Received: 15 August 2019 | Revised: 20 January 2020 | Accepted: 20 January 2020

DOI: 10.1111/jsr.12994

REGULAR RESEARCH PAPER

Journal of Sleep Research ESRS

Exploring scoring methods for research studies: Accuracy and variability of visual and automated sleep scoring

Christian Berthomier[1] | Vincenzo Muto[2,3,4] | Christina Schmidt[2,4] | Gilles Vandewalle[2] | Mathieu Jaspar[2,3,4] | Jonathan Devillers[2,3] | Giulia Gaggioni[2] | Sarah L. Chellappa[2] | Christelle Meyer[2,3] | Christophe Phillips[2,5] | Eric Salmon[2] | Pierre Berthomier[1] | Jacques Prado[1] | Odile Benoit[1] | Romain Bouet[6] | Marie Brandewinder[1] | Jérémie Mattout[6] | Pierre Maquet[2,3,7]

[1]PHYSIP, Paris, France

[2]GIGA-Cyclotron Research Centre-In vivo Imaging, University of Liège, Liège, Belgium

[3]Walloon Excellence in Life Sciences and Biotechnology (WELBIO), Liège, Belgium

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

[5]Department of Electrical Engineering and Computer Science, University of Liège, Liège, Belgium

[6]Lyon Neuroscience Research Center, INSERM U1028, CNRS UMR 5292, University of Lyon 1, Lyon, France

[7]Department of Neurology, CHU Liège, Liège, Belgium

Correspondence
Christian Berthomier, PHYSIP, 6 rue Gobert, 75011 Paris, France.
Email: C.Berthomier@physip.fr

Vincenzo Muto, GIGA-CRC-In vivo Imaging, Université de Liège, Allée du 6 Août, Bâtiment B30, Sart Tilman, 4000 Liège, Belgique.
Email: vincenzo.muto@uliege.be

Funding information
University of Liege; Fonds National de la Recherche Scientifique (FNRS); Walloon Excellence in Life sciences and Biotechnology

Abstract

Sleep studies face new challenges in terms of data, objectives and metrics. This requires reappraising the adequacy of existing analysis methods, including scoring methods. Visual and automatic sleep scoring of healthy individuals were compared in terms of reliability (i.e., accuracy and stability) to find a scoring method capable of giving access to the actual data variability without adding exogenous variability. A first dataset (DS1, four recordings) scored by six experts plus an autoscoring algorithm was used to characterize inter-scoring variability. A second dataset (DS2, 88 recordings) scored a few weeks later was used to explore intra-expert variability. Percentage agreements and Conger's kappa were derived from epoch-by-epoch comparisons on pairwise and consensus scorings. On DS1 the number of epochs of agreement decreased when the number of experts increased, ranging from 86% (pairwise) to 69% (all experts). Adding autoscoring to visual scorings changed the kappa value from 0.81 to 0.79. Agreement between expert consensus and autoscoring was 93%. On DS2 the hypothesis of intra-expert variability was supported by a systematic decrease in kappa scores between autoscoring used as reference and each single expert between datasets (.75–.70). Although visual scoring induces inter- and intra-expert variability, autoscoring methods can cope with intra-scorer variability, making them a sensible option to reduce exogenous variability and give access to the endogenous variability in the data.

KEYWORDS

automatic scoring, large datasets, scoring variability, visual scoring

1 | INTRODUCTION

Sleep studies are entering a new era, with datasets becoming larger and targetting a wider range of objectives, such as phenotypic (Van Dongen, Vitellaro, & Dinges, 2005), longitunal (Redline, Schluchter, Larkin, & Tishler, 2003), multicentric (Redline et al., 2011) or epidemiologic objectives (Castro, Poyares, Leger, Bittencourt, & Tufik, 2013). Reflection on the adequate ways to

C. Berthomier and V. Muto contributed equally to this work.

describe sleep accompanies the evolution of the science of sleep, both regarding analysis methods (Stephansen et al., 2018) and classification in terms of sleep stages. Sleep scoring consists of applying consensual criteria (Iber, et al. 2007) to identify stages and results in the hypnogram (i.e., the succession of sleep stages across time). Sleep scoring methods are long debated, visual and automatic scoring being both questionned and defended with respect to reliability. Visual sleep scoring is the reference standard for sleep analysis. However, it is affected by inter and intra-expert variability (Collop, 2002; Magalang et al., 2013; Penzel, Zhang, & Fietze, 2013; Zhang et al., 2015), because it is difficult for human experts to achieve the same scoring for a given recording (Grigg-Damberger, 2012; Himanen & Hasan, 2000; Morgenthaler, Deriy, Heald, & Thomas, 2016; Rosenberg & Van Hout, 2013; Van Dongen et al., 2005). Training sessions improve the homogeneous application of the scoring rules (Danker-Hopfe et al., 2009; Rosenberg & Van Hout, 2013) but must be repeated to prevent a scoring "drift" over time, identified by Redline et al. as a critical concern in clinical and epidemiological research (Redline, Dean, & Sanders, 2013). Periodic redefinition of scoring rules (Grigg-Damberger, 2012; Himanen & Hasan, 2000; Morgenthaler et al., 2016; Schulz, 2008), together with the difficulty of applying these rules consistently, adds another variability source (Danker-Hopfe et al., 2009). In conclusion, visual scoring introduces exogenous, confounding sources of variability to the intrinsic meaningful variability of the data. On the other hand, automatic scoring, unlike visual scoring, can be totally reproducible, and its reliability has significantly improved over the years to the point that it could appear as a sensible option for analysing research data. We explored both visual and automated scoring in terms of accuracy and variability, using each approach as a way to question the other one, in turns.

2 | METHODS

2.1 | Experimental design

2.1.1 | Participants and protocol

This was a retrospective study. The data were collected in 24 healthy male participants (aged 21.6 ± 2.5 years), randomly selected from a larger sample (364) of volunteers. All volunteers were free of medication or psychoactive drugs, non-smokers and moderate caffeine and alcohol consumers, devoid of sleep and cognitive disorders. The study was approved by the Ethics Committee of the Faculty of Medicine of the University of Liège. After a first habituation night aimed at ruling out participants with sleep disorders, four full-night polysomnographic recordings (PSG) were acquired under successive sleep conditions: baseline night (BAS), 12-hr extended sleep opportunity (EXT), 8-hr night preceding 40 hr of sleep deprivation (BEF) and a recovery night (REC), following the sleep deprivation (Figure 1).

2.1.2 | Recordings

Data were recorded using a V-Amp 16 amplifier (Brain Products GmbH). Electroencephalogram (EEG) data were digitized at Fs = 500 Hz with a low-pass filter at 180 Hz and a magnitude resolution of 0.049 µV/bit. Recordings included 10 EEG channels (F3, Fz, F4, C3, Cz, C4, Pz, O1, O2, A1), two electro-oculograms (EOGs), two chin electromyograms (EMGs) and two electrocardiograms (ECGs).

