Sleep and Breathing
https://doi.org/10.1007/s11325-019-01801-x

METHODS • ORIGINAL ARTICLE

Interrater agreement between American and Chinese sleep centers according to the 2014 AASM standard

Shujian Deng[1,2,3,4] • Xin Zhang[1,2,3,4] • Ying Zhang[1,2,3,4] • He Gao[5] • Eric I-Chao Chang[6] • Yubo Fan[1,2,3,4] • Yan Xu[1,2,3,4,6,7]

Received: 22 November 2018 / Revised: 16 January 2019 / Accepted: 6 February 2019

Abstract

Objectives To determine inter-lab reliability in sleep stage scoring using the 2014 American Academy of Sleep Medicine (AASM) manual. To understand in-depth reasons for disagreement and provide suggestions for improvement.

Methods This study consisted of 40 all-night polysomnographys (PSGs) from different samples. PSGs were segmented into 37,642 30-s epochs. Five doctors from China and two doctors from America scored the epochs following the 2014 AASM standard. Scoring disagreement between two centers was evaluated using Cohen's kappa (κ). After visual inspection of PSGs of deviating scorings, potential disagreement reasons were analyzed.

Results Inter-lab reliability yielded a substantial degree ($\kappa = 0.75 \pm 0.01$). Scoring for stage W ($\kappa = 0.89$) and R ($\kappa = 0.87$) achieved the highest agreement, while stage N1 ($\kappa = 0.45$) reflected the lowest. Considering the relative disagreement ratio, N2-N3 (22.09%), W-N1 (19.68%), and N1-N2 (18.75%) were the most frequent combinations of discrepancy. American and Chinese doctors showed certain characteristics in the scoring of discrepancy combination W-N1, N1-N2, and N2-N3. There are seven reasons for disagreement, namely "on-threshold characteristic" (29.21%), "context influence" (18.06%), "characteristic identification difficulty" (8.81%), "arousal-wake confusion" (7.57%), "derivation inconsistence" (2.15%), "on-borderline characteristic" (0.92%), and "misrecognition" (33.27%).

Conclusions This study demonstrated the sleep stage scoring agreement of the 2014 AASM manual and explored potential sources of labeling ambiguity. Improvement measures were suggested accordingly to help remove ambiguity for scorers and improve scoring reliability at the international level.

Keywords Sleep stage scoring · AASM manual · Interrater reliability (IRR) · Polysomnography (PSG) · Discrepancy

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11325-019-01801-x) contains supplementary material, which is available to authorized users.

✉ Yan Xu
xuyan04@gmail.com

[1] School of Biological Science and Medical Engineering and Research Institute, Beihang University, Shenzhen, China

[2] Key Laboratory of Biomechanics and Mechanobiology of Ministry of Education, Beihang University, Beijing 100191, China

[3] State Key Laboratory of Software Development Environment, Beihang University, Beijing 100191, China

[4] Beijing Advanced Innovation Centre for Biomedical Engineering, Beihang University, Beijing 100191, China

[5] Clinical Sleep Medicine Center, The General Hospital of the Air Force, Beijing 100142, China

[6] Microsoft Research Asia, Beijing 100080, China

[7] Beihang University, Xueyuan Road No. 37, Beijing, China

Abbreviations

AASM	American Academy of Sleep Medicine
PSG	Polysomnography
IRR	Interrater reliability
R & K	Rechtschaffen and Kales
N2	NREM2
R	REM
W	Wake
N1	NREM1
N3	NREM3
REM	Rapid eye movement
EOG	Electrooculogram
EEG	Electroencephalographic
EMG	Electromyogram
SWA	Slow-wave activity
SEM	Slow eye movement
EM	Eye movement
KC	K complex
SS	Sleep spindle

Published online: 19 February 2019

Introduction

In 2007, the American Academy of Sleep Medicine (AASM) published the AASM sleep stage scoring manual [1]. This manual gradually replaced the one published by Rechtschaffen and Kales (R & K standard) [2, 3], becoming the gold standard for the definition of sleep stages. Due to the subjectivity of visual scoring, it is necessary to research the reliability of the scoring manual and figure out issues for future improvements [4–9]. A number of studies [10–12] have investigated the inter-scorer reliability of the 2007 AASM manual statistically. Despite the variety of their datasets and scorers, the interrater reliabilities (IRRs) were about substantial ($0.61 \leq \kappa \leq 0.80$) according to Landis and Koch's arbitrary classification of κ [13]. However, the above research has not focused on root cause analysis of the scoring discrepancy. Rosenberg et al. [14] supposed that the agreement at sleep transitions would be low, probably due to the difficulty in recognizing key waveforms. However, results revealed that only the transition from stage NREM2 (N2) to REM (R) follows the expected pattern. The study also listed epochs with the highest disagreement in transition epochs, as well as corresponding explanations for discrepancy. Nonetheless, without visual inspection of the PSGs, the provided causes might be misleading and subjective, making the advised modifications unreliable.

The AASM manual is renewed almost annually. Although most of the framework is retained, slight revisions can influence sleep stage judgements. We have not found any discussions on scoring discrepancies for the 2014 AASM manual.

This article focuses on the inter-lab reliability of the 2014 AASM scoring manual. Scoring results were given by scorers from the USA and China. Besides statistical analysis of the scoring agreement, PSG observation towards controversial epochs and analysis of the causes for discrepancy were done. Accordingly, solutions were put forward for future improvements of the manual. Validation experiments were conducted on parts of the solutions afterwards. In addition, Chinese doctors' intra-lab reliability was also briefly explored.

Methods

The Air Force General Hospital sleep center (Beijing, China) and Compumedics Regional Sleep Disorder Center (Charlotte, USA) were involved in this research. Forty Chinese subjects (29 males and 11 females) were included in this study. Thirty-seven thousand six hundred forty-two epochs were analyzed. Full night PSGs were collected using Compumedics E-series (an attended in-laboratory PSG system) in the Chinese Air Force General Hospital sleep center over 2 months in 2015. The ethical committee of the Air Force General Hospital approved the study. Ten channels' biosignals were used for the sleep stage classification, including two electrooculogram (EOG) leads (E1-M2, E2-M2), six electroencephalographic (EEG) leads (F3-M2, F4-M1, C3-M2, C4-M1, O1-M2, O2-M1), and two submental electromyogram (EMG) leads (Chin1-Chin3, Chin2-Chin3) [15].

