https://doi.org/10.5664/jcsm.9538

JCSM | Journal of Clinical Sleep Medicine

SCIENTIFIC INVESTIGATIONS

Interrater reliability of sleep stage scoring: a meta-analysis

Yun Ji Lee, MD[1]; Jae Yong Lee, MD, PhD[1]; Jae Hoon Cho, MD, PhD[2]; Ji Ho Choi, MD, PhD[1]

[1]Department of Otorhinolaryngology—Head and Neck Surgery, College of Medicine, Soonchunhyang University, Bucheon Hospital, Bucheon, Korea; [2]Department of Otorhinolaryngology—Head and Neck Surgery, College of Medicine, Konkuk University, Seoul, Korea

Study Objectives: We evaluated the interrater reliabilities of manual polysomnography sleep stage scoring. We included all studies that employed Rechtschaffen and Kales rules or American Academy of Sleep Medicine standards. We sought the overall degree of agreement and those for each stage.

Methods: The keywords were "Polysomnography (PSG)," "sleep staging," "Rechtschaffen and Kales (R&K)," "American Academy of Sleep Medicine (AASM)," "interrater (interscorer) reliability," and "Cohen's kappa." We searched PubMed, OVID Medline, EMBASE, the Cochrane library, KoreaMed, KISS, and the MedRIC. The exclusion criteria included automatic scoring and pediatric patients. We collected data on scorer histories, scoring rules, numbers of epochs scored, and the underlying diseases of the patients.

Results: A total of 101 publications were retrieved; 11 satisfied the selection criteria. The Cohen's kappa for manual, overall sleep scoring was 0.76, indicating substantial agreement (95% confidence interval, 0.71–0.81; $P < .001$). By sleep stage, the figures were 0.70, 0.24, 0.57, 0.57, and 0.69 for the W, N1, N2, N3, and R stages, respectively. The interrater reliabilities for stage N2 and N3 sleep were moderate, and that for stage N1 sleep was only fair.

Conclusions: We conducted a meta-analysis to generalize the variation in manual scoring of polysomnography and provide reference data for automatic sleep stage scoring systems. The reliability of manual scorers of polysomnography sleep stages was substantial. However, for certain stages, the results were poor; validity requires improvement.

Keywords: interrater reliability, meta-analysis, sleep stage scoring

Citation: Lee YJ, Lee JY, Cho JH, Choi JH. Interrater reliability of sleep stage scoring: a meta-analysis. *J Clin Sleep Med.* 2022;18(1):193–202.

BRIEF SUMMARY

Current Knowledge/Study Rationale: The principal way to score polysomnography is manually. To increase the interrater reliability of sleep stage scoring, there have been consistent studies with a high degree of variance.

Study Impact: We conducted a meta-analysis to obtain comprehensive data and to identify the sleep stage that is most affected by manual scoring. In addition to presenting an integrated reference through this study, we summarized the points requiring improvement.

INTRODUCTION

Polysomnography (PSG) yields fundamental data on sleep architecture and aids the diagnosis of sleep disorders, which not only pose problems per se but also increase the risks of chronic diseases of the cardiovascular system and neurodegenerative diseases, including dementia and Parkinson disease.[1,2] PSG yields comprehensive information on sleep states based on various electrophysiological signals, including electroencephalogram (EEG), electro-oculogram, and electromyogram signals.

In 1968, Rechtschaffen and Kales (R&K) introduced rules for the scoring of sleep stages; manual scoring thus became systematic.[3] In 2007, the American Academy of Sleep Medicine (AASM) published an updated version of their earlier scoring rules; these are commonly used today (*The AASM Manual for the Scoring of Sleep and Associate Events: Rules, Terminology and Technical Specifications,* first edition).[4] Manual scoring is time-consuming, and scorer reliability has been questioned. Several reports have explored the extents of agreement among manual scorers.[5–11] The extent of overall agreement has varied greatly, from 61.1%–92.2%. Of the various sleep stages, the reliability of N1 identification was the lowest, ranging from 19.8%–38.18%. In this study, we explored the interrater reliability of manual sleep stage scoring and systematically reviewed the factors affecting reliability.

METHODS

Literature sources and study identification

We reviewed works on the interrater reliability of sleep stage scoring published since 1968, when the R&K manual became available. We comprehensively searched PubMed, OVID Medline, EMBASE, the Cochrane library, KoreaMed, KISS, and MedRIC for papers written in English. The keywords were "Polysomnography (PSG)," "sleep monitoring," "sleep staging," "Rechtschaffen and Kales (R&K)," "American Academy of Sleep Medicine (AASM)," "interrater (interscorer) reliability," "agreement," and "Cohen's kappa (κ)." The search was performed on June 3, 2020. After 2 reviewers independently checked all abstracts and titles, we excluded studies that were irrelevant or ineligible. The selected manuscripts were thoroughly reviewed, and data were collected.

Study selection

PSG data that were scored epoch by epoch and classified into stages were selected. We evaluated only manual (not automated) scoring. Studies of healthy adults or patients with various underlying conditions, including sleep-disordered breathing, were included. Studies of pediatric patients (younger than age 13 years) were excluded. Scorer histories and nationalities varied among the studies; we imposed no restriction.

Data collection

We extracted the number of PSGs and that of the epochs, the underlying patient diseases, the numbers and histories of scorers, the scoring method used (R&K or AASM), and the number of sleep stages.

Cohen's κ is a measure of the extent of agreement between two scorers[12] and is widely used to evaluate interrater reliability when the outcome is based on a categorical scale. Cohen's κ is considered more robust than a simple percentage because it considers the possibility that agreement occurred by chance. Cohen's κ can be simply calculated from a data matrix. The matrix and the formula for estimating κ when 2 independent scorers classify sleep stages into 5 levels (W, N1, N2, N3, and R) is shown in **Figure 1**.[12,13] According to Landis and Koch,[14] a $\kappa > 0.80$ represents near-perfect agreement (beyond chance); in comparison, a κ of 0.61–0.80 represents substantial agreement, 0.41–0.60 represents moderate agreement, 0.21–0.40 represents fair agreement, and 0.00–0.20 represents slight agreement.