2.2 | Automatic scoring (AS)

Aseega is an automatic sleep scoring method based on the analysis of the single EEG bipolar signal Cz-Pz, without using any information from either EOG or EMG. The data were automatically analysed without information on subject or sleep condition. Unlike evolutive methods, such as stochastic methods or methods using incremental learning, Aseega is fully deterministic, hence results are fully reproducible. It was validated on healthy subjects (Berthomier et al., 2007).

2.3 | Visual scoring (VS)

2.3.1 | Pool of independent scorers

Six expert visual scorers from the same laboratory were trained together to homogenize the application of AASM scoring rules (Iber et al., 2007). The training was considered achieved when 80% of scoring agreement was reached with the expert leader, considered

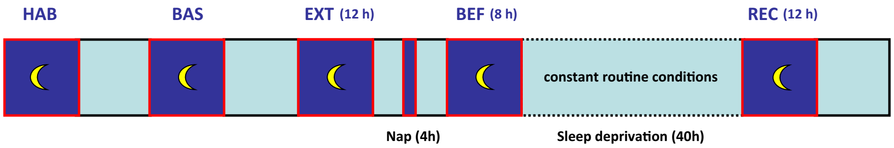

FIGURE 1 Experimental protocol. Participants underwent a 6-day laboratory protocol including: a habituation night, a baseline night (BAS), an extended sleep episode (EXT, i.e., 12 hr), a 4-hr afternoon nap, an 8-hr night (BEF), 40 hr of sleep deprivation, followed by a 12-hr recovery night (REC)

as the laboratory reference standard. Here, a human scorer is designated by "visual scorer" (VS) or "expert", whereas "AS" refers to automatic scoring. 'Scorer' and 'scoring' refer to automated and visual scorer/scoring.

2.3.2 | Setup of visual references

In order to get closer to the scoring ground truth, we took advantage of the multiple visual scorings by setting up two kinds of visual references.

First, we used the full consensus, which is the drastic restriction of the dataset to only the epochs on which all independent scorers agreed. This is determined by statistical means, not as the result of verbal discussion between experts.

As an alternative to full consensus, which is a harsh approach because any epoch where only one expert disagrees is rejected, we used the visual majority scoring, V_{Maj} (Rosenberg & Van Hout, 2013). For each epoch, we considered as V_{Maj} the scoring decision that brings together at least N_{Maj} experts, with N_{Maj} varying from two to six. Note that a majority of at least six experts is equivalent to the full consensus case.

2.4 | Datasets

The data consisted of two scoring sets, DS1 and DS2, referred to as scoring condition, which differed according to the temporal proximity to the training sessions. The scoring sessions of DS2 started 1.5 months after DS1 within a 6-month period. Both datasets excluded the recordings used for the training sessions. The composition of the two datasets is summarized in Table 1. DS1 included four recordings from two volunteers: three recordings (BAS, EXT and BEF) from one participant and one recording (BAS) from another one. Each recording was scored independently by all seven different scorers (six VS and one AS), resulting in 28 scorings.

DS2 included 88 recordings (four recordings from 22 different subjects), scored twice, once by one of the six VS and once by AS, resulting in 176 scorings. Contrary to DS1, each recording was visually scored only once, in addition to AS. When a subject was assigned to an expert the four nights were scored only by this expert (see Table 1).

2.5 | Statistics

Scorings were compared on an epoch-by-epoch basis. To avoid overweighting short nights and underweighting long nights, comparisons in each dataset were based on pooled night scorings. For each dataset separately, pooling was carried out by concatenating the hypnograms of a given scorer (AS and VS) into a single continuous sequence of sleep stages (Berthomier et al., 2007).

Two metrics of epoch-by-epoch agreement between scorers were used: first, the percentage agreement, defined as the percentage of epochs that were assigned the same label (i.e., sleep stage) by two or more scorers; and second, Conger's kappa coefficient (κ), which is the generalization of Cohen's kappa coefficient (Cohen, 1960) to the comparison of more than two raters. To have homogeneous statistical criteria we computed Conger's kappa instead of the Fleiss kappa, which is sometimes used in multi-scorer comparison (Danker-Hopfe et al., 2009) but which reduces to a different two-rater agreement coefficient called Pi (Gwet, 2012). In the case of two raters, Conger's kappa and Cohen's kappa are equivalent (Conger, 1980).

Distributions of agreements are displayed using box-and-whisker plots, where the central mark is the median and the edges of the box are the 25th and 75th percentiles. The whiskers extend to the most extreme data points, whereas the outliers are plotted individually.

TABLE 1 Cohort breakdown into two datasets, DS1 and DS2, according to the number of times recordings were visually scored by the six visual scorers involved (V_1 ...V_6)

	Subjects		Visual scorers						Auto
Dataset 1	Subject ID	# recordings per subject	V_1	V_2	V_3	V_4	V_5	V_6	A
	981	3	x	x	x	x	x	x	x
	1,004	1	x	x	x	x	x	x	x
Dataset 2	18, 67, 180, 314	4	x						x
	44, 110	4		x					x
	45, 112, 208, 365	4			x				x
	41, 106, 200, 330, 412, 439	4				x			x
	17, 66, 123, 273	4					x		x
	204, 316	4						x	x

Note: Each subject may provide for one to four nights. All scorers scored all recordings of DS1 (four nights). In DS2, subjects were assigned to scorers who scored the four nights of such subject. In other words, each recording of DS1 was visually scored six times, whereas each recording of DS2 was visually scored only once.

2.6 | Scoring variabilities

The inter-scorer variability is defined as the difference measured between the scorings of two or more scorers.

The intra-scorer variability is defined as the differences between the scorings of a given scorer and a reference. In a first approach, the scorer reads the same recording twice and the score–rescore agreement is based on the comparison between the two scorings produced by the same scorer, used as its own reference. This is the typical score–rescore agreement (Whitney et al., 1998). A second approach is to compare two scorings produced by two different scorers on a succession of different recordings, and to consider the scorings of one of the scorers, supposed to be stable over time, as the scoring reference; the variability over time of the inter-scorer agreement is supposed to reflect the intra-scorer variability of the tested scorer. This corresponds to the typical training situation, where the increase over time of the inter-scorer agreement (trainee vs. sleep expert) is interpreted as the intra-trainee variability and is taken as progress of the trainee (Chediak et al., 2006; Rosenberg & Van Hout, 2013). Here, we explored a variation of the second approach, where the reference to which visual scorings are compared to assess fluctuations over time is autoscoring.

2.7 | Dataset 1 analysis

We first compared how agreement between experts is affected by the number of experts involved in the process by investigating the evolutions of both the percentage agreement and kappa as functions of the number of experts involved, N_{VS}, first between two experts, then between three, up to all experts. For a given N_{VS} (two to six), all possible combinations, N_c, of experts were computed, providing distributions of agreements.