Seven sleep technologists (five Chinese and two Americans who had at least 2 years of experience in sleep stage scoring using AASM standard) conducted scoring. All the scorers followed the 2014 AASM standard without uniform interpretation. Five stages were distinguished: Wake (W), REM (R), NREM1 (N1), NREM2 (N2), and NREM3 (N3). Thirty-second epochs were presented on a computer display, with amplitude markers at ± 37.5 μV assisting the measurement of slow-wave activity (SWA). Five Chinese doctors assigned stages independently. Their majority votes were provided for inter-lab comparison. Two American doctors applied consensus scoring, then generated only one copy of scoring results [16].

After waveform analysis by PSG observation, re-judgment of some discrepant epochs was done. Figures containing 10 channels of biosignal (with amplitude and time markers) were plotted using MATLAB. Scorers were asked to mark key points (e.g., peak and borders of SWA) using "data cursor" function of MATLAB. MATLAB code is available at https://github.com/emergencyd/SLEEP.

The degree of agreement between American and Chinese doctors was analyzed at three levels: (1) visual inspection of sample hypnograms; (2) statistical analysis of overall, stage-specific, and discrepancy-combination-specific consistency; and (3) waveform analysis of discrepant epochs. Statistical analysis was also conducted for the evaluation of Chinese intra-lab scoring reliability.

Results

Hypnograms are shown in Fig. 1. It could be observed that there were some uncertainties among sleep stage scores despite a general agreement. The scoring differences among five Chinese doctors and between American and Chinese doctors represent intra- and inter-lab variability, respectively. Statistical comparison at epoch level was then conducted for further validation. Unless otherwise stated, (1) "CN" denotes China and "US" denotes America in all the figures and tables; (2) the default confidence interval (CI) is 95%.

Agreement between American and Chinese scorers

Comparing sleep stages given by American and Chinese centers, the overall level of agreement for 40 subjects was 82.06% (Cohen's kappa $\kappa = 0.75 \pm 0.01$). According to Fig. 2, the range of kappa was wide (0.55–0.90), yielding a moderate ($0.41 \leq \kappa \leq 0.60$) to perfect ($0.81 \leq \kappa \leq 1.00$) agreement

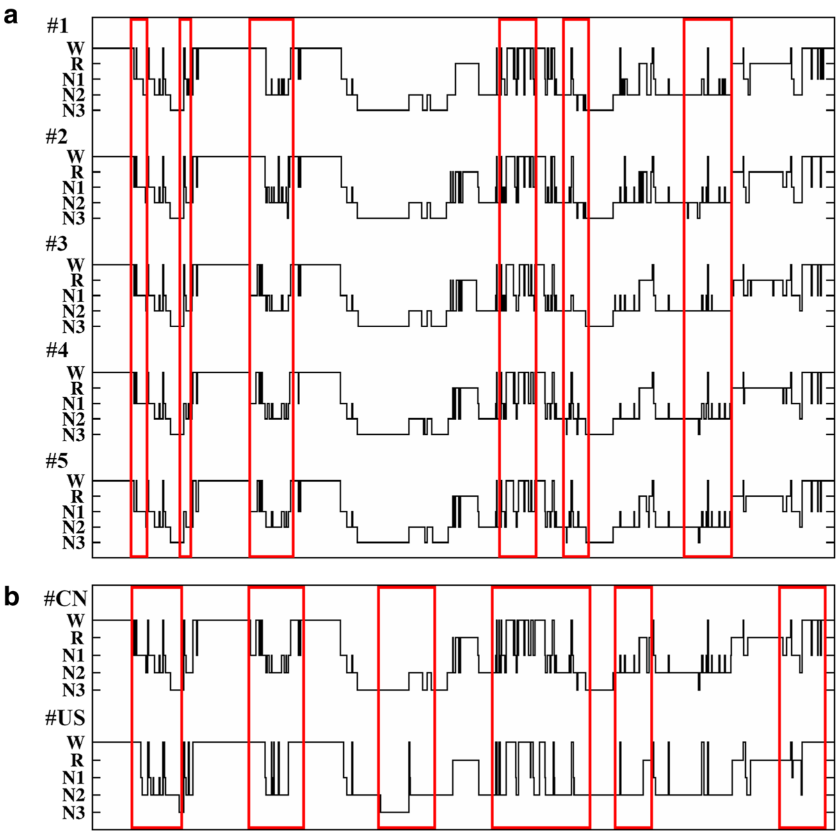

Fig. 1 Hypnograms of a 21-year-old male sample without OSA, which were (**a**) scored by 5 Chinese doctors independently; and (**b**) majority vote of 5 Chinese scores and consensus score from two American doctors. Some scoring disagreements were marked by red boxes

[13]. Figure 2 also shows the distribution of stage-specific individual kappa.

A sleep stage-specific analysis of the scoring discrepancy was also conducted. Table 1 shows the number of epochs for five stages scored by American and Chinese centers. Most of the epochs were labeled as stage N2 (43.39%), while only 8.48% were labeled as N1. This sleep architecture was similar to the results of Mitterling et al. [17].