Studies that reported information sufficient for κ calculations, and the standard errors (or estimates thereof) between 2 scorers, were included. Reports that used Fleiss's κ, which estimates the extent of agreement among 3 or more raters, or that gave only percentage agreements or mean Cohen's κ values obtained after the evaluation of different patients, were excluded.

For the analysis of the overall sleep stage, all available studies that provided essential data for performing a meta-analysis were included regardless of the number of sleep stages. For the analysis of each sleep stage, the staging criteria of the currently used AASM guideline was followed.[4] Studies based on the R&K system[3] that combined sleep stages S3 and S4 into 1 stage (slow-wave sleep, stage N3 sleep) were included.

Statistical analysis

We used the Dersimonian-Laird random-effects model to analyze Cohen's κ values. Heterogeneity was calculated employing the Cochran Q statistic and the I^2 test. The latter explores variation among studies; if $I^2 > 50\%$ and $P < .05$, then significant heterogeneity is in play.

We used a funnel plot to visually explore publication bias, followed by the Arcsine Thompson test. Publication bias is suspected when a funnel plot is asymmetrical. All analyses were performed with the aid of R software (version 3.1.3, R Foundation for Statistical Computing, Vienna, Austria).

RESULTS

Search results and characteristics

A flow chart of the study selection process is shown in **Figure 2**. A total of 101 relevant articles were retrieved. After screening of the titles and abstracts, 60 were subjected to a full-text review and 49 were excluded because they lacked the required data, contained only limited information without standard errors, or lacked a data

Figure 1—Data matrix and formula for calculating the Cohen's κ.

A

	W	N1	N2	N3	R
W	S_{11}	S_{12}	S_{13}	S_{14}	S_{15}
N1	S_{21}	S_{22}	S_{23}	S_{24}	S_{25}
N2	S_{31}	S_{32}	S_{33}	S_{34}	S_{35}
N3	S_{41}	S_{42}	S_{43}	S_{44}	S_{45}
R	S_{51}	S_{52}	S_{53}	S_{54}	S_{55}

B

(a) $$N = \sum_{i=1}^{5}\sum_{j=1}^{5} S_{ij}$$

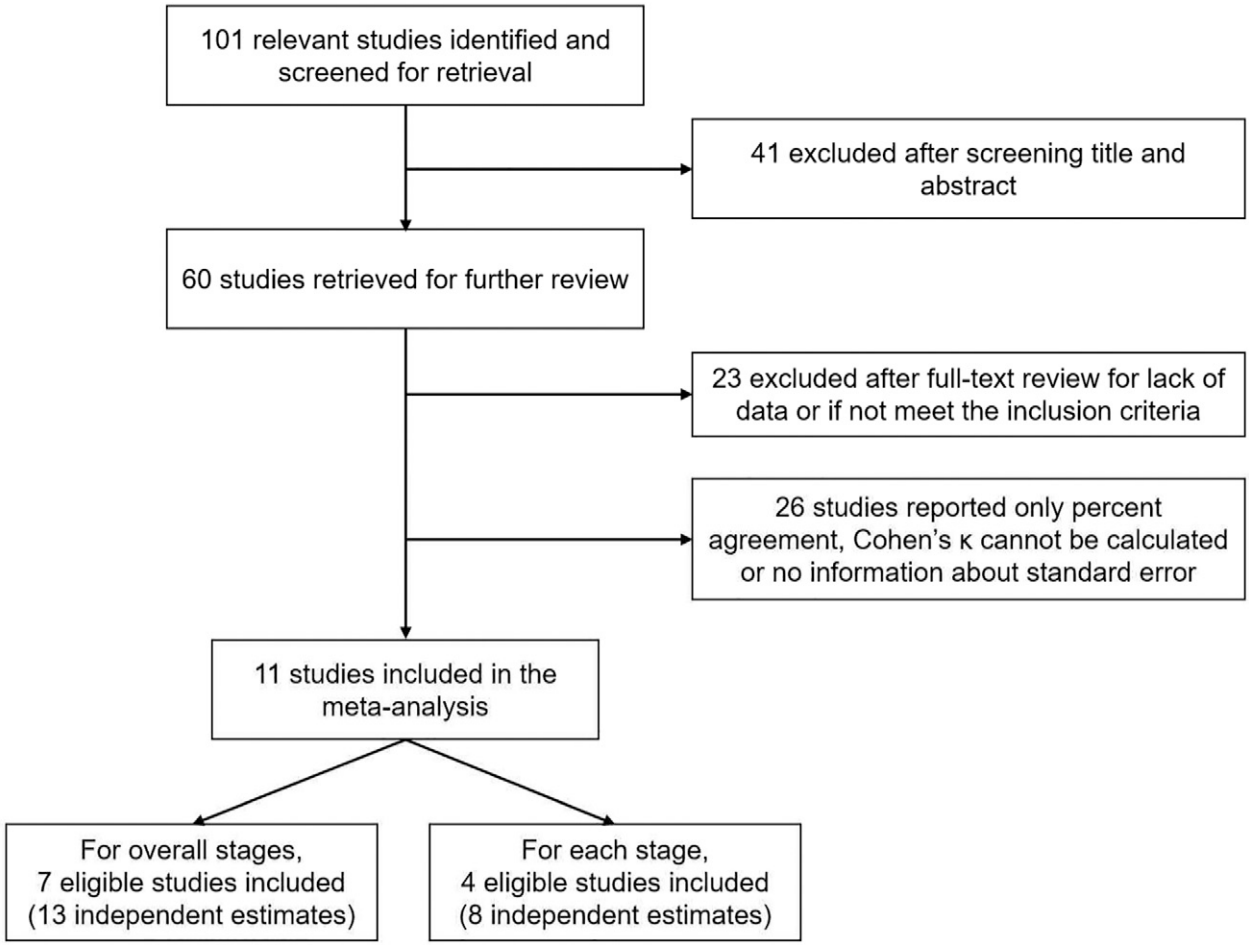
(b) $$\kappa = \frac{P_o - P_c}{1 - P_c}$$

(c) $$P_o = \frac{\sum_{i=1}^{5} S_{ii}}{N}$$

(d) $$P_c = \frac{\sum_{i=1}^{5}(\sum_{j=1}^{5} S_{ij} \sum_{j=1}^{5} S_{ji})}{N^2}$$

(A) The data matrix derived when sleep scoring sought to identify 5 sleep stage categories (W, N1, N2, N3, and R). *Sij* is the number of epochs. **(B)** The formula used to calculate the κ coefficient. **(a)** *N* is the total number of epochs scored. **(b)** *Po* is the observed agreement and *Pc* is the expected agreement. **(c, d)** *Po* and *Pc* are derived using these formulas.

