Discrepancy in Polysomnography Scoring for a Patient with Obstructive Sleep Apnea Hypopnea Syndrome

MASAAKI SUZUKI, HANAKO SAIGUSA, SHINTARO CHIBA, TOMOKO YAGI, KANA SHIBASAKI, MINEMOKO HAYASHI, MICHIKO SUZUKI, KIYOSHI MORIYAMA and KAZUOKI KODERA

Department of Otolaryngology, Teikyo University School of Medicine, Tokyo,
1 Ota Memorial Sleep Center, Ota General Hospital, Kanagawa,
2 Ikebukuro Sleep Clinic, Tokyo,
3 Department of Respiratory and Infectious Diseases, Tohoku University Graduate School of Medicine, Sendai, and
4 Sleep Respiratory Center, Komagamine Hospital, Tokyo, Japan


Overnight polysomnography (PSG) is indispensable for diagnosis of obstructive sleep apnea hypopnea syndrome. However, studies on interscorer agreement on PSG scoring between laboratories are few. The purpose of this study was to examine the reliability of interscorer agreement on PSG scoring among 16 sleep laboratories in Japan. We found a relatively moderate interscorer reliability of the index of oxygen desaturation and arousal during sleep, but a relatively low reliability of the index of transient reduction in and complete cessation of breathing (apnea hypopnea index). The median rate of interscorer coincidence of sleep staging was the lowest for slow wave (deep) sleep (23.5%), followed by those for Stage 1 (59.8%), Wake (73.2%) and Stage 2 (74.2%) in this order, and rapid eye movement was the most reliably identified stage (91.3%). The median rate of interscorer coincidence for all stages was 71.8%. The present study demonstrates that scorers tend to analyze PSG data according to a relatively empirical decision as opposed to a rule-dependent decision. Further detailed scoring manuals are required to decrease the interscorer discrepancy in PSG scoring.

Obstructive sleep apnea hypopnea syndrome (OSAHS) is a disorder characterized by repetitive partial or complete obstruction of the upper airway resulting in oxygen desaturation and sleep fragmentation (Kubin and Davis 2002). Recent investigations suggested that OSAHS is a risk factor and a predictor of a wide range of associated cardiovascular diseases, such as hypertension, angina, myocardial infarction, stroke and/ or congestive heart failure (Nieto et al. 2000; Shahar et
al. 2001; Tun et al. 2003). Polysomnography (PSG) is routinely indicated for evaluating sleep stages and sleep-related breathing events to detect sleep disorders. PSG has been accepted as the “gold standard” for evaluating individuals with sleep disorders and the severity of these disorders. According to Tachibana’s study, the estimated number of overnight PSG studies in Japan would be 18.3 PSG/year per 100,000 people, and 60.9 ± 28.2% of PSGs were performed for sleep disordered breathing (Tachibana et al. 2003). Since the Association for the Psychophysiological Study of Sleep established a standard system for visually scoring stages of sleep in 1967, this system has been known as the “Rechtschaffen and Kales standard” (Rechtschaffen and Kales 1968) and has been used worldwide. Originally the Rechtschaffen and Kales (1968) standard was meant to be a reference method, however, in practice it became both the major method of sleep analysis and the gold standard with which all other methods are compared (Himanen and Hasan 2000). Terashima’s study was presented in a Japanese Society of Sleep Research (JSSR) Conference in 1983, which investigated the discrepancy in sleep stage scoring using one PSG paper recording for a normal male (JSSR 2001). The results of multiple comparisons among four university research institutes and laboratories in Japan on the standard scoring system were inconsistent. The rate of coincidence for Stage 1 sleep was the lowest, with the second lowest rate of coincidence for slow-wave sleep stages (less than 70%). In 1993, Kim et al. (1993) reported sleep stage scoring using three PSG paper recordings for normal men in 10 laboratories in Japan, and demonstrated that the rates of coincidence for Stages 2 (85.2 – 90.1%) and Rapid eye movement (REM) (92.7 - 94%) were high, and that for Stage 1 was moderately high (65.9 – 74.1%), and those for Stages 3 and 4 were low (44.2 - 62.6%). Because the need to re-examine the definitions of sleep stages was emphasized by Terashima, the JSSR proposed supplementary definitions and amendments to the Rechtschaffen and Kales (1968) standard to increase both the intra- and interscorer agreements in sleep stage identification in 1995 (JSSR 2001).