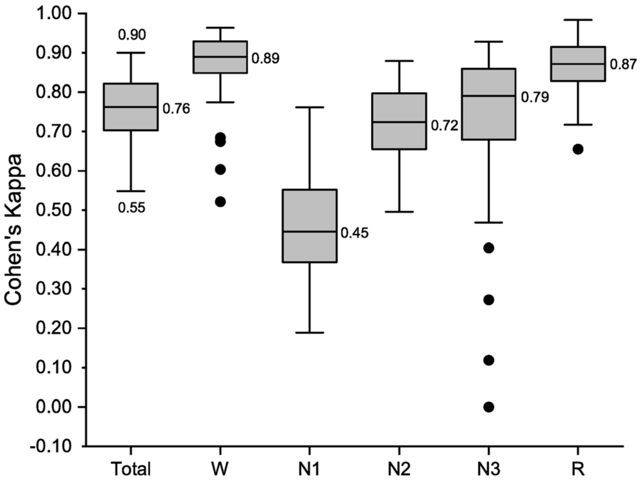

Fig. 2 Boxplot of recording-specific and stage-specific Cohen's kappa ($n = 40$ subjects), where medians, maximum, and minimum were marked

Taking length differences of five sleep stages into consideration, the relative disagreement ratio β for combination of stage M and N is calculated as:

$$\beta = \frac{2 \times L_{M-N}}{L_M + L_N} \quad (1)$$

where L_M, L_N, and L_{M-N} denote the number of epochs for stages M and N and the discrepancy combination M-N, respectively.

Inter-lab epoch-by-epoch comparison for specific stages is detailed in Table 2.

Combination M-N includes two cases—an epoch was labeled as: (1) stage M by Chinese scorers and N by American

Table 1 Number of epochs (L) classified as five sleep stages, with percentage of all stages in parentheses

Stage	Chinese	American	Average
W	6462 (17.17%)	7175 (19.06%)	6818.5 (18.11%)
R	5658 (15.03%)	5736 (15.24%)	5697 (15.13%)
N1	4098 (10.89%)	2289 (6.08%)	*3193.5 (8.48%)*
N2	15,203 (40.39%)	17,465 (46.40%)	*16334 (43.39%)*
N3	6221 (16.53%)	4977 (13.22%)	5599 (14.87%)

Values in italics indicate the most and least average numbers

Table 2 Combinations of discrepancies between American and Chinese centers: number of epochs L_{M-N} (upper triangle matrix); relative variance β (lower triangle matrix)

Stage	W	R	N1	N2	N3
W	–	93	985	178	45
R	1.49%	–	388	790	3
N1	*19.68%*	8.73%	–	1831	17
N2	1.54%	7.17%	*18.75%*	–	2423
N3	0.72%	0.05%	0.39%	*22.09%*	–

Values in italics indicate the highest deviation ratios

scorers; and (2) stage N by Chinese scorers and M by American scorers. Directed discrepancy combinations are provided in Table 3 to evaluate the scoring preference of scorers from two centers. W-N1, N1-N2, and N2-N3 were obvious unbalanced pairs.

Waveform analysis for discrepancy epochs

Three thousand one hundred fifty-six controversial epochs from 20 samples (15 healthy subjects and 5 OSA patients) were randomly selected. Visual inspection of the corresponding PSGs was carried out. Table 4 summarizes seven discrepancy reasons and their proportions in all cases: (1) On-threshold characteristic: the case when characteristics of a waveform reached a critical point and became hard to measure with limited time and much noisy signals (Fig. 3); (2) context influence: when there was a discrepancy for definite stage scoring, epochs around it were also scored differently (Fig. 4); (3) characteristic identification difficulty: the difficulty in distinguishing a characteristic wave from complex background EEG signals (Fig. 5); (4) arousal-wake confusion: some scorers tended to mark epochs with much arousal or high EMG as stage W (Fig. 6); (5) derivation inconsistence: discrepancy occurred when signals from different derivations presented different waveforms (Fig. 7); (6) on-borderline characteristic: confusion happened when characteristic waveforms of N1 or N2 crossed up the middle of the segment due to the continuing or preceding rules (Fig. 8); (7) misrecognition: obvious scoring errors.

Table 3 Directed combinations of discrepancies between American and Chinese centers: number of epochs

CN	US				
	W	R	N1	N2	N3
W	–	45	*165*	82	2
R	48	–	138	411	1
N1	*820*	250	–	*1438*	7
N2	96	379	*393*	–	*612*
N3	43	2	10	*1811*	–

Values in italics indicate the obvious unbalanced pairs

Rejudging with quantitative assistance

"On-threshold characteristic" and "on-borderline characteristic" are difficult to measure without the assistance of quantitative tools. After categorizing discrepancy epochs, two technologists from America and China relabeled 951 epochs whose difference reasons are "on-threshold characteristic" and "on-borderline characteristic." Rather than giving a final decision for a single epoch, they marked key points of the characteristic waveforms within 5 s. After that, we calculated the amplitude, frequency, duration, and location of the marked waveforms, then scored the epoch to a certain stage accordingly. Table 5 showed the scoring agreement.

Agreement among Chinese scorers

The agreement within five Chinese doctors was also analyzed statistically to evaluate intra-lab reliability. The overall level of agreement was 86.01% (Fleiss' kappa $\kappa = 0.86 \pm 0.00$). The kappa range of 40 subjects was wide (0.46–0.91), as detailed in Fig. 9. Thirty-seven subjects yielded a perfect ($0.81 \leq \kappa \leq 1.00$), 2 yielded a substantial, and 1 yielded a moderate ($0.41 \leq \kappa \leq 0.60$) agreement [13].

A sleep stage-specific analysis of the scoring discrepancy was also conducted. The distribution of stage-specific individual kappa is displayed in Fig. 9.

Discussion

This study first evaluated the scoring agreement between American and Chinese sleep centers. Cohen's kappa was calculated for the evaluation of inter-lab reliability. A substantial agreement was found for the pool of all epochs, which agreed with the results of Danker-Hopfe et al. [10]. Despite the modifications that the 2014 AASM manual has made on the 2007 AASM manual, little overall reliability improved. Because this research differs from previous ones in terms of the dataset, labeling settings, etc., further research is required to figure out the influence of guidance version on deviating scoring rate.