Figure 2—Flow diagram of study selection.



matrix (κ could not be calculated). Seven studies were finally included for an analysis of the overall stages[5,6,8,15–18]. They contained 13 independent estimates. Four studies featured 8 independent estimates of the reliability of specific sleep stage scoring.

Overall interrater reliability

Thirteen independent datasets were analyzed in terms of the interscorer reliability of overall sleep staging (**Table 1**). If there are two or more data included in a paper under different conditions (eg, comparison among another scorers or subgroups of different patient characteristics), they are indicated as (a), (b), and (c). The Sleep Heart Health Study was a multicenter, longitudinal study conducted to relate sleep-disordered breathing to cardiovascular outcomes. Participants were recruited regardless of their sleep apnea status. Participants who met the inclusion criteria (aged ≥ 40 years, no history of sleep apnea treatment, no tracheostomy, and no current home oxygen therapy) were recruited from the parent cohort derived from 9 epidemiological studies. Significant among-study heterogeneity was evident ($I^2 = 99.8\%$; $P < .001$ according to the Q test). We used a random-effects model to derive overall pooled estimates. The overall Cohen's κ was 0.76 (95% confidence interval [CI], 0.71–0.81), indicating substantial reliability according to Landis and Koch (**Figure 3**).[14]

Subgroup analysis and publication bias

Because the evaluation rules that were employed differed, we predicted that there would be heterogeneity. And because few data were derived using AASM standards, subgroup analysis was impossible. When we analyzed the data by the number of sleep stages scored, significant differences were found between groups scored using 4–5 and 6–7 sleep stages. The estimates were 0.73 (95% CI, 0.67–0.79) for the former group and 0.83 (95% CI, 0.81–0.85) for the latter group (**Figure 4**). The difference was significant ($Q = 10.56$; $P < .05$). However, residual heterogeneity remained significant ($P < .001$), suggesting that other factors also influenced interrater reliability. We explored whether patient comorbidities, scorer histories, and the number of EEG derivations might be in play but found no significant between-group difference.

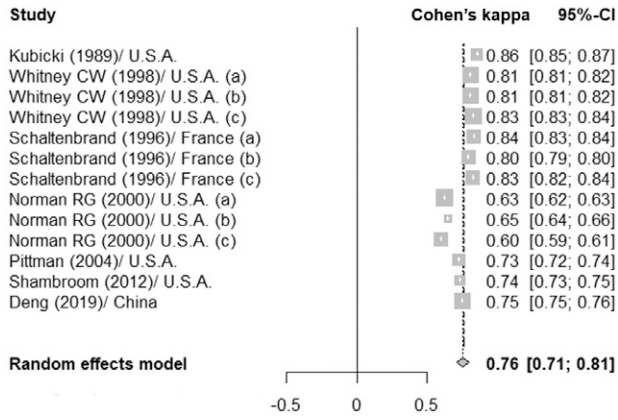
A funnel plot and the Arcsine Thompson test were used to evaluate publication bias. The P value of the Arcsine Thompson test was >.05, and no clear asymmetry was evident on a

Table 1—Studies for analysis of overall interrater reliability.

Study/Location	Rule/Number of Stages	EEGs	Scorers/Scorers' Characteristics	PSGs	Epochs	Patient Characteristics	Percentage Agreement (%)	Cohen's κ	SE
Kubicki et al, 1989[15]/U.S.A.	R&K/7	1	2/very experienced	10	13,850	Healthy	91.3	0.8584	0.0039
Whitney et al, 1998[6]/U.S.A.			3/trained and certified according to rigorous methods (Scorer ID 912, 914, 915)						
(a)	R&K/5	2	scorer 912 vs. scorer 914	30	29,507	NR, SHHS	86.8	0.8108	0.0028
(b)	R&K/5	2	scorer 912 vs. scorer 915	30	29,507	NR, SHHS	86.7	0.8107	0.0028
(c)	R&K/5	2	scorer 914 vs. scorer 915	30	29,507	NR, SHHS	88.1	0.8308	0.0027
Schaltenbrand et al, 1996[8]/ France			2/NR						
(a)	R&K/6	2		20	21,138	Healthy	88.2	0.8364	0.0031
(b)	R&K/6	2		20	20,080	Depressed	85.3	0.7977	0.0034
(c)	R&K/6	2		20	20,731	Insomnia	88.9	0.8297	0.0033
Norman et al, 2000[5]/U.S.A.			5/average experience 13.4 y, range 7–24 y						
(a)	R&K/5	2		62	53,735	Healthy, OSA	73.0	0.6263	0.0026
(b)	R&K/5	2		48	11,762	Healthy	76.0	0.6502	0.0057
(c)	R&K/5	2		14	33,628	SDB	71.0	0.6010	0.0034
Pittman et al, 2004[16]/U.S.A.	R&K/5	4	2/RPSGT	31	26,876	SDB	82.1	0.7299	0.0035
Shambroom et al, 2012[17]/U.S.A.	R&K/4	6	2/trained in a research setting	26	19,556	Healthy	83.2	0.7370	0.0042
Deng et al, 2019[18]/China	AASM/5	6	7/at least 2 y of AASM scoring	40	37,642	Healthy, OSA	82.1	0.7537	0.0027
Random-effects model (95% CI)								0.76 (0.71–0.81)	
Heterogeneity (I^2, %)								99.8	
P Value								< .001	

AASM = American Academy of Sleep Medicine, CI = confidence interval, EEG = electroencephalogram, NR = not reported, OSA = obstructive sleep apnea, PSG = polysomnography, R&K = Rechtschaffen and Kales, RPSGT = registered polysomnographic technologist, SDB = sleep-disordered breathing, SE = standard error, SHHS = Sleep Heart Health Study.