In a second step, we added autoscoring to the pool. We performed pairwise comparisons using kappa and percentage agreement, with every possible pairs of scorings (VS and AS). We also performed comparisons between autoscoring and visual full consensus, VS_{all}, as well as between each visual scoring, VS_i, and the partial consensus of all the other experts (VS_j, i ≠ j).

In a third step, we compared all scorers (VS and AS) with respect to their individual contribution to the overall agreement, questioning whether automated analysis contributes differently to the overall agreement than visual experts. A global kappa, κ_{G1}, was computed on all seven scorings available (six VS + one AS). Partial kappa coefficients were computed for all possible pools of six scorings where only one was left out of the pool. The specific case where autoscoring was removed stands for an assessment of the impact of adding AS to the pool of experts.

The fourth step consisted of replacing the visual full consensus by the visual majority scoring. In case of *ex aequo* majority scorings (e.g., rapid eye movement [REM] decision for three experts and N2 decision for the other three experts for a given epoch), the ground truth was considered as undetermined and the epoch was discarded. Likewise, when an epoch did not meet the minimum number of majority members requested (e.g., if, in the previous case, the minimum number of experts requested to constitute a majority was $N_{Maj} = 4$), the epoch was discarded.

Finally, the contingency matrix between automatic scoring and the visual full consensus was computed to provide sensitivity and positive predictive value for sleep stages.

2.8 | Dataset 2 analysis

As a first step, the global kappa, κ_{G2}, and the percentage agreement between automatic and visual scorings were computed. The recordings scored by a given expert were then pooled together to provide for each expert a pairwise comparison with autoscoring.

In a second step the auto-visual agreements on DS2 were compared with the corresponding pairwise agreements obtained on DS1.

The third step aimed to assess the intra-expert variability according to the sleep condition (BAS, EXT, BEF and REC). We again used autoscoring as reference, because it is also blind to the sleep condition. Kappa coefficients and percentage agreement were computed for each expert and in each sleep condition.

Finally, the contingency matrix yielded sensitivity and positive predictive value for all sleep stages, without distinguishing between experts.

3 | RESULTS

3.1 | Dataset 1

In the first step, the inter-expert agreement as a function of the number of experts N_e (Figure 2) showed that the larger the number of experts, the lower percentage agreement and kappa variance, whereas mean kappa remained stable. The percentage agreement across all six visual scorers was 68.7% and the corresponding kappa was 0.81.

In the second step, concerning the inter-scorer variability ([min–max], μ = mean), the six pairwise kappa coefficients between autoscoring and each scorer separately, AS versus VS_i, ranged from 0.72 to 0.79, $\mu = 0.75$, and the 15 pairwise kappa coefficients between visual scorers, VS_i versus VS_j, ranged from 0.73 to 0.87, $\mu = 0.81$ (Figure 3a). The corresponding percentage agreements for AS versus VS_i ranged from 79.1% to 84.7%, $\mu = 81.7\%$, whereas for VS_i versus VS_j they ranged from 79.8% to 90.5%, $\mu = 85.6\%$ (Figure 3b).

As expected, pairwise comparisons yielded lower agreements than comparisons involving full consensus scoring because the latter discards ambiguous epochs. Accordingly, the kappa between AS and the visual full consensus VS_{all} was 0.91, whereas kappa between each individual expert and the partial consensus of the others experts ranged from 0.90 to 0.98, $\mu = 0.95$ (Figure 3a, third boxplot).

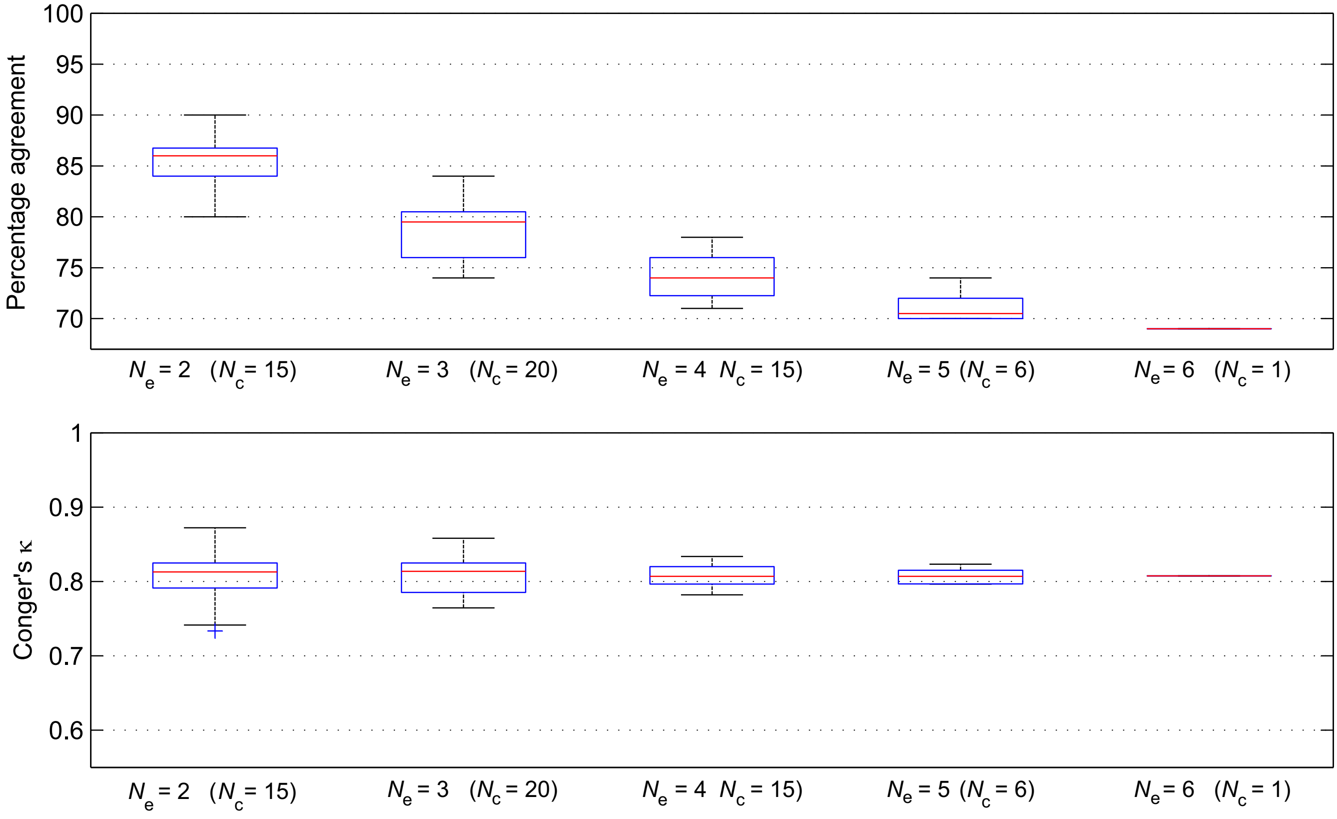

FIGURE 2 Inter-expert agreement. Evolution of the agreement according to the number of visual scorers involved. The consensus agreement (percentage agreement, upper plot) and the inter-expert agreement (Conger's kappa, lower plot) are drawn according to the number of experts, N_e, included in the pool. All the expert combinations, N_c, have been computed (N_c = 15 for agreement between N_e = 2 experts out of 6; N_c = 20 for three experts out of six, etc.) for each possible number of experts (N_e = 2–6), yielding distributions of agreements. Note that for N_e = 6 experts, the number of possible combinations is reduced to N_c = 1

The percentage agreement was 93.1% between AS and VS_{all}, and 92.9% to 98.5%, μ = 96.5%, between each VS_i and the partial consensus of the other visual scorings (Figure 3b, third boxplot).