Not only sleep stages but also sleep-related breathing events and arousals are important for detection and evaluation of the OSAHS. It is currently a standard in clinical practice and epidemiologic studies to assess the severity of OSAHS by combining the numbers of apneas, hypopneas and respiratory-effort-related arousals to obtain apnea hypopnea index (AHI). However, Redline (2000) analyzed 5046 PSG data on the basis of different AHI definitions available and found that median AHI varied from 6 to 29.3 /h. A patient with a given apnea hypopnea index determined in one laboratory should be comparable with a patient determined to have the same AHI in another laboratory. In Japan, studies on interscorer agreement on respiratory disturbance and arousal indices between laboratories are few. The purpose of this study was to examine the reliability of interscorer agreement on respiratory indices, oxygen desaturation index of more than 3% (ODI3), and arousal index (Ar-I), and epoch-by-epoch comparison of sleep staging among institutions. This study is a multicenter trial in Japan designed to investigate the discrepancy in digital PSG scoring for OSAHS.

**METHODS**

An attended overnight sleep study was carried out on an OSAHS patient, using digital PSG (Alice 4: Respironics, Marietta, GA, USA). The patient was a 38-year-old male without any complications and whose BMI was 25.9 kg/m². The study protocol was approved by the Institutional Ethics Committee of our institute, and informed consent was obtained from the patient. PSG data were analyzed manually by polysomnographic technologists from 16 sleep clinical institutes in Japan. PSG data analysis was performed by one scorer in each sleep laboratory. Seven of the 16 scorers were approved by the Board of Registered Polysomnographic Technologists (RPSGT) in the United States. All the scorers were informed that this was a reliability study when they analyzed the PSG data.

For digital PSG study, electroencephalography (EEG; C4/A1, C3/A2), electrooculography, submental electromyography (EMG) and electrocardiography using surface electrodes, measurement of air flow at the nose and mouth using a thermistor, measurement of respirato-
ry movements of the rib cage and abdomen by inductive
plethysmography, and percutaneous arterial oxygen satu-
rating using a finger pulse oximeter were simultaneously
carried out. Some software montage displays may be
slightly different from each other among the 16 institutes.
PSG data were scored manually on a high-resolution
monitor using 30-second epochs for staging and arousal
detection, and using 2- or 5-minute epochs for respiratory
data. Sleep data were scored without visualization of re-
spiratory channels, and respiratory data were scored
without visualization of EEG and EMG channels. The
predominant sleep stage was scored for each 30-second
epoch coded as Wake, Stage 1, Stage 2, Stage 3/4 (slow
wave sleep), or REM according to the criteria established
by Rechtschaffen and Kales (1968). Arousals were identi-
fied according to the criteria of the American Sleep
Disorders Association (ASDA) (1992). Potential apneas
were identified by most of the scorers as a nearly flat air-
flow (< 25% of baseline, where baseline amplitude is
identified during the nearest preceding period of regular
breathing with stable oxygen saturation) lasting 10 sec-
onds. Two scorers used < 20% of baseline for apnea
definition. Potential hypopneas were identified by most
of the scorers as an airflow or a thoracoabdominal excur-
sion of approximately < 70% of baseline associated with
either an oxygen desaturation of > 3% or an arousal for
at least 10 seconds. One scorer used < 74% of baseline
and oxygen desaturation of > 4% for hypopnea defini-
tion.

Interscorer reliabilities of the respiratory disturbance
indices (as measures of sleep-related breathing events),
the index of oxygen desaturation of more than 3%
(O DI3), and arousal index ( Ar-I ) were estimated using
coefficient of variation (CV; equal to the standard error
devided by the average). The durations of obstructive
breathing events were also compared among each scorer
using CV. The interscorer epoch-by-epoch comparison
of sleep staging was assessed. For each stage, the num-
ber of pairwise agreements and disagreements between
scorers are presented in tables, then the rates of coinci-
dence within each scorer pair was calculated. The rates
of coincidence of sleep stages between one and the other
15 technologists were calculated in each matching. The
method of analysis of the rate of coincidence of sleep
stages was according to the method of Terashima (Kim et
al. 1993). In brief, the rate of coincidence was defined as
the percent agreement on a particular stage among ep-
ocli in which at least one of the scorer pair indicated that
stage. For example, if the total numbers of epochs
scored as Stage 2 by scorer A and scorer B are 451 and
396, respectively, and the total number of epochs which
both scorers agreed on as Stage 2 is 288, the rate of coin-
cidence associated with Stage 2 is 288 × 2/ (451 + 396)
multiplied by 100, that equals 68.0%.