Figure 3—Forest plot for overall interrater reliability.



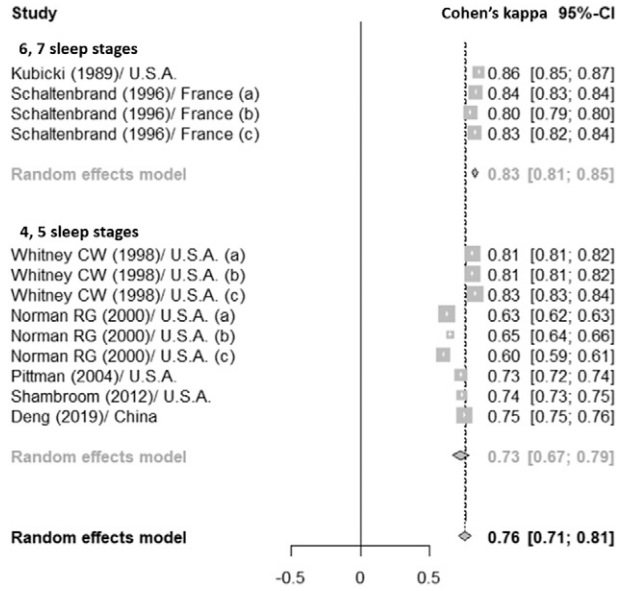
CI = confidence interval.

self-reported evaluation of the funnel plot, implying the absence of publication bias (**Figure 5**).

Interrater reliabilities by stage

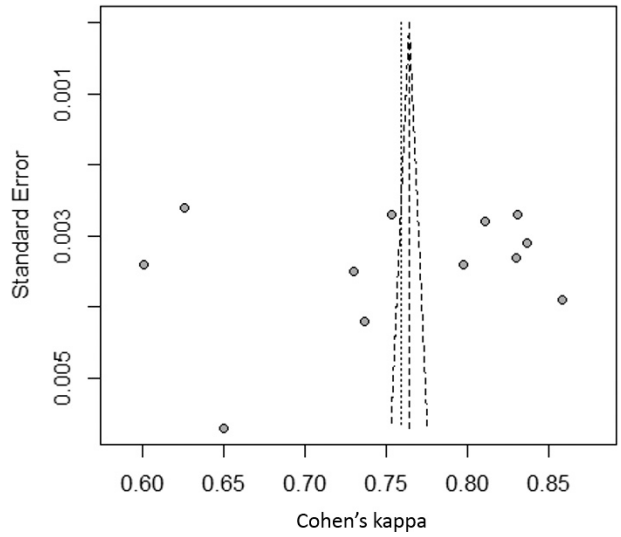
Four papers provided data allowing us to analyze agreement by stage (**Table 2**). We evaluated 8 datasets from these studies. The κ coefficients for each sleep stage (derived using random-effects models) were 0.70 (95% CI, 0.63–0.77) for stage W, 0.24 (95% CI, 0.15–0.33) for stage N1, 0.57 (95% CI, 0.54–0.60) for stage N2, 0.57 (95% CI, 0.42–0.71) for stage N3, and 0.69 (95% CI, 0.58–0.81) for stage R (**Figure 6**). The interrater reliabilities for stages W and R were substantial, those for stages N2 and N3 were moderate, and that for stage N1 was only fair. Significant among-study heterogeneities were evident for all 5 stages (I^2 = 87.2%–99%; P < .001 according to the Q test). We drew a funnel plot and performed the Arcsine Thompson test and found no obvious publication bias for any stage (all P > .05).

Figure 4—Forest plot for interrater reliability of different sleep stages.



CI = confidence interval.

Figure 5—Funnel plot for overall interrater reliability.



DISCUSSION

We collected data on the reliability of manual PSG scorers published since 1968 (when the R&K manual was published). A whole-night record usually contains 8 hours of data, which, when split into 30-second epochs yields 960 epochs that require 2–4 hours for manual scoring, which is both time-consuming and error-prone. To overcome this problem, the first automatic sleep staging system was reported by Martin et al[19] in 1972, and many other systems have since been developed and evaluated.[10,11,20,21] Currently, algorithms using artificial intelligence and machine learning are being developed. Despite certain automated systems that are highly accurate and consistent, manual scoring is still considered to be the gold standard. One of the reasons why the latest equipment still cannot be accepted is the low degree of reliability.[22] However, there is no clinical standard for the level of reliability for automated sleep staging. Thus, our analysis of interrater agreement among manual scorers is important; we derived generalizable outcomes that can be compared to those of the automatic systems.