In the third step, the global kappa over DS1 obtained by comparing all scorings (VS + AS) was κ_{G1} = 0.79, which is barely lower than the kappa obtained by comparing visual scorings (VS) only. The distribution of partial kappa coefficients when removing one scorer at a time by permutation ranged from 0.78 to 0.81 (μ = 0.79), the upper bound being the specific case of “visual” partial kappa, computed without autoscoring.

We then compared, in a fourth step, the scorings with various majority scorings, VS_{Maj}. A high N_{Maj} was associated with a high agreement with VS_{Maj}, but implied a high number of rejected epochs for which too few experts agreed to reach a consensus (Figure 4). No expert obtained a 100% agreement with the majority scoring.

For the last step, the subset of the most reliable epochs according to visual scoring was built by discarding the 1,349 epochs (31.3%) of disagreement between experts, leaving 2,959 epochs of full visual consensus. The contingency matrix is shown in Table 2. Automated scoring used as a benchmark (first column) shows that the highest disagreement among experts is observed for epochs autoscored N1, whereas higher agreement is observed among experts for Wake (W) and N3.

Among these 1,349 non-consensual epochs between experts, Aseega agreed 1,255 times (93.0%) with at least one expert.

Proceeding from the contingency matrix, the sensitivity and positive predictive value of all sleep stages are reported in Table 3.

3.2 | Dataset 2

Two recordings were rejected by Aseega because of low signal quality. Out of the remaining 102,141 recorded epochs, 764 (0.80%) were classified as artefacst (Arts) by automatic scoring and 445 (0.46%) were classified as Arts by the experts; 115 Art epochs were common to automatic and visual scoring. The total number of discarded epochs was 1,094 (1.14%), leaving 101,047 epochs for subsequent analysis.

Regarding the AS versus VS comparison, the global kappa was κ_{G2} = 0.70 and the percentage agreement 79.0%. The six pairwise kappa coefficients between automatic scoring and each expert ranged from 0.67 to 0.73, μ = 0.70 (Figure 3a, right boxplot). The corresponding percentage agreements ranged from 76.4% to 80.8%, μ = 78.7 (Figure 3b, right boxplot). Pairwise agreements between DS1 and DS2 systematically decreased from DS1 to DS2 for all experts (μ = −3.7%, Figure 5).

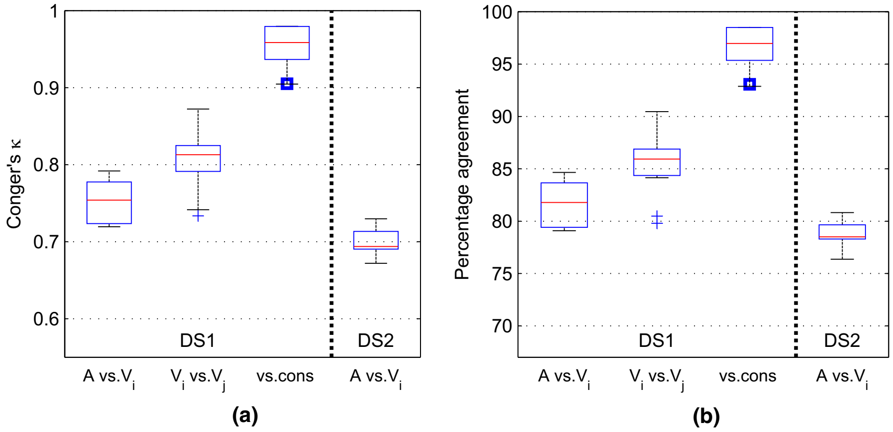

FIGURE 3 Inter-scorer and intra-scorer variability, pairwise comparisons. Comparisons are presented via (a) Conger's and (b) percentage agreement. These two graphs are divided into results on DS1 (the three boxplots on the left) and on DS2 (fourth box on the right). The comparison between automatic scoring and each visual scorer (A versus V$_i$) is represented on the first plot; the comparison between each pair of visual scorers (V$_i$ vs. V$_{j,}$ with i≠j) is represented on the second plot. On the third plot are reported the agreements between each visual scoring and partial consensus built by the other experts. The specific case of the agreement between autoscoring and the full visual consensus is highlighted in bold squares. The fourth boxplot on the right represents the A versus V$_i$ pairwise comparison on DS2. For each of the two graphs, the three plots on the left illustrate the inter-scorer variability on DS1 data and confirm that comparisons with full consensus provide far better agreements because the doubtful epochs are discarded. The left-most plot, together with the right-most plot, illustrates the intra-scorer variability between datasets (scoring condition; i.e., temporal proximity with the training sessions), using automatic analysis as the reference