RESULTS

The apnea index ( AI ), hypopnea index ( HI )
and AHI ranged from 1.3 to 6.5 /h, 10.5 to 26.5
/h, and 13.5 to 29.5 /h, respectively, and CVs
were 47.2%, 31.2%, and 24.8%, respectively (Fig.
1). On the other hand, O DI3 and Ar-I ranged
from 9.2 to 18.3 /h, and 13.3 to 36.8 /h, respec-
tively, and CVs were 21.5% and 22.9%, respec-
tively, suggesting that O DI3 and Ar-I are more re-
liable than respiratory disturbance indices. The
respiratory disturbance indices and O DI3 obtained
by autoscoring (3.1 /h for AI, 5.2 /h for HI, 8.8 /h
for AHI, 9.2 /h for O DI3 ) were relatively lower
than those obtained by manual scoring, however,
the arousal index obtained by autoscoring (36.3
/h) was relatively higher than that by manual scor-
ing. CVs of AI, HI, AHI, O DI3, and Ar-I among
7 Registered Polysomnographic Technologist
(RPSGT) were 49.4, 31.9, 25.1, 26.3, and 31.2,
respectively, showing no significant difference
comparing with CVs among all scorers.

The durations of maximum apnea and hy-
popnea ranged from 36.5 to 105.5 sec and 47 to
207.5 sec, respectively, and CVs were as high as
33.9% and 46.5%, respectively (Fig. 2). On the
other hand, the durations of mean apnea and hy-
popnea ranged from 19.6 to 30.6 sec and 18.8 to
39.0 sec, respectively, and CVs were 12.8% and
20.3%, respectively, suggesting their higher reli-
ability than maximum duration. The respiratory
disturbance durations obtained by autoscoring (34
sec for maximum apnea duration, 55 sec for max-
imum hypopnea duration, 16.6 sec for mean ap-
nea duration, and 23.1 sec for mean hypopnea du-
ration) were relatively lower than those obtained
by manual scoring.

The numbers of epochs for each sleep stage
calculated by 16 scorers are shown in Fig. 3. The
numbers of epochs for Wake, Stage 1, Stage 2,
Stage 3, Stage 4, and REM, and movement time
(MT) ranged from 14 to 149 epochs, 162 to 532
epochs, 66 to 582 epochs, 0 to 97 epochs, 0 to 52 epochs, 163 to 206 epochs, and 0 to 14 epochs, respectively. The median numbers of epochs for Wake, Stage 1, Stage 2, Stage 3, Stage 4, REM and MT were 100 epochs, 244 epochs, 478 epochs, 18 epochs, 0 epochs, 181 epochs, and 0 epochs, respectively.

The rates of interscorer coincidence for each sleep stage and all sleep stages between one RPSGT and the other scorers on epoch-by-epoch
Fig. 3. Scatterplot of numbers of epochs for each sleep stage by 16 scorers. Thick bars and bold numbers indicate median values. Thin bars and italic numbers indicate autoscoring values. Other numbers indicate maximum and minimum values. Wake, wake stage; I, Stage 1; II, Stage 2; III, Stage 3; IV, Stage 4; REM, rapid eye movement; MT, movement time.

Fig. 4. Scatterplot of the interscorer rate of coincidence for each sleep stage and all stages between one and the other scorers on epoch-by-epoch comparison in each matching. Each plot indicates the rate of interscorer coincidence for each sleep stage and all sleep stages between one and the other scorers on epoch-by-epoch comparison. Thick bars and bold numbers indicate median values. Thin bars and italic numbers indicate autoscoring values. Other numbers indicate maximum and minimum values. SW, slow wave stage; All, all sleep stages.