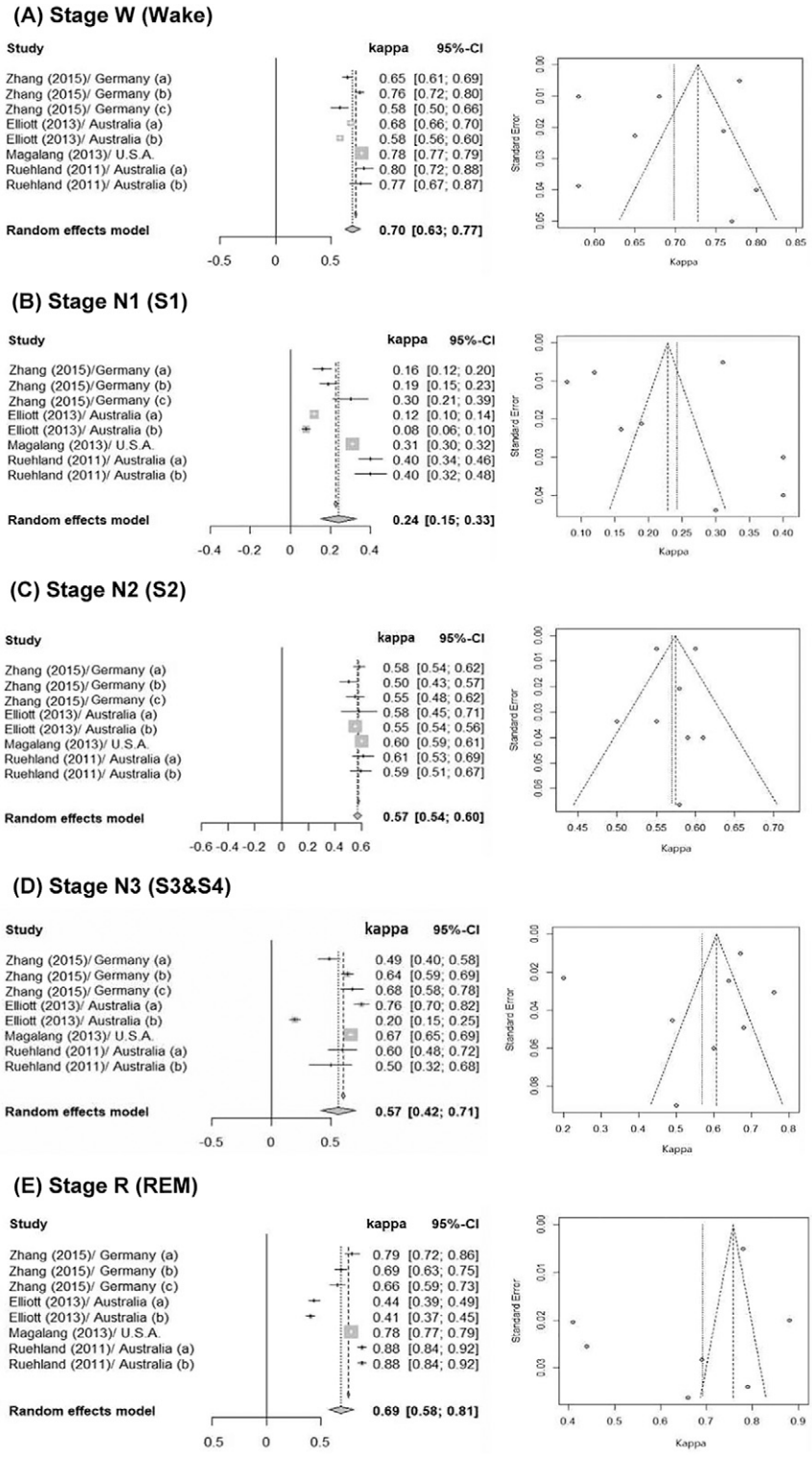
The R&K manual,[3] which has been accepted for decades, is only applicable to adults. Other recommendations have been used to reflect differences in the sleep characteristics and structures of infants and children.[23,24] A section for children (≥ 2 months postterm) was proposed in the 2007 AASM scoring manual, but the upper age limit that the pediatric rule applies to is controversial.[4,25] Because the scoring rules for infants and children are distinct from those for adults and have been applied separately, we decided to analyze studies including only adults.

Table 2—Studies for analysis of interrater reliabilities by stage.

Study/Location	Rule	Scorers/ Scorers' Characteristics	PSGs	Epochs	Patient Characteristics	Agreement (κ, SE)				
						Stage W (Wake)	Stage N1 (S1)	Stage N2 (S2)	Stage N3 (S3&S4)	Stage R (REM)
Zhang et al, 2015[26]/ Germany		4/2 Chinese & 2 German, at least 5 y experience								
(a)	AASM		7	7,250	Healthy	0.65 (0.0227)	0.16 (0.0227)	0.58 (0.0208)	0.49 (0.0454)	0.79 (0.0340)
(b)	AASM		8	7,231	SAHS	0.76 (0.0212)	0.19 (0.0212)	0.50 (0.0336)	0.64 (0.0247)	0.69 (0.0283)
(c)	AASM		15	14,897	Narcolepsy	0.58 (0.0387)	0.30 (0.0439)	0.55 (0.0336)	0.68 (0.0491)	0.66 (0.0362)
Elliott et al, 2013[27]/ Australia		3/qualified sleep technologists								
(a)	R&K	technologists 1 vs. technologists 2	16	18,644	ICU patients	0.68 (0.0102)	0.12 (0.0077)	0.58 (0.0663)	0.76 (0.0306)	0.44 (0.0255)
(b)	R&K	technologists 2 vs. technologists 3	16	25,908	ICU patients	0.58 (0.0102)	0.08 (0.0102)	0.55 (0.0051)	0.20 (0.023)	0.41 (0.0204)
Magalang et al, 2013[28]/U.S.A.	AASM	9/SAGIC members, at least 5 y experience	15	12,712	AHI 0–20 events/h (n=5); 21–30 events/ h (n=5); >30 events/ h (n=5)	0.78 (0.0051)	0.31 (0.0051)	0.60 (0.0051)	0.67 (0.0102)	0.78 (0.0051)
Ruehland et al, 2011[29]/ Australia		3/2 with at least 10 y experience; 1 with 1 y experience								
(a)	AASM/ 3 EEG		10	NR	NR	0.80 (0.0400)	0.40 (0.0300)	0.61 (0.0400)	0.60 (0.0600)	0.88 (0.0200)
(b)	AASM/ 1 EEG		10	NR	NR	0.77 (0.0500)	0.40 (0.0400)	0.59 (0.0400)	0.50 (0.0900)	0.88 (0.0200)
Random-effects model (95% CI)						0.70 (0.63–0.77)	0.24 (0.15–0.33)	0.57 (0.54–0.60)	0.57 (0.42–0.71)	0.69 (0.58–0.81)
Heterogeneity (I^2, %)						98.1	99.0	87.2	98.2	98.7
P Value						< .001	< .001	< .001	< .001	< .001

AASM = American Academy of Sleep Medicine, AHI = apnea-hypopnea index, CI = confidence interval, EEG = electroencephalogram, ICU = intensive care unit, NR = not reported, PSG = polysomnography, R&K = Rechtschaffen and Kales, REM = rapid eye movement, SAGIC = Sleep Apnea Genetics International Consortium, SAHS = sleep apnea hypopnea syndrome, SE = standard error.

Figure 6—Forest plot and funnel plot for interrater reliabilities by stage.



CI = confidence interval, REM = rapid eye movement.