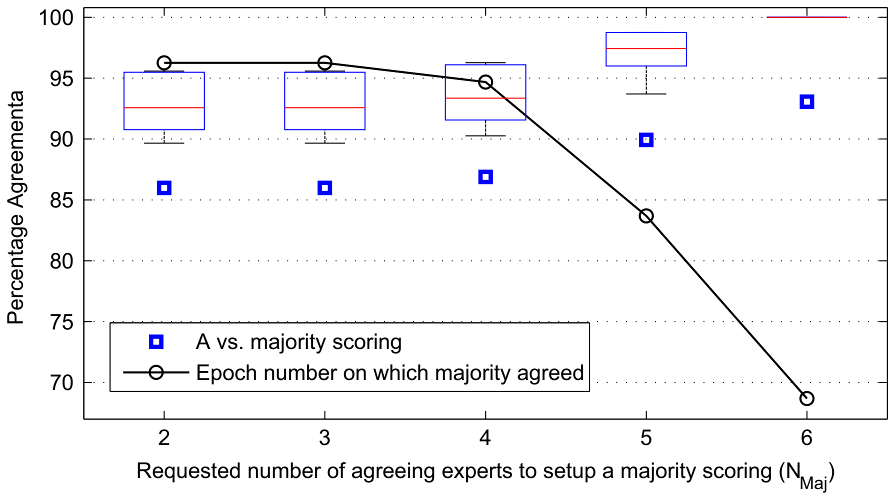

FIGURE 4 Majority scoring. Comparison between scorings and visual majority scoring, V$_{Maj}$, used as the scoring reference. The evolution of the agreement between V$_{Maj}$ and A (bold squares) or V (boxplot) is drawn according to the minimal number of agreeing experts, N$_{Maj}$, requested to set up a majority. The number of valid epochs on which the comparisons are computed is also plotted (circle) according to N$_{Maj}$. When more agreeing experts are requested to set up a majority, agreement increases, but on a number of valid epochs that decreases. The last case, N$_{Maj}$ = 6, equivalent to the visual full consensus, V$_{all}$, rejected a third of epochs. Second learning, the experts composing the majority scoring can be different at each epoch. For instance, when N$_{Maj}$ = 3, the V versus V$_{Maj}$ agreement ranged from 89.7% to 95.6%, μ = 92.8%. Apart from the trivial N = 6 consensus case, no expert reaches 100% agreement with the majority scoring, thus disagreements always exist between “real” visual scorings and “virtual” majority scoring (i.e., nobody scores like the majority scoring)

Regarding the impact of the sleep condition, Figure 6 shows that agreements on long-duration nights (EXT and REC) were globally lower than the ones obtained on shorter-duration nights (BAS and BEF).

The contingency matrix of dataset DS2 (Table 4) provides the sensitivity and positive predictive value of all sleep stages (Table 5).

4 | DISCUSSION

Scoring variability and accuracy cannot be measured in the typical scoring routine where recordings are scored once, and yet their effects still apply. Our study proposes new results and elements to

TABLE 2 Contingency matrix for the dataset DS1

		Full consensus of visual scorings						
Number of epochs (dataset DS1)		Artefact or disagree	W	REM	N1	N2	N3	Total
Automatic scoring	Art	**3**	33	–	–	–	3	39
	W	91	**575**	–	–	1	–	667
	REM	341	9	**450**	9	28	–	837
	N1	134	8	8	**12**	5	–	167
	N2	685	11	20	12	**1,023**	45	1,796
	N3	135	1	–	–	48	**694**	878
	Total	1,389	637	478	33	1,105	742	**4,384**

Note: The values on the first diagonal have special meaning (agreement) and should be in bold font for an easy reading of the table. 1,389 epochs out of 4,384 have been discarded; 76 for artefact (Art) labelling by AS or VS and 1,389 due to disagreement between visual scorers (first column). Wake, W.

TABLE 3 Sensitivity (Se) and positive predictive value (PPV) of automatic scoring versus visual full consensus (DS1)

Dataset DS1	W	REM	N1	N2	N3
Se	95.2	94.1	36.4	92.6	93.9
PPV	99.8	90.7	36.4	92.1	93.4

Note: These results were computed after the removal of the 1,389 epochs of partial disagreement between experts. Wake, W.

support the hypothesis on scoring variabilities, in order to identify the best ways to mitigate them.

4.1 | Visual–automatic agreement is similar to visual–visual agreement

Scoring agreement between scorers (all VS and AS) in DS1 was excellent (Landis & Koch, 1977), Conger's kappa of 0.79, whereas pairwise kappa between visual and automatic scoring ranged between 0.72 and 0.79, corresponding to an 82% mean agreement, in line with recent works (Fiorillo et al., 2019). The best kappa coefficient was observed when automatic scoring was left out. Likewise, better pairwise kappa and percent agreement were observed between visual scorers (respectively, 0.81 and 86%). High agreement between experts from the same centre guarantees high homogeneity in local scoring but is likely to lead to inter-site variability (not assessed here). Agreement between autoscoring and the consensus of visual scoring was similar to results published by Stephansen (Stephansen et al., 2018). A significant drop in automatic–visual agreement was noted between DS1 and DS2 for the REM positive predictive value, whereas the corresponding sensitivity remains quite stable (Table 3 and 5). Qualitative appreciation of scorings showed that the algorithm used tends to smooth REM episodes, unlike visual scoring where REM episodes are more fragmented. This fragmentation lowers automatic–visual agreement when automatic is compared to one expert only. However, whereas commonly marked by all experts, this fragmentation is located differently by all experts. It therefore disappears from visual consensus scoring, and thus from the comparison between AS and V_{Sall}.

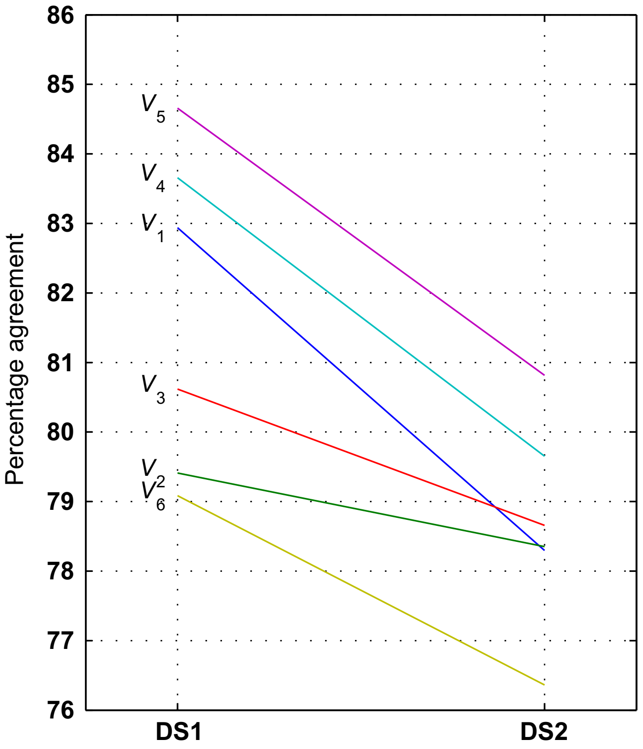

FIGURE 5 Evolution of the pairwise auto versus visual agreement over time. Impact of the scoring condition, temporal proximity of the dataset with the training sessions. For each visual scorer, V_i, evolution of the percentage agreement with automatic scoring according to the dataset. The agreement decrease between datasets is not only a mean effect (μ = −3.7%), as illustrated in Figure 3, but also a systematic effect for every single expert, ranging from −1.3% to −5.6%

4.2 | Inter-expert variability: visual disagreement is not just noise

As stated by Silber: “no visual-based scoring system will ever be perfect, as all methods are limited by the physiology of the human eye and visual cortex, individual differences in scoring experience, and the ability to detect events viewed using a 30-s epoch” (Silber et al.,