When looking at five sleep stages separately (Fig. 2), it emerged that the inter-lab agreement was the worst for N1, whose median kappa barely fell into a moderate interval. On the contrary, both stage W and R reached a perfect agreement. This order was similar to the ranking (W and R changed position) of Danker-Hopfe et al. [10]. With regard to the extremely low agreement for stage N1, Danker-Hopfe et al. [10] and Basner et al. [4] proposed two hypotheses: (1) transition from W to N1 is difficult to identify, especially for those

Table 4 Detailed difference reasons for deviating combinations: number of epochs. Percentages of all reasons were listed inside the parentheses in the last line

	On-threshold characteristic			Context influence	Characteristic identification difficulty		Arousal-wake confusion	Derivation inconsistence		On-borderline characteristic		Misrecognition	Total
N2-N3	SWA	Amplitude	267	–	–		–	SWA	66	–		266	1177
		Frequency	315										
		Duration	263										
W-N1	(Alpha and EM) duration		45	–	–		133	–		–		194	372
N1-N2	(Arousal) duration		10	319	Arousal	94	–	–		Arousal	29	385	939
					KC	90							
					SS	12							
N1-R	(Arousal) duration		4	73	REM	9	–	EMG	2	–		82	180
	(High EMG) duration		7		Arousal	3							
N2-R	–			160	KC	16	–	–		–		62	289
					REM	42							
					SS	9							
N2-W	–			8	KC	3	63	–		–		25	99
N3-W	–			–	–		29	–		–		4	33
N1-N3	SWA	Amplitude	2	–	–		–	–		–		9	15
		Frequency	4										
R-W	(High EMG) duration		5	10	–		14	–		–		23	52
Total	922 (29.21%)			570 (18.06%)	278 (8.81%)		239 (7.57%)	68 (2.15%)		29 (0.92%)		1050 (33.27%)	3156

SWA slow-wave activity, *EM* eye movement, *KC* K complex, *SS* sleep spindle, *REM* rapid eye movement

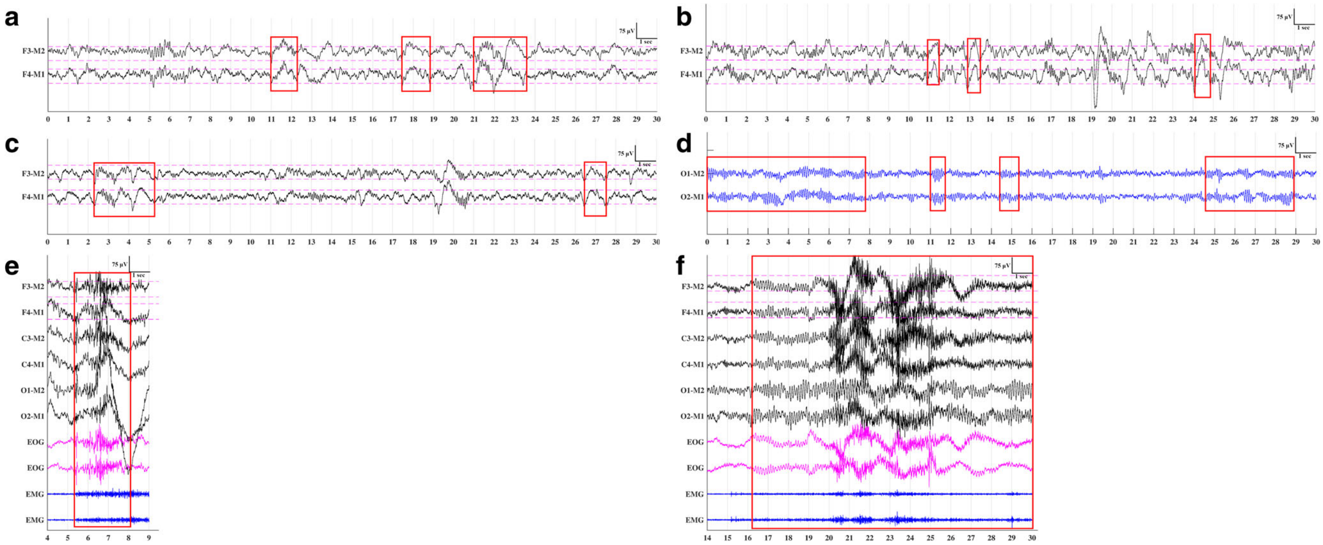

Fig. 3 **a** N2 (CN) vs. N3 (US), waves highlighted by red boxes are slow waves. The total duration is about 5.5–6 s, while the border of the slow-wave activity (SWA) blurs. **b** N3 (CN) vs. N2 (US), waves highlighted by red boxes are suspected slow waves. Their frequencies are about 1–2 Hz, while in most cases, the frequency of slow wave is 0.5–1 Hz. **c** N2 (CN) vs. N3 (US), waves highlighted by red boxes are suspected slow waves. Their amplitudes are more or less 75 μV, making these waves difficult to classify. **d** W (CN) vs. N1 (US), alpha rhythm (highlighted by red boxes) roughly lasts 15 s, with indistinct borders and discrete distributions. **e** N1 (CN) vs. N2 (US), suspected arousal activity (highlighted by red boxes) roughly lasts 3 s. **f** N1 (CN) vs. R (US), arousal and raise of EMG tone (highlighted by the red box) roughly lasts 14–15 s, nearly half the epoch. Slow eye movements (SEMs) could be observed in the last half of this epoch