We derived Cohen's κ coefficients between pairs of independent evaluators who scored sleep stages epoch by epoch. This measure assesses categorical results (ie, sleep stages). The κ coefficient can be calculated from the data matrix as follows: κ = (observed agreement–expected agreement)/(1–expected agreement). If 3 or more raters are to be compared, Fleiss's κ is commonly used.[30] The interpretation is the same as that for Cohen's κ. This statistic is the most important and widely

accepted measure of interrater reliability, especially in medical literature.[31]

Most studies have reported percentage agreements among scorers. Alternatively, agreement can be quantitatively assessed by deriving the intraclass correlation coefficient (ICC) of sleep parameters, including the apnea-hypopnea index, rapid eye movement sleep latency, and time-in-stage. Several reports have explored consistency using ICC.[7,10,28] In this analysis, we focused on Cohen's κ; this is the preferred method of reliability estimation when using categorical variables.

The differences between the R&K and AASM systems are as follows. The R&K guidelines[3] identify sleep stages as wake, rapid eye movement sleep, and nonrapid eye movement sleep (stages 1, 2, 3, and 4); 1 paper included stage 0 sleep (awake after falling asleep but before final awakening).[15] The AASM guidelines (AASM scoring manual)[4] recognize 5 sleep stages. Wake and rapid eye movement sleep stages are named W and R, respectively, and the nonrapid eye movement sleep stages are N1, N2, and N3. Stage 3 and stage 4 of the R&K standard are considered to form a single stage and are incorporated into N3 (deep delta-wave sleep). When analyzing overall agreement, all studies were included (regardless of the number of sleep stages). Finally, studies that applied 4–7 stages were selected, as shown in the second column of **Table 1**. We performed subgroup analysis based on the number of sleep stages; a significant difference was evident between the 2 groups. Otherwise, a consistent criterion for the number of sleep stages was required to analyze interrater reliabilities by stage, because this study was devised to include both the R&K and the AASM scoring manual guidelines. We targeted studies that classified sleep into 5 stages as recommended by the AASM, which is the standard currently used.

Reliability estimates will be influenced by the study setting, scorer experience, and any underlying disease of a patient. The AASM scoring manual[4] recommends the use of a montage that includes 3 standard EEGs (frontal, central, and occipital). In contrast, the R&K guidelines[3] suggest the use of at least 1 central EEG channel. According to a previous report, overall agreement improved slightly (from Cohen's κ 0.65 to 0.67) when the number of EEGs rose from 1 to 3.[29] Subgroup analysis in this study was not possible given the small number of papers. However, the data included in our meta-analysis were obtained using different EEG montages, thus creating heterogeneity.

A scorer's credibility is influenced by experience and qualifications. The scorers were described as "trained" or "certified," or "with several years of experience." In some studies, the scorers were described as registered polysomnographic technologist–certified, but this was not the case in other studies. The registered polysomnographic technologist has been an international expert certification for those who clinically assess sleep disorders since the Board of Registered Polysomnographic Technologists was established in 1978. This certification is accepted as the global standard measure of competency for estimating the ability of a scorer above a certain level. The registered polysomnographic technologist is the only objective reference for proving the scorers' qualifications in the included studies. We could not perform subgroup analysis by this factor because the number of studies was limited and among-study diversity rendered objectification difficult. We strongly suspect that scorer skill markedly affects reliability.

Several studies have suggested that interrater agreement falls when a patient has a condition that disrupts sleep.[5,26,27] Scoring of patients with sleep-disordered breathing or narcolepsy was less reliable than scoring of healthy patients. In patients in an intensive care unit, the overall κ was 0.51–0.56, thus only moderate.[27] Because the current analysis targeted all participants regardless of their underlying conditions, the results will be heterogeneous.

Participant age was considered as a contributing factor to the heterogeneity. However, in this study, it was difficult to conduct a subgroup analysis considering age. There was a fundamental limitation in terms of performing this subgroup analysis because age was reported in various forms, such as a range or the mean, or it was not reported. Therefore, it was impossible to classify the studies according to age categories.

Danker-Hopfe, Anderer, et al[32] compared the AASM and R&K standards[3,4] and found that the former were better for all stages except stage N2 sleep. In our results, the interrater reliability for stage N2 sleep was lower than that for stages W and R sleep. Because many of the studies included in the analysis of each sleep stage followed the AASM scoring manual, a similar line of reasoning may have worked. Another factor that may have contributed to the relatively low reliability for stage N2 sleep is that this stage can be confused with stage N1 sleep when sleep stages are scored. After the K complex and sleep spindle—typical features of stage N2 sleep—are detected once, the stage is scored as stage N2 sleep even if the characteristic N2 features are no longer visible.

The interrater reliability for stage N3 was moderate. The definition of a slow wave, which plays a major role when classifying stage N3 sleep, is specified by the amplitude (>75 μV). When the EEG amplitude is being visually determined, scoring errors can be introduced by human factors (manual scoring), various EEG channel derivations, and different software programs. In addition, the slow-wave sleep amplitude decreases significantly with age after approximately age 40 years.[33,34] The wide range of ages could have affected the reliability of rating stage N3 sleep.

In several previous studies, stage N1 sleep scores tended to exhibit the poorest agreement.[7,32,35] One report hypothesized that the cause of this phenomenon was that the transition from stage W to stage N1 sleep was difficult to recognize, especially in patients exhibiting sparse α activity. We found that stage N1 sleep was the least reliably detected, consistent with previous results.

Our meta-analysis could not include several recent large-scale studies because of a lack of data.[11,32,36] Furthermore, subgroup analysis was not possible because of the small number of studies and the methodological variations among them. Future studies should consider percentage agreement and ICC as alternative methods for determining the degree of agreement between scorers. The percentage agreement is the proportion of epochs scored as the same stage by the scorers. Percentage agreement could be calculated using the data matrix and formula in **Figure 1**. It has the same value as the "observed agreement," with *Po* converted into a percentage.

Figure 7—Data matrix and formula for calculating the ICC.