Comparison in each matching are shown in Fig. 4. Seven of the 16 scorers were on the Board, and the total number of matchings was 84. Because median numbers of epochs for MT was 0, MT was not analyzed in this study. The rates of interscorer coincidence for Wake, Stage 1, Stage 2, slow wave (SW), and REM ranged from 17.2% to 92.7%, 34.6% to 81.9%, 16.2% to 89.2%, 0 to 76.5%, and 84 to 97.9%, respectively. The rank order of the median rates of interscorer coincidence was as
follows: REM (91.3%), Stage 2 (74.2%), Wake (73.2%), Stage 1 (59.8%), and SW (23.5%). The median rate of coincidence for REM had the highest reliability, however, that for SW sleep was less than 25%. SW and Stage 2 were the most interchangeably misidentified, followed by Stages 1 and 2. REM was rarely misidentified by another scorer as Wake, Stage 1, 2 or SW. The rates of interscorer coincidence for all sleep stages between one and the other scorers ranged from 36.7 to 87.4%, and the median, demonstrating the most representative rate of coincidence for sleep staging in the present study, was 71.8%. The median rate of coincidence between the scorers and autoscoring for REM showed moderate reliability (80.4%), although those for Wake (29.7%), Stage 1 (36.6%), Stage 2 (46.3%), and SW (15.7%) were not reliable. The median rates of coincidence between the scorers and autoscoring for all stages was as low as 51.2%.

**DISCUSSION**

**Discrepancy in arousals**

An EEG arousal is an abrupt shift in EEG frequency, which may include theta, alpha, and/or frequencies greater than 16 Hz but not spindles, according to the 11 EEG arousal scoring rules (ASDA scoring rules) (ASDA 1992). Since ASDA suggested arousal identification in 1992, arousals have been determined by most scorers according to these international standard criteria. However, CV of Ar-I in the present study was 22.9% ranging from 13.3 to 36.8/h, suggesting a moderate reliability. There are several explanations for this discrepancy in arousal scoring in this study. First, arousals that proceeded to the Wake Stage were evaluated differently by the scorers. Although arousal scoring is independent of the Rechtschaffen and Kales standard epoch scoring in ASDA scoring rules, no criteria for arousal termination are established. Second, delta wave bursts associated with the termination of apnea/hypopnea were not considered as arousals by many scorers, however, they were considered as arousals by some. According to ASDA scoring rules, delta waves are not scored as arousals. However, the use of respiratory tracings may be important in the scoring of arousals. Third, there was a disagreement in whether or not a leg EMG shift that precedes an EEG frequency shift should be scored as an arousal. Fourth, there was also disagreement in whether or not theta, alpha, or delta wave changes that precedes a shift in EEG frequency should be scored as arousals. According to three-second duration criteria, it may have to be considered as an arousal when such activities and a EEG frequency shift amount to 3 seconds. Identification and agreement on events of shorter duration were difficult to achieve.

**Discrepancy in sleep stages**

The present study showed that SW and Stage 2 were most interchangeably misidentified followed by Stages 1 and 2. The reason for this discrepancy in slow-wave stage scoring may be, first, voltage measurements for detecting delta waves were not performed by many scorers, thus other waves such as sweat artifacts were evaluated as delta. Second, an epoch in which delta waves amounted to 20 or 50% were identified as Stage 3 or 4, however, there was a discrepancy in whether only one EEG channel or two to four channels are used for this calculation. The reason for this interchangeable misidentification between Stages 1 and 2 may be, first, the discrepancy in whether or not the three-minute duration criteria was strictly observed. Second, an epoch in which more than 50% of low-voltage-mixed frequency accompanied with spindles at the end of the epoch was identified as Stage 1 by some, and identified as Stage 2 by others. Third, there was a discrepancy in whether an epoch with spindles but having mental EMG shifts was identified as Stage 1 or 2.

The stage-specific results obtained in the present study demonstrated that the reliability of stage scoring had not improved compared with that obtained by previous investigators (Kim et al. 1993; JSSR 2001). Supplementary definitions and amendments proposed in 1995 were established for PSG paper recordings (personal communications). Digital PSG is already widely used, therefore, new supplementary definitions and amendments for digital recordings is neces-
sary.

**Discrepancy in respiratory disturbance indices**

In the present study, CVs of AI (47.2%), HI (31.2%), and AHI (24.8%) were higher than we expected. The Sleep Heart Health Study (SHHS) investigated the reliability of PSG scoring measured using an unattended, in-home portable monitor (Whitney et al. 1998). This study was carried out by three scorers who were the staff of the Reading Center of SHHS, and trained and certified according to their Reading Center Manual of Operation (SHHS 1996). The intraclass correlation (ICC) of respiratory disturbance indices in the SHHS study was from 0.74 to 0.99, suggesting that the reliability was substantial to almost perfect, and much higher than that of arousal index with an ICC of 0.54. In contrast, the reliability of respiratory disturbance indices in this study was lower than that of Ar-I. Apnea, characterized by a cessation of airflow for 10 sec or more, seems to be easily recognized. However, in the present study, the CV of AI was 47.2% suggesting a low reliability. Not only between hypopneas and normal breathings, but also between apneas and hypopneas, interchangeable misidentification was noted in this study because of varied definitions of the degree of airflow or respiratory effort reduction for breathing event scoring. The definition of hypopnea, characterized by a reduction without cessation in air flow or effort, is not consistent. More variable definition features include the degree of airflow or respiratory effort reduction, inclusion of degree of oxygen desaturation, and inclusion of arousal from sleep.