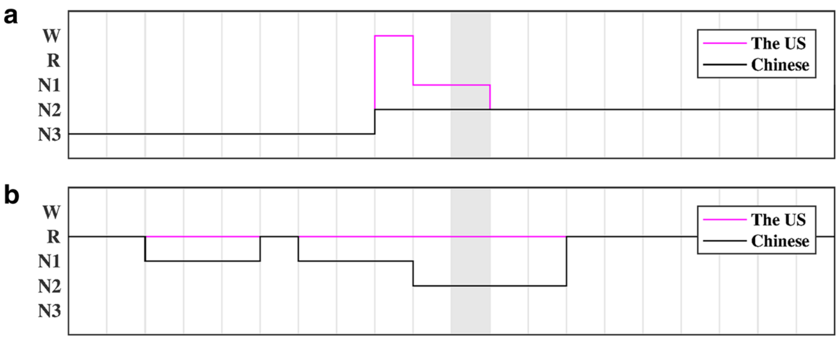

Fig. 4 Hypnogram fragments. The shaded epoch was scored as **a** N2 (CN) and N1 (US) and **b** N2 (CN) and R (US). Both of them followed the labels of the preceding or/and following epochs

from people who generated little or no α activity; and (2) low proportion of N1 sleep results in low kappa, which was introduced by Danker-Hopfe et al. [5] as well. The first hypothesis is reasonable. When looking at relative variance β in Table 2, disagreement for W-N1 was slightly greater than N1-N2. However, we could also find that stage N1 was mainly confused with N2, then W. This paradox is the result of the extremely imbalanced distribution of sleep stages, since almost half the epochs were labeled as stage N2, while less than one-fifth were stage W (as shown in Table 1). When discussing the low agreement evaluated by kappa, the number of discrepant epochs rather than β should be decisive. The research of Rosenberg et al. [14] also demonstrated that the agreement in W-N1 transitions was better than the average for stage N1, which revealed that the first hypothesis did not seem to be the dominant reason. But still, the tendencies to confuse N1 with W or N2 were almost the same. The second hypothesis was soon rejected by Danker-Hopfe et al. [10] themselves, since the highest proportion of N2 did not yield the highest kappa, which was also evidenced by our research. N1 is the only stage with almost no characteristic grapho-elements. Besides the role in the initial sleep onset, N1 behaves as a transitional stage throughout the night [18]. Therefore, a high degree of disagreement for N1 is primarily due to the ambiguity in recognizing other stages (especially stage W and N2), including but not limited to transition from W to N1.

Taking the total duration of a particular stage into consideration, N2-N3, N1-N2, and W-N1 showed the greatest disagreement (presented in bold type in Table 2). A possible explanation for this phenomenon is that sleep development is gradual, so characteristics between adjacent stages might overlap. This was also discovered by Rosenberg et al. [14]. Table 2 revealed the lowest deviation ratio for W-R, W-N2, W-N3, N3-R, and N1-N3. This order of discrepancy combination was basically in concordance with the result of Danker-Hopfe et al. [10]. It can be observed from Table 2 that stage N3 was seldom confused with other stages except for N2. Likewise, stage W was predominantly confused with N1. The probable reasons are (1) characteristics of both stage W and N3 are highly discriminative from most other stages; and (2) they do not have preceding or continuation rules.

A pilot study for interrater reliability (IRR) research, PSG observation on difference epochs, was also done. Except for "misrecognition," detailed reasons for scoring disagreement regarding particular deviation combinations are elaborated below.

N2-N3 had the highest β. Most of the deviations were caused by "on-threshold characteristic", since the definition

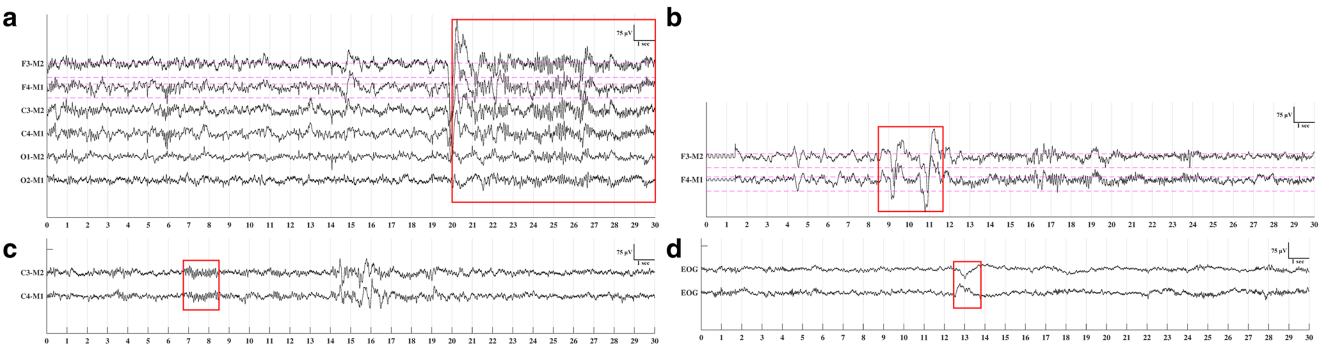

Fig. 5 **a** This epoch was scored as stage N2. The subsequent epoch was scored as N1 (CN) and N2 (US). Frequency change can be observed in the last half of the epoch, as highlighted by the red box, but is not a big difference compared with the previous signal. **b** N1 (CN) vs. N2 (US), waves highlighted by red boxes are suspected K complexes, the shapes of which do not exactly meet the requirement of negative-positive patterns. **c** N1 (CN) vs. N2 (US), wave highlighted by the red box is suspected sleep spindle (SS), the shape of which is not strictly sinusoidal. **d** This epoch was scored as R (CN) and N2 (US). Wave highlighted by the red box is suspected REM, the tone of which is low and thus less distinguishable from background activity

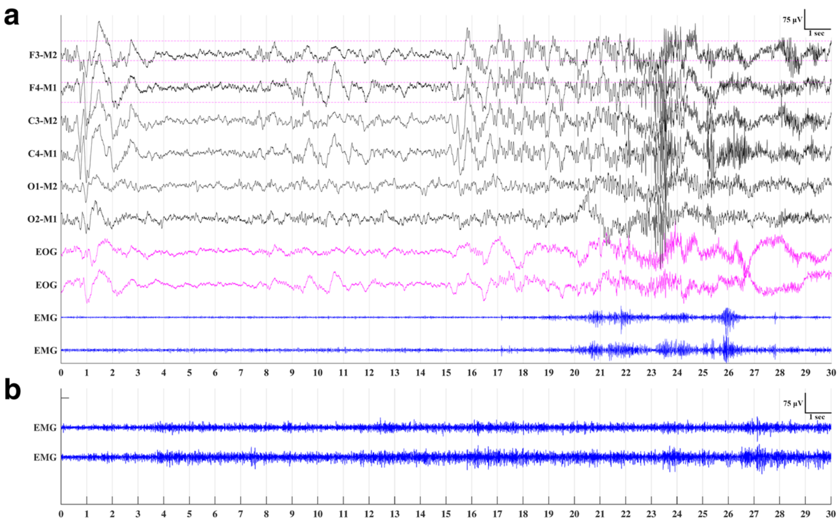

Fig. 6 **a** N2 (CN) vs. W (US), much arousal could be observed in the last half of the epoch. **b** N1 (CN) vs. W (US), EMG is high in most parts of this segment, and slow eye movements (SEMs). SEMs could be observed in full image (in Supplymentary material)