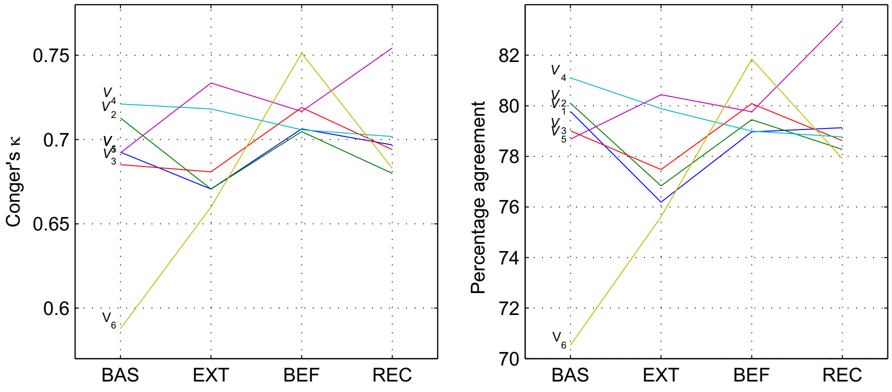

FIGURE 6 Evolution of the pairwise auto versus visual agreement over sleep condition. For each visual scorer, V_i, agreements (Conger's kappa on the left and percentage agreement on the right) are reported according to the sleep condition: baseline (BAS), extended night (EXT), before sleep deprivation (BEF) or recovery night (REC)

TABLE 4 Contingency matrix for the dataset DS2

		Visual scorings						
Number of epochs (dataset DS2)		Art	W	REM	N1	N2	N3	Total
Automatic scoring	Art	**115**	302	65	117	127	38	764
	W	38	**8,460**	140	714	154	7	9,513
	REM	148	776	**17,099**	4,341	2,430	0	24,794
	N1	15	481	329	**1,023**	458	11	2,317
	N2	95	1,013	1,488	2,790	**38,377**	2,817	46,580
	N3	34	75	1	36	3,144	**14,883**	18,173
	Total	445	11,107	19,122	9,021	44,690	17,756	**102,141**

Note: The values on the first diagonal have special meaning (agreement) and should be in bold font for an easy reading of the table. 764 epochs out of 102,141 have been discarded for Art labelling by AS and 445 by VS. Art, artefact; Wake, W.

TABLE 5 Sensitivity (Se) and positive predictive value (PPV) of automatic scoring versus visual scoring (DS2)

Dataset DS2	W	REM	N1	N2	N3
Se	78.3	89.7	11.5	86.1	84.0
PPV	89.3	69.4	44.4	82.6	82.0

2007). There is a continuous effort to improve the guidelines (GRADE program in 2009 (Morgenthaler et al., 2016)) and to homogenize their enforcement. Based on this program, over 2,500 technicians showed 82.6% agreement compared to a reference built as the majority score (Rosenberg & Van Hout, 2013). Even if this inter-scoring reliability may be overestimated, for protocol reasons (Redline et al., 2013) or by the use of majority scoring, these efforts are needed.

Here, when limiting the number of experts to two, their percentage agreement is about 85%, whatever the pairing. However, this apparent homogeneity is misleading because it suggests that only 15% of epochs raise doubts. Increasing the number of experts significantly decreases the overall agreement and shows that the contentious epochs and scoring consensus depend strongly on which pair of experts is considered. On DS1, the asymptotic consensus agreement is closer to 65%, meaning that one third of the epochs raised doubts. Our results show that the inter-expert disagreement cannot be considered as a low-level constant noise. It is not only a matter of specific epochs that are difficult to score (Younes, Raneri, & Hanly, 2016), otherwise adding more experts to build the scoring consensus would not affect the number of consensus epochs. The variability in inter-expert-agreement comes from both epoch-specific content (difficulty in applying the scoring rules) and expert-specific sensitivity to signal content.

Consensus and majority scoring are costly mitigations for inter-expert variability in two ways: they require several experts and entail large amounts of rejected epochs. Majority scoring (three experts minimum) suggests that a unified majority of agreeing experts exists, next to a minority of disagreeing experts. But the majority scoring can be composed of different experts at each epoch: no expert scores like the majority scoring (Figure 4). Among alternative approaches to cope with the inter-expert variability, the "discussed consensus" implies that experts discuss contentious epochs in order to reach a consensus (Danker-Hopfe et al., 2009; Penzel et al., 2013). If this time-costly approach does reinstate the non-consensual epochs, inter-expert variability is abolished at the cost of losing the independence of experts. Another alternative is computer-assisted scoring, which has been explored for decades (Anderer et al., 2010; Ktonas & Smith, 1976; Younes et al., 2016). However, it only copes partially with expert variability because it involves human expertise.

4.3 | Evolution of the auto–visual agreement across datasets: intra-expert variability?

Intra-expert variability, rooted in learning process, experience and fatigue, is usually demonstrated by comparing different scorings performed on the same data by a given expert (score–rescore). The common option of having all recordings read by the same scorer in a sleep laboratory mitigates the inter- but not the intra-expert variability. Relying on the reliability of the autoscoring used in the study, we decided to explore intra-expert variability by using autoscoring as the reference, where scorers are observed right after their training and a few months later. Whereas studies on training generally focus on the training period (Danker-Hopfe et al., 2009; Rosenberg & Van Hout, 2013), we gained insight into post-training situations, when scorers go back to their scoring routine and effects of training tend to at least partially fade away. As noticed by Danker-Hopfe et al. (2009), training sessions represent an essential tool to achieve a homogeneous interpretation of the scoring rules: in our study, prior training sessions probably led to the excellent inter-scorer agreement in DS1. We have no direct evidence as to visual–visual agreement in DS2, but the systematic decrease we observed in visual–automatic agreement for each expert between DS1 and DS2 (Figure 5) could be consistent with the hypothesis of a "drift" over time in visual scoring (Redline et al., 2013).

Regarding the sleep conditions, long nights (EXT and REC) showed globally lower agreements compared to shorter ones (Figure 6), suggesting that the intra-expert variability is possibly determined by fluctuations in attention. The intrinsic composition of each night appears to matter. For instance, the REC night, which as expected showed more consolidated sleep, gave rise to a better agreement between automatic and visual scoring. By contrast, the EXT night, which is associated with more fragmented sleep (more transitions) at the end of the night due to the decreased sleep debt, yields a lower agreement with a wider dispersion. The issues of visual scoring could result from the interaction between a specific content and the sensitivity of the scorer to this content.

