual scorers are relying on the AASM rules and should observe the 75 μV criterion, in practice, scoring habits may emerge in sleep research centres (e.g., due to reliance on other criteria such as slow wave continuity or because of the typical population most recorded). Importantly, when considering both N2 and N3 duration together (SWS as whole), staging methods do not differ anymore, further suggesting that it is the identification of lower amplitude oscillations as slow waves and whether they cover more than 20% of a 30-s epoch that drives the differences between raters. This means that computation of SWS parameters, including slow-wave activity (SWA) (i.e., the EEG power within the delta band [0.5-4 Hz] during NREM sleep, which is considered to reflect sleep need) (Achermann, Dijk, Brunner, & Borbély, 1993), based on either of the experts or on ALGO would be similar. Because agreement was also high between raters for REM (Table 2, Figure 2), computation of EEG power during REM (e.g., in the theta band, 4–8 Hz) should also be similar across staging methods.

We also find low agreement for N1 sleep. The number of epochs scored N1 by any rater was low and significantly differed between experts, but it was even lower when using ALGO. This may not come as a surprise and might result from the low EEG characterizability of N1, as, apart from the occasional vertex sharp wave, stage N1 is mainly characterized by a reduction of alpha rhythm or activity in the theta range. It is thus a relatively unspecific stage, and ALGO might tend to prioritize classification in other stages. Indeed, 31% and 38% of epochs scored by both experts as N1 are scored as wakefulness and N2 by ALGO, respectively. This first suggests that the disagreement over N1 epochs happens in part at the beginning of the sleep opportunity, where ALGO is inclined to favour wakefulness longer and tends to score sleep onset directly to N2. The second may reflect the fact that ALGO better detects spindles embedded into a mixed frequency background. Importantly, only 16% of N1 epochs of experts are categorized as REM by ALGO. Notably, SOL, the time between lights off and the first N2 epoch, which is an important sleep variable, including in clinical practice, was not significantly different across staging methods. Also, TST and SE differed between all three scorers.

Overall, we report non-negligible inter-expert variability on several classically used sleep variables when considering sleep recordings of individuals in their late midlife/early old age scored by experts from different centres, with no prior alignment process. Although beyond the scope of the current paper, because age-related changes are progressive, we suspect that this variability would be further exacerbated in even older individuals. This would, however, deserve further investigation. Inter-expert scoring variability across research centres does not imply that the effect of an experiment manipulation cannot be detected in older individuals within a research centre where sleep recordings are usually scored the same way (and often by the same