of stage N3 relies on the recognition and duration measurement of SWAs. The judgment of SWA strongly depends on the amplitude and frequency of the waveform. It can be concluded that clear stipulation of a particular waveform does not necessarily bring a higher agreement. "Derivation inconsistence" would influence reliability as well, because fine distinctions could influence the feature measurement. When facing ambiguity, the Chinese doctors tended to score N3 while the American doctors preferred N2 (Table 3).

W-N1 had the second highest β, mostly due to the "arousal-wake confusion." Arousal is considered sleep disruption, which indicates the ending of stage N2 or R and a switch to a shallower sleep state—usually stage N1 sleep [19]. According to R & K criteria, an arousal could be scored as a wake state [3]. Although not supported by the AASM manual, scoring epochs with much arousal as stage W might be reasonable. The same goes for epochs with an extremely high chin EMG tone. In addition, the "on-threshold characteristic" was also a source for confusion, as the definition of stage W relies on the duration of α activity and eye movement. When facing ambiguity, the Chinese doctors tended to score N1 while the American doctors preferred W (Table 3).

N1-N2 had the third highest β. (1) This was mainly a result of "context influence." Epochs following a controversial stage N2 would also be scored differently in the absence of other evidence. (2) "Characteristic identification difficulty" was also a big problem. Concise and clear as the definitions in the guideline are, PSG signal in reality is rather complex. Divergence in the judgment of K complex would happen, when the shape of the suspected waveform did not meet the criteria "negative...followed by positive." The sleep spindle also faced the challenge of shape recognition, due to the lack of strict restrictions on "sinusoidal" [20]. On the grounds that no quantitative criteria is made for SS, the classification could become quite subjective. However, additional criteria with precise measurement might not necessarily bring higher reliability because of the limitation in visual scoring, taking the definition of SWA as an example (also pointed out by Rosenberg et al. [14]). Moreover, disagreement existed on the classification of arousal (a crucial influence factor on whether or not to end stage N2), especially when the frequency shift was not significant. (3) Another reason for the discrepancy of N1-N2 was "on-borderline characteristic," since both stage N1 and N2 strictly follow the rule that the present

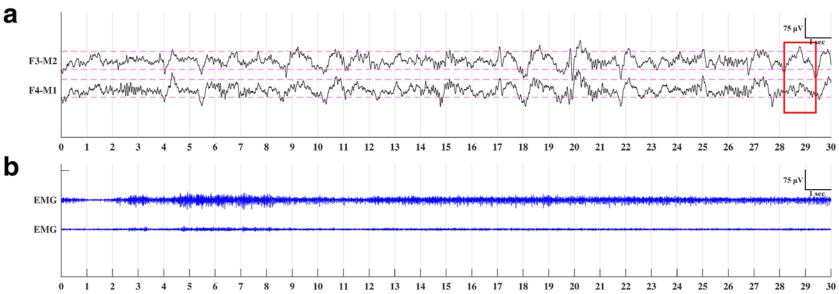

Fig. 7 **a** N2 (CN) vs. N3 (US), wave highlighted by the red box could be determined as a slow-wave activity (SWA) according to the "F3-M2" derivation, but the decision could be quite the opposite based on the "F4-M1" derivation. **b** This epoch was scored as N1 (CN) and R (US). Only upper chin EMG derivation demonstrated high tone

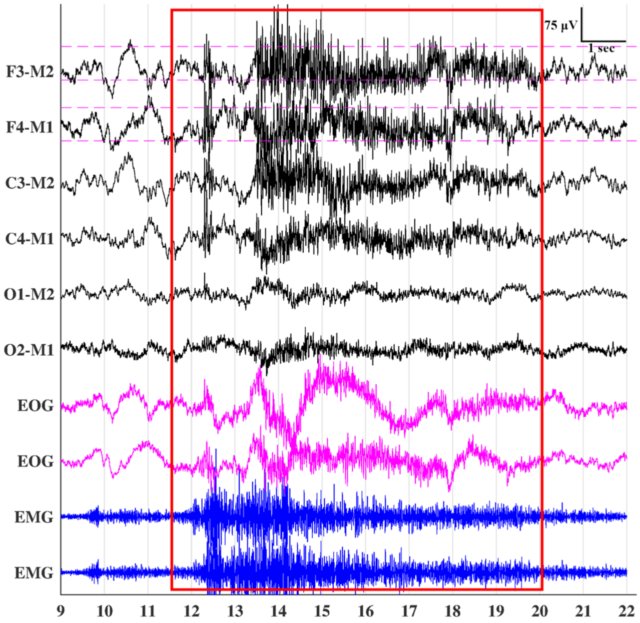

Fig. 8 This epoch was scored as N1 (CN) and N2 (US), while the preceding epoch was scored as N2. Arousal could be observed in the middle of this epoch, as highlighted by the red box

epoch is defined by the featured waveform in its first half, or in the last half of the preceding epoch. (4) The "on-threshold characteristic" was another cause, because the classification of arousal needs a duration measurement. When facing ambiguity, the Chinese doctors tended to score N1 while the American doctors preferred N2 (Table 3).

N1-R had a much lower β. "Context influence" must not be neglected, for the reason that stage R has preceding and continuation rules. "Characteristic identification difficulty" was observed when suspected arousals had inconspicuous frequency shifts, or suspected REM had low amplitudes. "On-threshold characteristic" was also a source for disagreement, because the duration of high EMG or suspected arousal could reach a critical point. "Derivation inconsistence" occurred when signal tones from two EMG derivations were completely opposite.