A

	Scorer 1	Scorer 2	⋯	Scorer k
Subject 1	x_{11}	x_{12}	⋯	x_{1k}
Subject 2	x_{21}	x_{22}	⋯	x_{2k}
⋯	⋯	⋯	⋯	⋯
Subject n	x_{n1}	x_{n2}	⋯	x_{nk}

B

(a) $x_{ij} = \mu + r_i + c_j + rc_{ij} + e_{ij}$

(b) $$ICC = \frac{variance\ of\ interest}{total\ variance} = \frac{variance\ of\ interest}{(variance\ of\ interest + error\ variance)}$$

(A) When the scorers (j = 1, 2, … , k) evaluate the PSG results of the patients (i = 1, 2, … , n), the data matrix can be filled in with target variables x_{ij}. Values of the target variables should fall along a continuous scale, such as the AHI and sleep stage (% or minutes). **(B)** The basic formula used to calculate the ICC in a 2-way random model. **(a)** Each measurement x_{ij} is assumed to be composed of a true component and a measurement error component. The model can be regarded as the sum of 5 terms: μ = mean of the patient's scores, r_i = deviation from the mean for patient i, c_j = bias of scorer j, rc_{ij} = interaction between patient deviation and scorer deviation, and e_{ij} = measurement error. **(b)** The ICC was calculated as a ratio of variance based on the results of an analysis of variance. The total variance is equal to the sum of the variance of interest (true score variance) and the error variance. The ICC is unitless and has a value between 0 and 1; an estimate of 1 indicates perfect reliability and 0 indicates no reliability. AHI = apnea-hypopnea index, ICC = intraclass correlation coefficient, PSG = polysomnography.

The ICC can be used to assess reliability between 2 or more raters when measurements yield continuous data (ie, apnea-hypopnea index [events/h], total sleep time [minutes], sleep efficiency [%]). Previous studies used units of time (minutes) or percentages to apply the ICC to assess the reliability of sleep stage scoring.[37–39] There are 6 types of ICC models, and the appropriate one is selected according to the study setting. A 2-way random model is suitable when both the patient effects and scorer effects are random. This model was applied by Norman et al.[5] We obtained the ICC using the data matrix and formula in **Figure 7**.[40,41]

The AASM interscorer reliability program implemented in 2010 to increase the reliability of scorers and to find the cause of disagreement involved thousands of participants and reported that the extent of sleep stage agreement was 82.6%.[9] This study identified that the scoring discrepancies were most common in epochs of transition from one stage to another and suggested the use of detailed brain wave definitions and other criteria for stage classification. Silber et al[42] considered that visual scoring is necessarily imperfect. However, it is possible to improve accuracy, and this method is the way forward.

CONCLUSIONS

PSG optimally evaluates sleep structure and aids the diagnosis of various sleep disorders. Our meta-analysis revealed that the overall extent of agreement among manual scorers was substantial ($\kappa = 0.76$). When we compared each sleep stage, we found that stages W and R sleep exhibited a substantial level of agreement; stages N2 and N3 sleep were moderate; stage N1 sleep was fair. Significant heterogeneity was apparent among the κ estimates, and there was a statistically significant difference between subgroups using a different number of sleep stages.

These results can serve as baseline data with which to judge whether the guidelines to be updated improve the interrater reliability of manual scoring. In addition, the nonrapid eye movement sleep stages were associated with a relatively low level of reliability, which should be considered when revising or supplementing the scoring rules. Many efforts are being made to improve agreement among scorers; analyses of the disagreements are helpful in this context.

ABBREVIATIONS

AASM, American Academy of Sleep Medicine
CI, confidence interval
EEG, electroencephalogram
ICC, intraclass correlation coefficient
PSG, polysomnography
R&K, Rechtschaffen and Kales

REFERENCES

1. Javaheri S, Redline S. Sleep, slow-wave sleep, and blood pressure. *Curr Hypertens Rep.* 2012;14(5):442–448.
2. Pillai JA, Leverenz JB. Sleep and neurodegeneration: a critical appraisal. *Chest.* 2017;151(6):1375–1386.
3. Kales A, Rechtschaffen A. *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages in Human Subjects.* Washington, DC: U.S. Government Printing Office; 1968.
4. Iber C, Ancoli-Israel S, Chesson AL Jr, Quan SF; for the American Academy of Sleep Medicine. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications.* 1st ed. Westchester, IL: American Academy of Sleep Medicine; 2007.
5. Norman RG, Pal I, Stewart C, Walsleben JA, Rapoport DM. Interobserver agreement among sleep scorers from different centers in a large dataset. *Sleep.* 2000;23(7):901–908.
6. Whitney CW, Gottlieb DJ, Redline S, et al. Reliability of scoring respiratory disturbance indices and sleep staging. *Sleep.* 1998;21(7):749–757.
7. Danker-Hopfe H, Kunz D, Gruber G, et al. Interrater reliability between scorers from eight European sleep laboratories in subjects with different sleep disorders. *J Sleep Res.* 2004;13(1):63–69.
8. Schaltenbrand N, Lengelle R, Toussaint M, et al. Sleep stage scoring using the neural network model: comparison between visual and automatic analysis in normal subjects and patients. *Sleep.* 1996;19(1):26–35.
9. Rosenberg RS, Van Hout S. The American Academy of Sleep Medicine inter-scorer reliability program: sleep stage scoring. *J Clin Sleep Med.* 2013;9(1):81–87.
10. Stepnowsky C, Levendowski D, Popovic D, Ayappa I, Rapoport DM. Scoring accuracy of automated sleep staging from a bipolar electrooocular recording compared to manual scoring by multiple raters. *Sleep Med.* 2013;14(11):1199–1207.