4.4 | Automatic scoring: pros and cons

Visual scoring is variable and yet, in the current state of the art, it is the reference standard for sleep scoring. This inevitably raises the issue of the ground truth of sleep: if two scorers, or even one, even if they are highly trained and experienced, can provide different scorings of the same night, what is the score that corresponds adequately to the content of the night and to the real state of the subject? This uncertainty holds for sleep scoring in general. Indeed, it is rooted in the very principle of sleep scoring. As already pointed out, sleep scoring is structurally flawed, for instance by the 30-s (or 20-s) windows that are based only on contingent technical constraints with no physiological ground. Visual scoring provides access to a ground truth, which is extensive (i.e., on the whole night) but questionable when only one expert is involved, or strong but partial (many epochs rejected) when more experts are involved, yielding a consensus. This situation obviously makes the evaluation of automated analysis complicated, when there is no other way to proceed apart from to refer to an uncertain reference, made all the more difficult because different approaches coexist that differ in the way they are assessed (Anderer et al., 2005; Koupparis, Kokkinos, & Kostopoulos, 2014; Ktonas & Smith, 1976; Malhotra et al., 2013; Pittman et al., 2004; Popovic, Khoo, & Westbrook, 2014; Sun et al., 2017; Wang, Loparo, Kelly, & Kaplan, 2015), in protocol (population studied and numbers of experts) and in comparison methodology (reference setup and statistics). Precise assessment and comparison on common datasets using common metrics (de Zambotti et al., 2016; Dean et al., 2016; Penzel et al., 2013; Redline et al., 2013) remains an open question. Providing public sleep databases has been an ongoing and useful process for several years in the USA (www.sleepdata.org), in Europe (www.physionet.org) and more recently in Canada (www.ceams-carsm.ca/en/MASS). In addition, autoscoring methods differ in nature or by their objective: some use multichannel data analysis (Anderer et al., 2005; Malhotra et al., 2013; Pittman et al., 2004), others a single EEG channel (Berthomier et al., 2007; Popovic et al., 2014; Wang et al., 2015) or EOG only (Virkkala, Hasan, Varri, Himanen, & Muller, 2007), and some are limited to wake–sleep scoring (Kaplan, Wang, Loparo, Kelly, & Bootzin, 2014), making autoscoring also accountable for inter-scorer variability. The variabilities of visual scoring delineate the range of the acceptable uncertainty of automated scoring. Indeed, our results are in line with recent works (Fiorillo et al., 2019), which show that several automated analysis algorithms have reached this level of acceptable uncertainty. Automated analysis, being based on standard sleep scoring, remains unable to give access to the ground truth, but addresses the key issue of exogenous and potentially confounding variability. Indeed, the investigated variability of sleep characteristics, $V_{characteristic}$, is only reachable via analysis methods that introduce additional noise, V_{method}, such that the variability of the measurement $V_{observed}$ can be written as:

$$V_{observed} = V_{characteristic} + V_{method}$$

The stake is to minimize the contribution of the noise originating from the method. By neutralizing external sources of variance (intra-expert variability), autoscoring avoids possible masking effects of visual scoring, giving closer access to the intrinsic meaningful data variance. Automated scoring, based on alternative criteria derived from quantitative and artificial intelligence methods instead of conventional visual criteria, can then provide interesting new ways to describe sleep while carrying on building on the existing science of sleep.

ACKNOWLEDGMENTS

This study was funded by University of Liège research funds, Walloon Excellence in Life Sciences and Biotechnology (WELBIO) and Fonds National de la Recherche Scientifique (Belgium).

CONFLICT OF INTEREST

C. Berthomier, P. Berthomier and M. Brandewinder have ownership/directorship and are employees of Physip.

AUTHOR CONTRIBUTIONS

V.M., C.B. and P.M. designed the research; V.M., M.J., J.D., G.V., S.L.C. and C.M. performed the research; C.S., C.P., E.S., P.B., J.P., O.B., M.B., R.B., J.M. and P.M. provided expertise and feedback; V.M., C.B. and J.M. analysed data; V.M., C.B., M.B. and P.M. wrote the paper. All authors revised the manuscript.

ORCID

Christian Berthomier https://orcid.org/0000-0002-2300-9476
Vincenzo Muto https://orcid.org/0000-0001-5100-9927
Christina Schmidt https://orcid.org/0000-0002-5563-4671
Gilles Vandewalle https://orcid.org/0000-0003-2483-2752
Christophe Phillips https://orcid.org/0000-0002-4990-425X
Eric Salmon https://orcid.org/0000-0003-2520-9241
Jérémie Mattout https://orcid.org/0000-0003-2659-6984

REFERENCES

Anderer, P., Gruber, G., Parapatics, S., Woertz, M., Miazhynskaia, T., Klosch, G., … Dorffner, G. (2005). An E-health solution for automatic sleep classification according to Rechtschaffen and Kales: Validation study of the Somnolyzer 24 x 7 utilizing the siesta database. *Neuropsychobiology*, *51*, 115–133.

Anderer, P., Moreau, A., Woertz, M., Ross, M., Gruber, G., Parapatics, S., … Dorffner, G. (2010). 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*, *62*, 250–264.

Berthomier, C., Drouot, X., Herman-Stoïca, M., Berthomier, P., Prado, J., Bokar-Thire, D., … 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

Castro, L. S., Poyares, D., Leger, D., Bittencourt, L., & Tufik, S. (2013). Objective prevalence of insomnia in the Sao Paulo, Brazil epidemiologic sleep study. *Annals of Neurology*, *74*, 537–546.

Chediak, A., Esparis, B., Isaacson, R., Cruz, L. D. L., Ramirez, J., Rodriguez, J. F., … Abreu, A. (2006). How many polysomnograms must sleep fellows score before becoming proficient at scoring sleep? *Journal of Clinical Sleep Medicine*, *2*, 427–430. https://doi.org/10.5664/jcsm.26659

Cohen, J. (1960). A coefficient of reliability for nominal scales. Educational and psychological measurement. *Educational and Psychological Measurement*, *20*, 37–46.

Collop, N. A. (2002). Scoring variability between polysomnography technologists in different sleep laboratories. *Sleep Medicine*, *3*, 43–47. https://doi.org/10.1016/S1389-9457(01)00115-0

Conger, A. J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, *88*, 322–328. https://doi.org/10.1037/0033-2909.88.2.322

Danker-Hopfe, H., Anderer, P., Zeitlhofer, J., Boeck, M., Dorn, H., Gruber, G. … 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.

De Zambotti, M., Godino, J. G., Baker, F. C., Cheung, J., Patrick, K., & Colrain, I. M. (2016). The boom in wearable technology: Cause for alarm or just what is needed to better understand sleep? *Sleep*, *39*, 1761–1762. https://doi.org/10.5665/sleep.6108

Dean, D. A., Goldberger, A. L., Mueller, R., Kim, M., Rueschman, M., Mobley, D., … Redline, S. (2016). Scaling up scientific discovery in sleep medicine: The national sleep research resource. *Sleep*, *39*, 1151–1164. https://doi.org/10.5665/sleep.5774

Fiorillo, L., Puiatti, A., Papandrea, M., Ratti, P.-L., Favaro, P., Roth, C., … Faraci, F. D. (2019). Automated sleep scoring: A review of the latest approaches. *Sleep Medicine Reviews*, *48*, 101204.