N2-R was primarily caused by "context influence," considering that both stage **N2** and R have continuation or preceding rules. "Characteristic identification difficulty" also existed in the recognition of K complex, SS, and REMs.

To sum up:

(1) "On-threshold characteristic" occurred when scorers needed to accurately measure the following:

   a. Slow-waves' duration (Fig. 3a), or a wave's frequency (Fig. 3b) and amplitude (Fig. 3c) to score N3
   b. Duration of α rhythm and eye movement to score W (Fig. 3d)
   c. Duration of frequency changes to score arousal (Fig. 3e)
   d. Duration of different stages in an epoch to score the primary stage (Fig. 3f)

(2) "Context influence" mainly happened in the confusion of stage N2/R due to their continuation and/or preceding rules

(3) "Characteristic identification difficulty" including the determination of the following:

   a. Arousal when the frequency change was not abrupt (Fig. 5a)
   b. K complex when the shape of a suspected wave was not strictly biphasic (Fig. 5b)
   c. Sleep spindle while the shape of a suspected wave was not sinusoidal (Fig. 5c)
   d. REM when the amplitude of a suspected wave was rather low (Fig. 5d)

(4) "Arousal-wake confusion" could lead to the discrepancy in the judgment of W

(5) "Derivation inconsistence" would influence the judgment of the following:

   a. N3 when the duration of SWAs or the tone of a suspected SWA reached a critical point (Fig. 7a)
   b. R if EMG tones from two leads showed opposite patterns (Fig. 7b)

Table 5 Second-round judgment of discrepancy epochs (caused by "on-threshold characteristic" and "on-borderline characteristic") with quantitative assistance: number of epochs and agreement rate

Discrepancy combination	Key points	Total	Agreement	Rate
On-threshold characteristic				
N2-N3	Peak and bottom boundaries of SWA	845	630	
W-N1	Boundaries of alpha rhythm and REM	45	35	
N1-N2	Boundaries of arousal	10	9	
N1-R	Boundaries of high EMG	11	9	
N1-N3	Peak and bottom boundaries of SWA	6	6	
R-W	Boundaries of high EMG	5	3	
On-borderline characteristic				
N1-N2	Boundaries of arousal	29	25	
Total		951	717	75.39%

SWA slow-wave activity, *REM* rapid eye movement

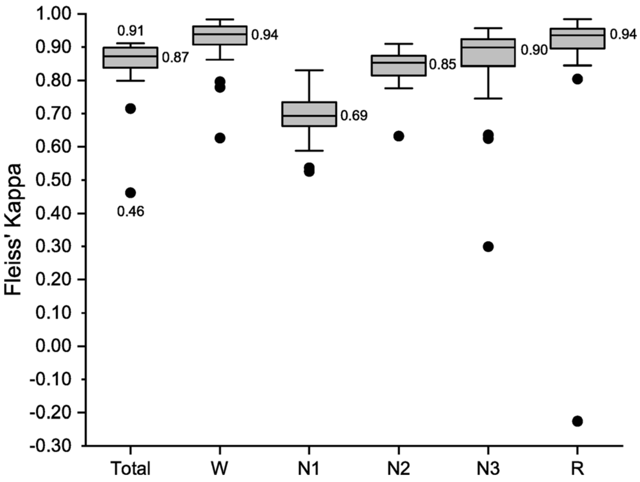

Fig. 9 Boxplot of recording-specific and stage-specific Fleiss' kappa ($n = 40$ subjects), where medians, maximum, and minimum were marked

(6) "On-borderline characteristic" had a significant impact on the scoring of N2 and N1, since they were determined by the characteristic waveforms in the first half of the epoch (Fig. 8)

(7) "Misrecognition" was the most common case. It was possible that doctors tended to label a single stage consecutively in light of sleep completeness. When most of the epochs were labeled as a particular stage during a period, few exceptions were ignored. Additionally, it was presumably due to scorers' fatigue or carelessness.

Possible measures to deal with these cases were proposed. First, it might help if scorers could mark the key points of a certain waveform (such as the peak and the bottom boundary points of SWA), then let the software system calculate the corresponding frequency, amplitude, or duration automatically. In this way, much disagreement caused by "on-threshold characteristic" might be avoided. By marking key points, characteristic waves could also be located, thus reducing "on-borderline characteristic." As verified in Table 5, with quantitative tools calculating the waveform length, amplitude, duration, and location, scoring results of two technologists agreed on 75.39% of previously controversial epochs, even though the boundary points were hard to mark accurately under complex background signals. Second, commonly used scoring software would need the scorers to label every epoch. With the existence of preceding and continuation rules, scorers could just label definite stages, and let the software system replenish the remaining epochs. It might effectively reduce "context influence" when conducting consensus scoring, as the source epoch of a series of unreasonable labeling would be easily pointed out by others. Furthermore, much time and effort could be saved. It would also bring convenience for further IRR research, because "context influence" could become easier to figure out. Third, "characteristic identification difficulty" and "arousal-wake confusion" could possibly be avoided by adding detailed and clear statements in the scoring manual, such as (1) strict requirement for the shape of K complex and SS; (2) quantitative evaluation of frequency shift for arousal; (3) amplitude requirement for REM; and (4) declaration of the relationship between arousal and wake, especially when an epoch contains much arousal. Fourth, the main source of deviation "misrecognition" should be addressed by intensive and standardized training. There is still room for the reassessment of the AASM scoring rules towards the solution of "derivation inconsistence." This research brings some valuable insights for future revisions.