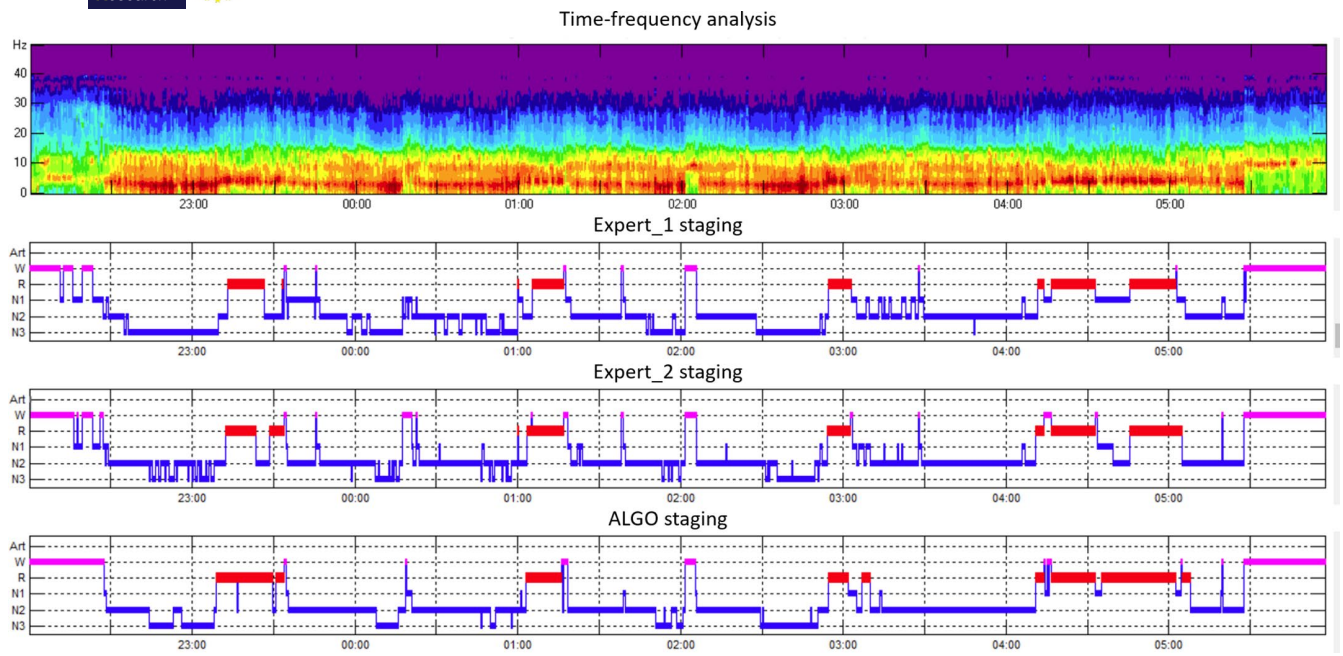

FIGURE 3 Representative example of scorings for each rater. From top to bottom: time frequency analysis, hypnograms of EXPERT_1, EXPERT_2 and algorithm automatic staging (ALGO). Although the macrostructure seems identical, sleep is heavier in the case of EXPERT_2 than for EXPERT_1, and ALGO seems to lie in between. At sleep onset, both experts put N1, whereas ALGO stays in wakefulness before going straight to N2

experts). Caution is required, however, when comparing results across research centres (e.g., in BIG data studies), and may prevent normative values being obtained from the literature. It also highlights the importance of scoring rules and scoring criteria for sleep stage scoring. Although sleep unarguably undergoes a series of changes with aging, an important question remains over to what extent the reported changes are influenced by the potentially considerable variability between sleep staging experts. It may be particularly important to regularly set a common ground for scoring, as recommended by the AASM, or to turn to automatic algorithms. The latter may not be better than human experts but will be systematic and consistent across and within research centres. Without such standardization, it may prove difficult to compare scorings and thus extracted sleep variables, probably particularly so in aged populations where sleep is more fragile.

ACKNOWLEDGEMENTS

This work was supported by the Canadian Institutes of Health Research (CIHR) (grant number 190750 [JC]), Fonds National de la Recherche Scientifique (FRS-FNRS, Belgium, T.0242.19), Actions de Recherche Concertées (ARC SLEEPDEM 17/27-09) of the Fédération Wallonie-Bruxelles, University of Liège (ULiège) and the European Research Council (COGNAP project-GA 757763); GV is supported by the F.R.S.-FNRS Belgium.

CONFLICT OF INTEREST

C. Berthomier and M. Brandewinder have ownership/directorship and are employees of Physip, which owns Aseega.

AUTHOR CONTRIBUTIONS

DC, CB, JC, GV and VM designed the research; SF and EL scored sleep stages; DC, CB and VM analysed data; DC, CB, MB, GV and VM wrote the paper. All authors revised the manuscript.

DATA AVAILABILITY STATEMENT

Data can be made available upon request.

ORCID

Daphne Chylinski https://orcid.org/0000-0002-7319-0859
Christian Berthomier https://orcid.org/0000-0002-2300-9476
Gilles Vandewalle https://orcid.org/0000-0003-2483-2752
Vincenzo Muto https://orcid.org/0000-0001-5100-9927

REFERENCES

Achermann, P., Dijk, D.J., Brunner, D.P., & Borbély, A.A. (1993). A model of human sleep homeostasis based on EEG slow-wave activity: Quantitative comparison of data and simulations. *Brain Research Bulletin*, *31*, 97–113. https://doi.org/10.1016/0361-9230(93)90016-5

Altman, D.G. (1997). *Practical statistics for medical research*. CRC press.

Basner, M., Griefahn, B., & Penzel, T. (2008). Inter-rater agreement in sleep stage classification between centers with different backgrounds. *Somnologie*, *12*, 75–84. https://doi.org/10.1007/s11818-008-0327-y

Berry, R.B., Brooks, R., Gamaldo, C., Harding, S., Lloyd, R., Quan, S., Troester, M., & Vaughn, B. (2017). *The AASM Manual for the Scoring of Sleep and Associated Events*. American Academy of Sleep Medicine.