11. Younes M, Raneri J, Hanly P. Staging sleep in polysomnograms: analysis of inter-scorer variability. *J Clin Sleep Med.* 2016;12(6):885–894.
12. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas.* 1960;20(1):37–46.
13. Fraiwan L, Lweesy K, Khasawneh N, Wenz H, Dickhaus H. Automated sleep stage identification system based on time-frequency analysis of a single EEG channel and random forest classifier. *Comput Methods Programs Biomed.* 2012; 108(1):10–19.
14. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33(1):159–174.
15. Kubicki S, Höller L, Berg I, Pastelak-Price C, Dorow R. Sleep EEG evaluation: a comparison of results obtained by visual scoring and automatic analysis with the Oxford sleep stager. *Sleep.* 1989;12(2):140–149.
16. Pittman SD, MacDonald MM, Fogel RB, et al. Assessment of automated scoring of polysomnographic recordings in a population with suspected sleep-disordered breathing. *Sleep.* 2004;27(7):1394–1403.
17. Shambroom JR, Fábregas SE, Johnstone J. Validation of an automated wireless system to monitor sleep in healthy adults. *J Sleep Res.* 2012;21(2):221–230.
18. Deng S, Zhang X, Zhang Y, et al. Interrater agreement between American and Chinese sleep centers according to the 2014 AASM standard. *Sleep and Breathing.* 2019;23(2):719–728.
19. Martin WB, Johnson LC, Viglione SS, Naitoh P, Joseph RD, Moses JD. Pattern recognition of EEG-EOG as a technique for all-night sleep stage scoring. *Electroencephalogr Clin Neurophysiol.* 1972;32(4):417–427.
20. Levendowski DJ, Ferini-Strambi L, Gamaldo C, Cetel M, Rosenberg R, Westbrook PR. The accuracy, night-to-night variability, and stability of frontopolar sleep electroencephalography biomarkers. *J Clin Sleep Med.* 2017;13(6):791–803.
21. Jensen PS, Sorensen HB, Leonthin HL, Jennum P. Automatic sleep scoring in normals and in individuals with neurodegenerative disorders according to new international sleep scoring criteria. *J Clin Neurophysiol.* 2010;27(4):296–302.
22. Fiorillo L, Puiatti A, Papandrea M, et al. Automated sleep scoring: a review of the latest approaches. *Sleep Med Rev.* 2019;48:101204.
23. Scholle S, Schäfer T. Atlas of states of sleep and wakefulness in infants and children. *Somnologie (Berl).* 1999;3(4):163–241.
24. Anders TF, Emde RN, Parmelee AH. *A Manual of Standardized Terminology, Techniques and Criteria for Scoring of States of Sleep and Wakefulness in Newborn Infants.* Los Angeles, CA: UCLA Brain Information Service/BRI Publications Office, NINDS Neurological Information Network; 1971.
25. Grigg-Damberger M, Gozal D, Marcus CL, et al. The visual scoring of sleep and arousal in infants and children. *J Clin Sleep Med.* 2007;3(2):201–240.
26. Zhang X, Dong X, Kantelhardt JW, et al. Process and outcome for international reliability in sleep scoring. *Sleep Breath.* 2015;19(1):191–195.
27. Elliott R, McKinley S, Cistulli P, Fien M. Characterisation of sleep in intensive care using 24-hour polysomnography: an observational study. *Crit Care.* 2013;17(2): R46.
28. Magalang UJ, Chen N-H, Cistulli PA, et al; SAGIC Investigators. Agreement in the scoring of respiratory events and sleep among international sleep centers. *Sleep.* 2013;36(4):591–596.
29. Ruehland WR, O'Donoghue FJ, Pierce RJ, et al. The 2007 AASM recommendations for EEG electrode placement in polysomnography: impact on sleep and cortical arousal scoring. *Sleep.* 2011;34(1):73–81.
30. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull.* 1971;76(5):378–382.
31. Chmura Kraemer H, Periyakoil VS, Noda A. Kappa coefficients in medical research. *Stat Med.* 2002;21(14):2109–2129.
32. Danker-Hopfe H, Anderer P, Zeitlhofer J, et al. Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. *J Sleep Res.* 2009;18(1):74–84.
33. Smith JR, Karacan I, Yang M. Ontogeny of delta activity during human sleep. *Electroencephalogr Clin Neurophysiol.* 1977;43(2):229–237.
34. Tan X, Campbell IG, Feinberg I. Internight reliability and benchmark values for computer analyses of non-rapid eye movement (NREM) and REM EEG in normal young adult and elderly subjects. *Clin Neurophysiol.* 2001;112(8):1540–1552.
35. Anderer P, Gruber G, Parapatics S, et al. 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.* 2005;51(3): 115–133.
36. Anderer P, Moreau A, Woertz M, et al. 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 × 7. *Neuropsychobiology.* 2010; 62(4):250–264.
37. Malhotra A, Younes M, Kuna ST, et al. Performance of an automated polysomnography scoring system versus computer-assisted manual scoring. *Sleep.* 2013;36(4):573–582.
38. Muzet A, Werner S, Fuchs G, et al. Assessing sleep architecture and continuity measures through the analysis of heart rate and wrist movement recordings in healthy subjects: comparison with results based on polysomnography. *Sleep Med.* 2016;21:47–56.
39. Punjabi NM, Shifa N, Dorffner G, Patil S, Pien G, Aurora RN. Computer-assisted automated scoring of polysomnograms using the Somnolyzer system. *Sleep.* 2015;38(10):1555–1566.
40. Liljequist D, Elfving B, Skavberg Roaldsen K. Intraclass correlation—a discussion and demonstration of basic features. *PLoS One.* 2019;14(7):e0219854.
41. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull.* 1979;86(2):420–428.
42. Silber MH, Ancoli-Israel S, Bonnet MH, et al. The visual scoring of sleep in adults. *J Clin Sleep Med.* 2007;3(2):121–131.

SUBMISSION & CORRESPONDENCE INFORMATION

Submitted for publication March 31, 2021
Submitted in final revised form July 2, 2021
Accepted for publication July 2, 2021
Address correspondence to: Ji Ho Choi, MD, PhD, 14584, 170, Jomaru-ro, Bucheon-si, Gyeonggi-do, Republic of Korea; Tel: +82-32-621-5054; Email: handsomemd@hanmail.net

DISCLOSURE STATEMENT

All authors have seen and approved the manuscript. Work for this study was performed in the Department of Otorhinolaryngology—Head and Neck Surgery, College of Medicine, Soonchunhyang University, Bucheon Hospital, Bucheon, Korea. This study was funded by the Soonchunhyang University Research Fund. The authors report no conflicts of interest.