The discrepancy in hypopnea index is due to the definition used and perhaps methods of detection that have been extensively explored. Tsai et al. (1999) compared different AHI definitions including a discernable reduction in respiratory effort combined with > 4% desaturation alone (AHI-A), 4% desaturation or arousal (AHI-B), or arousal alone (AHI-C). The results revealed that AHI-A and AHI-B showed a high correlation and agreement with one another, on the other hand, the correlation and agreement between AHI-C and AHI-A were poor, indicating that a definition of hypopnea requiring associated oxygen desaturation yields substantially different results compared with that based on associated arousal alone. In 1999, the American Academy of Sleep Medicine (AASM) Task Force recommended the definition of obstructive hypopnea as a clear decrease (> 50%) from baseline in the amplitude of a valid measure of breathing during sleep or a clear amplitude reduction that does not fulfill the above criterion, but associated with either an oxygen desaturation of > 3% or an arousal (AASM 1999). Two years later, the Clinical Practice Review Committee (CPRC 2001) reported in a position paper that hypopnea is defined as a 30% reduction in thoracoabdominal movement or airflow as compared with baseline, and with at least a 4% oxygen desaturation. Although this definition does not include all patients who may benefit from treatment, it may be useful for obtaining higher inter- and intrascorer reliabilities.

On the other hand, the maximal esophageal pressure (Pes) swing at the inspiratory peak during sleep is an absolute value and rarely influenced by the difference in scorer or method of detection (Suzuki et al. 2005). An increasing negativity of Pes directly reflects the respiratory effort during sleep. Not only respiratory disturbance indices, but also Pes was found to be significant in evaluating sleep-related breathing events. A more detailed scoring manual for apnea and hypopnea is necessary in Japan to increase intra- and interscorer reliabilities.

In conclusion, a standard PSG scoring among institutions in Japan is as yet not established. The reliability of PSG scoring is very important for multicenter trials with the scope of establishing new evidence that may contribute to sleep medicine. Although gold standard manuals may restrict the development of subsequent sleep research, a detailed scoring manual for breathing events, arousals, and sleep stages should be adopted by every institution in Japan to reduce the interscorer discrepancy in PSG scoring.

**Acknowledgments**

The authors want to express their deepest appreciation to Dr. Masayoshi Terashima, Nagoya
Family Court, for providing technical support for this research. The authors acknowledge other participating institutions and PSG scorers: Ms. Ayaka Sakamoto, Division of Sleep Medicine, Setagaya Sleep Center, Tokyo; Ms. Wakana Imamura, Ueno Sleep Clinic, Tokyo; Ms. Chiyoko Konishi, Snoring and Sleep Disordered Breathing Laboratory, Suetsugu Clinic, Nagoya; Mr. Katsunori Yamamoto, Sleep Disorders Center, Toyohashi Mates Clinic, Toyohashi; Mr. Koutaro Miyake, Department of Psychiatry, Shiga University of Medical Science, Otsu; Ms. Setsuko Yamashita, Kyotani Clinic, Osaka; Mr. Keisuke Yorozu, Aijinkai General Hospital, Osaka; Ms. Noriko Maruta, Osaka Sleep Fuji-Respirronics; Ms. Naoko Suekichi, Okayama Sleep Fuji-Respirronics; Ms. Mitsuko Kuchiishi, Haradoi Hospital, Fukuoka; Ms. Mayumi Nakanaga, Kego Hospital, Fukuoka and Ms. Ritsuko Miyagi, Saga Insurance Hospital, Saga. This work was supported by Research Grant No. 17591802 to M. S. from the Ministry of Education, Science and Culture, Japan.

References


Sleep Heart Health Study (SHHS) (1996) Sleep Heart Health Study Reading Center Manual of Operations. SHHS Coordinating Center, Seattle, WA.