Berthomier, C., Drouot, X., Herman-Stoïca, M., Berthomier, P., Prado, J., Bokar-Thire, D., Benoit, O., Mattout, J., & d'Ortho, M.P. (2007). Automatic analysis of single-channel sleep EEG: Validation in healthy individuals. *Sleep*, *30*, 1587–1595. https://doi.org/10.1093/sleep/30.11.1587

Berthomier, C., Muto, V., Schmidt, C., Vandewalle, G., Jaspar, M., Devillers, J., Gaggioni, G., Chellappa, S.L., Meyer, C., Phillips, C., Salmon, E., Berthomier, P., Prado, J., Benoit, O., Bouet, R., Brandewinder, M., Mattout, J., & Maquet, P. (2020). Exploring scoring methods for research studies: Accuracy and variability of visual and automated sleep scoring. *Journal of sleep research*, *29*(5), e12994.

Březinová, V. (1975). The number and duration of the episodes of the various EEG stages of sleep in young and older people. *Electroencephalography and Clinical Neurophysiology*, *39*, 273–278. https://doi.org/10.1016/0013-4694(75)90149-2

Cajochen, C., Münch, M., Knoblauch, V., Blatter, K., & Wirz-Justice, A. (2006). Age-related changes in the circadian and homeostatic regulation of human sleep. *Chronobiology International*, *23*, 461–474. https://doi.org/10.1080/07420520500545813

Carrier, J., Viens, I., Poirier, G., Robillard, R., Lafortune, M., Vandewalle, G., … Filipini, D. (2011). Sleep slow wave changes during the middle years of life. *European Journal of Neuroscience*, *33*, 758–766. https://doi.org/10.1111/j.1460-9568.2010.07543.x

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Cuffin, N. (1993). Effects of local variations in skull and scalp thickness on EEG's and MEG's. *IEEE Transactions on Biomedical Engineering*, *40*(1), 42–48.

Danker-Hopfe, H., Anderer, P., Zeitlhofer, J., Boeck, M., Dorn, H., Gruber, G., Heller, E., Loretz, E., Moser, D., Parapatics, S., Saletu, B., Schmidt, A., & Dorffner, G. (2009). Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. *Journal of Sleep Research*, *18*, 74–84. https://doi.org/10.1111/j.1365-2869.2008.00700.x

Danker-Hopfe, H., Kunz, D., Gruber, G., Klösch, G., Lorenzo, J.L., Himanen, S.L., Kemp, B., Penzel, T., Röschke, J., Dorn, H., Schlögl, A., Trenker, E., & Dorffner, G. (2004). Interrater reliability between scorers from eight European sleep laboratories in subjects with different sleep disorders. *Journal of Sleep Research*, *13*, 63–69. https://doi.org/10.1046/j.1365-2869.2003.00375.x

Do Kim, Y., Kurachi, M., Horita, M., Matsuura, K., & Kamikawa, Y. (2007). Agreement in visual scoring of sleep stages among laboratories in Japan. *International Journal of Phytoremediation*, *20*, 135–136. https://doi.org/10.1111/j.1365-2869.1992.tb00011.x

Dube, J., Lafortune, M., Bedetti, C., Bouchard, M., Gagnon, J.F., Doyon, J., Evans, A., Lina, J.M., & Carrier, J. (2015). Cortical thinning explains changes in sleep slow waves during adulthood. *Journal of Neuroscience*, *35*, 7795–7807. https://doi.org/10.1523/JNEUROSCI.3956-14.2015

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Gaudreau, H., Carrier, J., & Montplaisir, J. (2001). Age-related modifications of NREM sleep EEG: From childhood to middle age. *Journal of Sleep Research*, *10*, 165–172. https://doi.org/10.1046/j.1365-2869.2001.00252.x

Gaudreau, H., Morettini, J., Lavoie, H.B., & Carrier, J. (2001). Effects of a 25-h sleep deprivation on daytime sleep in the middle-aged. *Neurobiology of Aging*, *22*, 461–468. https://doi.org/10.1016/S0197-4580(00)00251-7

Iber, C., Ancoli-Israel, S., Chesson, A., & Quan, S. (2007). *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications*. American Academy of Sleep Medicine.