Grigg-Damberger, M. M. (2012). The AASM scoring manual four years later. *Journal of Clinical Sleep Medicine*, *8*, 323–332. https://doi.org/10.5664/jcsm.1928

Gwet, K. L. (2012). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among multiple raters*, 3rd ed. Gaithersburg, MD: Advanced Analytics Press.

Himanen, S. L., & Hasan, J. (2000). Limitations of Rechtschaffen and Kales. *Sleep Medicine Reviews*, *4*, 149–167. https://doi.org/10.1053/smrv.1999.0086

Iber, C., Ancoli-Israel, S. Jr, Chesson, A. L., & Quan, S. F. (2007) *The AASM Manual for the scoring of sleep and associated events: Rules, terminology and technical specifications*. Westchester, Illinois: American Academy of Sleep Medicine.

Kaplan, R. F., Wang, Y., Loparo, K. A., Kelly, M. R., & Bootzin, R. R. (2014). Performance evaluation of an automated single-channel sleep-wake detection algorithm. *Nature and Science of Sleep*, *6*, 113–122.

Koupparis, A. M., Kokkinos, V., & Kostopoulos, G. K. (2014). Semi-automatic sleep EEG scoring based on the hypnospectrogram. *Journal of Neuroscience Methods*, *221*, 189–195. https://doi.org/10.1016/j.jneumeth.2013.10.010

Ktonas, P. Y., & Smith, J. R. (1976). Semi-automatic analysis of rapid eye movement (REM) patterns: A software package. *Computers and Biomedical Research, an International Journal*, *9*, 109–124. https://doi.org/10.1016/0010-4809(76)90034-3

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

Magalang, U. J., Chen, N.-H., Cistulli, P. A., Fedson, A. C., Gíslason, T., Hillman, D., … Pack, A. I. (2013). Agreement in the scoring of respiratory events and sleep among international sleep centers. *Sleep*, *36*, 591–596. https://doi.org/10.5665/sleep.2552

Malhotra, A., Younes, M., Kuna, S. T., Benca, R., Kushida, C. A., Walsh, J., … Pien, G. W. (2013). Performance of an automated polysomnography scoring system versus computer-assisted manual scoring. *Sleep*, *36*, 573–582. https://doi.org/10.5665/sleep.2548

Morgenthaler, T. I., Deriy, L., Heald, J. L., & Thomas, S. M. (2016). The evolution of the AASM clinical practice guidelines: Another step forward. *Journal of Clinical Sleep Medicine*, *12*, 129–135. https://doi.org/10.5664/jcsm.5412

Penzel, T., Zhang, X., & Fietze, I. (2013). Inter-scorer reliability between sleep centers can teach us what to improve in the scoring rules. *Journal of Clinical Sleep Medicine*, *9*, 89–91. https://doi.org/10.5664/jcsm.2352

Pittman, S. D., MacDonald, M. M., Fogel, R. B., Malhotra, A., Todros, K., Levy, B., … White, D. P. (2004). Assessment of automated scoring of polysomnographic recordings in a population with suspected sleep-disordered breathing. *Sleep*, *27*, 1394–1403. https://doi.org/10.1093/sleep/27.7.1394

Popovic, D., Khoo, M., & Westbrook, P. (2014). Automatic scoring of sleep stages and cortical arousals using two electrodes on the forehead: Validation in healthy adults. *Journal of Sleep Research*, *23*, 211–221. https://doi.org/10.1111/jsr.12105

Redline, S., Amin, R., Beebe, D., Chervin, R. D., Garetz, S. L., Giordani, B., … Ellenberg, S. (2011). The Childhood Adenotonsillectomy Trial (CHAT): Rationale, design, and challenges of a randomized controlled trial evaluating a standard surgical procedure in a pediatric population. *Sleep*, *34*, 1509–1517. https://doi.org/10.5665/sleep.1388

Redline, S., Dean, D. 3rd, & Sanders, M. H. (2013). Entering the era of "big data": Getting our metrics right. *Sleep*, *36*, 465–469. https://doi.org/10.5665/sleep.2524

Redline, S., Schluchter, M. D., Larkin, E. K., & Tishler, P. V. (2003). Predictors of longitudinal change in sleep-disordered breathing in a nonclinic population. *Sleep*, *26*, 703–709. https://doi.org/10.1093/sleep/26.6.703

Rosenberg, R. S., & Van Hout, S. (2013). The American Academy of Sleep Medicine inter-scorer reliability program: Sleep stage scoring. *Journal of Clinical Sleep Medicine*, *9*, 81–87. https://doi.org/10.5664/jcsm.2350

Schulz, H. (2008). Rethinking sleep analysis. *Journal of Clinical Sleep Medicine*, *4*, 99–103. https://doi.org/10.5664/jcsm.27124

Silber, M. H., Ancoli-Israel, S., Bonnet, M. H., Chokroverty, S., Grigg-Damberger, M. M., Hirshkowitz, 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

Stephansen, J. B., Olesen, A. N., Olsen, M., Ambati, A., Leary, E. B., Moore, H. E., … Mignot, E. (2018). Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nature Communications*, *9*, 5229. https://doi.org/10.1038/s41467-018-07229-3

Sun, H., Jia, J., Goparaju, B., Huang, G.-B., Sourina, O., Bianchi, M. T., & Westover, M. B. (2017). Large-scale automated sleep staging. *Sleep*, *40*(10). https://doi.org/10.1093/sleep/zsx139

Van Dongen, H. P., Vitellaro, K. M., & Dinges, D. F. (2005). Individual differences in adult human sleep and wakefulness: Leitmotif for a research agenda. *Sleep*, *28*, 479–496. https://doi.org/10.1093/sleep/28.4.479

Virkkala, J., Hasan, J., Varri, A., Himanen, S. L., & Muller, K. (2007). Automatic sleep stage classification using two-channel electro-oculography. *Journal of Neuroscience Methods*, *166*, 109–115. https://doi.org/10.1016/j.jneumeth.2007.06.016

Wang, Y., Loparo, K. A., Kelly, M. R., & Kaplan, R. F. (2015). Evaluation of an automated single-channel sleep staging algorithm. *Nature and Science of Sleep*, *7*, 101–111. https://doi.org/10.2147/NSS.S77888

Whitney, C. W., Gottlieb, D. J., Redline, S., Norman, R. G., Dodge, R. R., Shahar, E., … 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

Younes, M., Raneri, J., & Hanly, P. (2016). Staging sleep in polysomnograms: Analysis of inter-scorer variability. *Journal of Clinical Sleep Medicine*, *12*, 885–894. https://doi.org/10.5664/jcsm.5894

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

How to cite this article: Berthomier C, Muto V, Schmidt C, et al. Exploring scoring methods for research studies: Accuracy and variability of visual and automated sleep scoring. *J Sleep Res*. 2020;00:e12994. https://doi.org/10.1111/jsr.12994