Intra-lab reliability of five Chinese doctors was perfect incorporating all sleep stages, which was higher than that of inter-lab comparison. Note that even at the individual level, almost all subjects yielded a perfect agreement. Penzel et al. [6] stated that a certain flavor exists towards scoring of controversial epochs, which was also confirmed in our study. As shown in Table 3, American and Chinese doctors gave tendentious labels for W-N1, N1-N2, and N2-N3. It could be preliminarily assumed that normalized training could help improve IRR by universalizing the judgment when the rules are not obvious. As for sleep stage-specific analysis of intra-lab reliability, it could be observed from Fig. 9 that the agreement was best for stage W, followed by stages R, N3, N2, and N1, which was exactly the same as the order of inter-lab agreement.

This research was limited by the sample distribution. In the future, a greater diversity of subjects from multiple countries and more international scorers with different backgrounds will be included to generate a more reliable conclusion. Our research focused on the reliability of sleep stage scoring. In the future, we will also analyze other sleep events such as respiratory events and PLMs using the PSG waveform analysis method.

More examples (discrepancy epochs in EPS format) for detailed cases could be accessed on https://github.com/emergencyd/SLEEP.

Author contributions Conception and the design of the project: SD, YX. The analysis and interpretation of the data: SD, XZ, YZ, YX. Drafted, wrote, and revised the paper: SD, XZ, YZ, HG, EC, YF, YX. Provided the data: HG, EC, YF, YX.

Funding information This work was supported by the Microsoft Research under the eHealth program, the National Natural Science Foundation in China under Grant 81771910, the National Science and Technology Major Project of the Ministry of Science and Technology in China under Grant 2017YFC0110903, the Beijing Natural Science Foundation in China under Grant 4152033, the Technology and Innovation Commission of Shenzhen in China under Grant shenfagai2016-627, the Beijing Young Talent Project in China, the Fundamental Research Funds for the Central Universities of China under Grant SKLSDE-2017ZX-08 from the State Key Laboratory of Software

Development Environment in Beihang University in China, and the 111 Project in China under Grant B13003.

Compliance with ethical standards

Conflicts of interest The authors declare that they have no conflict of interest.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. This article does not contain any studies with animals performed by any of the authors.

Informed consent Informed consent was obtained from all individual participants included in the study.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. 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, vol 4849. American Academy of Sleep Medecine, Westchester
2. Himanen SL, Hasan J (2000) Limitations of Rechtschaffen and Kales. Sleep Med Rev 4(2):149–167
3. Hobson JA (1969) A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects: a Rechtschaffen and a Kales. Electroencephalogr Clin Neurophysiol 26(6):644
4. Basner M, Griefahn B, Penzel T (2008) Inter-rater agreement in sleep stage classification between centers with different backgrounds. Somnologie-Schlafforschung und Schlafmedizin 12(1): 75–84
5. Danker-Hopfe H, Kunz D, Gruber G, Klösch G, Lorenzo JL, Himanen SL, Kemp B, Penzel T, Röschke J, Dorn H et al (2004) Interrater reliability between scorers from eight European sleep laboratories in subjects with different sleep disorders. J Sleep Res 13(1):63–69
6. Penzel T, Zhang X, Fietze I (2013) Inter-scorer reliability between sleep centers can teach us what to improve in the scoring rules. J Clin Sleep Med 9(01):89–91
7. Silber MH, Ancoli-Israel S, Bonnet MH, Chokroverty S, Grigg-Damberger MM, Hirshkowitz M, Kapen S, Keenan SA, Kryger MH, Penzel T, Pressman MR, Iber C (2007) The visual scoring of sleep in adults. J Clin Sleep Med 3:121–131
8. Suzuki M, Saigusa H, Chiba S, Yagi T, Shibasaki K, Hayashi M, Suzuki M, Moriyama K, Kodera K (2005) Discrepancy in polysomnography scoring for a patient with obstructive sleep apnea hypopnea syndrome. Tohoku J Exp Med 206(4):353–360
9. Whitney CW, Gottlieb DJ, Redline S, Norman RG, Dodge RR, Shahar E, Surovec S, Nieto FJ (1998) Reliability of scoring respiratory disturbance indices and sleep staging. Sleep 21(7):749–757
10. Danker-Hopfe H, Anderer P, Zeitlhofer J, Boeck M, Dorn H, Gruber G, Heller E, Loretz E, Moser D, Parapatics S et al (2009) Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. J Sleep Res 18(1):74–84
11. Magalang UJ, Chen NH, Cistulli PA, Fedson AC, Gíslason T, Hillman D, Penzel T, Tamisier R, Tufik S, Phillips G et al (2013) Agreement in the scoring of respiratory events and sleep among international sleep centers. Sleep 36(4):591–596
12. Zhang X, Dong X, Kantelhardt JW, Li J, Zhao L, Garcia C, Glos M, Penzel T, Han F (2015) Process and outcome for international reliability in sleep scoring. Sleep Breath 19(1):191–195
13. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174
14. Rosenberg RS, Van Hout S (2013) The American academy of sleep medicine inter-scorer reliability program: sleep stage scoring. J Clin Sleep Med 9(01):81–87
15. Ruehland WR, O'Donoghue FJ, Pierce RJ, Thornton AT, Singh P, Copland JM, Stevens B, Rochford PD (2011) The 2007 AASM recommendations for EEG electrode placement in polysomnography: impact on sleep and cortical arousal scoring. Sleep 34(1):73–81
16. Hare AP. Consensus versus majority vote: A laboratory experiment. Small Group Behavior, 1980; 11(2):131-143.
17. Mitterling T, Högl B, Schönwald SV, Hackner H, Gabelia D, Biermayr M, Frauscher B (2015) Sleep and respiration in 100 healthy Caucasian sleepers—a polysomnographic study according to American Academy of Sleep Medicine standards. Sleep 38(6): 867–875
18. Carskadon MA, Dement WC et al (2005) Normal human sleep: an overview. Principles and Practice of Sleep Medicine 4:13–23
19. Parrino L, Ferri R, Zucconi M, Fanfulla F (2009) Commentary from the Italian association of sleep medicine on the AASM manual for the scoring of sleep and associated events: for debate and discussion. Sleep Med 10(7):799–808
20. Berry RB, Brooks R, Gamaldo CE et al (2014) The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications, version 2.1. American Academy of Sleep Medicine, Darien