Klerman, E.B., & Dijk, D.J. (2008). Age-related reduction in the maximal capacity for sleep-implications for insomnia. *Current Biology*, *18*, 1118–1123. https://doi.org/10.1016/j.cub.2008.06.047

Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Journal of Biometrics*, *33*, 159–174. https://doi.org/10.2307/2529310

Landolt, H.P., & Borbély, A.A. (2001). Age-dependent changes in sleep EEG topography. *Clinical Neurophysiology*, *112*, 369–377. https://doi.org/10.1016/S1388-2457(00)00542-3

Landolt, H.P., Dijk, D.J., Achermann, P., & Borbély, A.A. (1996). Effect of age on the sleep EEG: Slow-wave activity and spindle frequency activity in young and middle-aged men. *Brain Research*, *738*, 205–212. https://doi.org/10.1016/S0006-8993(96)00770-6

Mander, B.A., Winer, J.R., & Walker, M.P. (2017). Sleep and human aging. *Neuron*, *94*, 19–36. https://doi.org/10.1016/j.neuron.2017.02.004

Monroe, L.J. (1969). Inter-rater reliability and the role of experience in scoring EEG sleep records: phase 1. *Psychophysiology*, *5*, 376–384. https://doi.org/10.1111/j.1469-8986.1969.tb02836.x

Ohayon, M.M., Carskadon, M.A., Guilleminault, C., & Vitiello, M.V. (2004). Meta-analysis of quantitative sleep parameters from childhood to old age in healthy individuals: Developing normative sleep values across the human lifespan. *Sleep*, *27*, 1255–1273. https://doi.org/10.1093/sleep/27.7.1255

Pannese, E. (2011). Morphological changes in nerve cells during normal aging. *Brain Structure and Function*, *216*, 85–89. https://doi.org/10.1007/s00429-011-0308-y

Peter-Derex, L., Berthomier, C., Taillard, J., Berthomier, P., Bouet, R., Mattout, J., Brandewinder, M., & Bastuji, H. (2020). Automatic analysis of single-channel sleep EEG in a large spectrum of sleep disorders. *Journal of Clinical Sleep Medicine*, *17*(3), 393–402. https://doi.org/10.5664/jcsm.8864

Redline, S., Kirchner, H.L., Quan, S.F., Gottlieb, D.J., Kapur, V., & Newman, A. (2004). The effects of age, sex, ethnicity, and sleep-disordered breathing on sleep architecture. *Archives of Internal Medicine*, *164*, 406–418. https://doi.org/10.1001/archinte.164.4.406

Rosenberg, R.S., & Van Hout, S. (2013). The American Academy of Sleep Medicine inter-scorer reliability program: Respiratory events. *Journal of Clinical Sleep Medicine*, *10*, 447–454. https://doi.org/10.5664/jcsm.3630

Rosinvil, T., Bouvier, J., Dubé, J., Lafrenière, A., Bouchard, M., Cronier, J., Gosselin, N., Carrier, J., & Lina, J.M. (2020). Are age and sex effects on sleep slow waves only a matter of EEG amplitude? *Sleep*, *44*(3), 1–33. https://doi.org/10.1093/sleep/zsaa186

Silber, M.H., Ancoli-Israel, S., Bonnet, M.H., Chokroverty, S., Grigg-Damberger, M.M., Hirshkowitz, M., Kapen, S., Keenan, S., Kryger, M., Penzel, T., Pressman, M., & Iber, C. (2007). The visual scoring of sleep in adults. *Journal of Clinical Sleep Medicine*, *3*, 121–131. https://doi.org/10.5664/jcsm.26814

Webb, W.B., & Dreblow, L.M. (1982). A modified method for scoring slow wave sleep of older subjects. *Sleep*, *5*, 195–199. https://doi.org/10.1093/sleep/5.2.195

Whitney, C.W., Gottlieb, D.J., Redline, S., Norman, R.G., Dodge, R.R., Shahar, E., Surovec, S., & Nieto, F.J. (1998). Reliability of scoring respiratory disturbance indices and sleep staging. *Sleep*, *21*, 749–757. https://doi.org/10.1093/sleep/21.7.749

Zhang, X., Dong, X., Kantelhardt, J.W., Li, J., Zhao, L., Garcia, C., … Han, F. (2015). Process and outcome for international reliability in sleep scoring. *Sleep Breath.*, *19*, 191–195. https://doi.org/10.1007/s11325-014-0990-0

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Chylinski, D., Berthomier C., Lambot E., Frenette S., Brandewinder M., Carrier J., et al. Variability of sleep stage scoring in late midlife and early old age. *Journal of Sleep Research*. 2021;00:e13424. https://doi.org/10.1111/jsr.13424