ELSEVIER

Sleep Medicine 3 (2002) 43–47

SLEEP MEDICINE

www.elsevier.com/locate/sleep

Original article

Scoring variability between polysomnography technologists in different sleep laboratories

Nancy A. Collop*

Division of Pulmonary/Critical Care Medicine, University of Mississippi Medical Center, Jackson, MS 39216, USA

Received 25 July 2000; received in revised form 7 February 2001; accepted 27 February 2001

Abstract

Objective: Examine the variability of polysomnography technologists from different sleep laboratories regarding scoring of polysomnograms.

Background: Polysomnography is the gold standard to diagnose obstructive sleep apnea–hypopnea syndrome. There are criteria to score sleep stages and respiratory events. We sought to determine how different technologists would score the same tests using their laboratory's criteria.

Methods: Eleven technologists in nine different sleep laboratories which used the Oxford Medilog SAC® system scored eleven sleep studies performed in the Medical University of South Carolina Sleep Disorders Laboratory utilizing their respective laboratory's scoring rules. All sleep studies were performed for evaluation of obstructive sleep apnea–hypopnea syndrome (OSAHS). The scored studies were returned and analyzed for variability.

Results: Significant variability was present in scoring of both sleep and respiratory events with more variability demonstrated in respiratory event scoring. In four of the studies, diagnoses based on apnea–hypopnea indices (AHI) varied from none to moderate OSAHS depending on which technologist scored the study and in one study the diagnosis varied from none to severe OSAHS.

Conclusions: Clinicians should be aware that there is tremendous variability among polysomnography technologists regarding the scoring of polysomnograms. These differences are likely due to different rules used to score events as well as differences in the technologist's interpretation of the rules. Published by Elsevier Science B.V.

Keywords: Polysomnography; Sleep staging; Variability; Scoring; Obstructive sleep apnea–hypopnea syndrome; Polysomnography technologist; Apnea–hypopnea index

1. Introduction

Obstructive sleep apnea–hypopnea syndrome (OSAHS) is a common syndrome afflicting millions. It has been estimated to affect up to 4% of the working male population of the United States and the prevalence is higher in specific populations such as those with obesity or habitual snoring. Sleep-disordered breathing defined by polysomnography alone, is even more prevalent, with 24% of males and 9% females having an apnea–hypopnea index >5 [1]. Overnight polysomnography (PSG) remains the reference test for the diagnosis of OSAHS. As with any reference test, it is important to test its reliability. Polysomnography itself is a reliable test in that the equipment is relatively standard in its ability to measure sleep, respiration, heart rate and other variables. The scoring of polysomnography however, is at high risk for poor reliability because some of its variables require subjective interpretation and many of its definitions are non-standardized.

Polysomnography measures a variety of physiologic variables. Sleep is broken down into stages which are determined by fairly rigorous criteria developed by Rechtschaffen and Kales [2] utilizing electroencephalography, electromyography of the chin muscle and electrooculography. Respiratory parameters are measured utilizing airflow, chest and abdominal effort, and oxygen saturation. These variables are used to determine the presence of apneas and hypopneas, which are further subdivided into central, mixed, or obstructive. The definition of an apnea is defined as absence of airflow for ≥10 s. Hypopneas have variable definitions between laboratories [3]. An apnea–hypopnea index (AHI) is derived from the total of apneas + hypopneas divided by the total sleep time measured in hours. This index is the most important variable generated from a PSG for the diagnosis of OSAHS and usually the treatment, and the reimbursement for treatment,

* Tel.: +1-601-984-5650; fax: +1-601-984-5658.
E-mail address: ncollop@aol.com (N.A. Collop).

1389-9457/02/$ - see front matter. Published by Elsevier Science B.V.
PII: S1389-9457(01)00115-0

is dependent on this index. Another important variable measured in PSG is the degree of oxygen desaturation and this is often incorporated into the definition of obstructive events.

A few studies have evaluated interrater reliability for PSG [4,5]. Only one has investigated differences between laboratories and it specifically examined differences related to scoring electroencephalographic arousals [5]. The purpose of this study is to measure the differences between the scoring of sleep studies between technologists by sending sample PSGs to different sleep laboratories and analyzing how their individual technologists score utilizing their sleep laboratory's scoring rules.

2. Methods

Oxford Instruments provided the names of fifteen sleep laboratories that utilized the Oxford Medilog SAC System® which was compatible with the system used in Medical University of South Carolina (MUSC) Sleep Disorders Laboratory. Eleven polysomnograms (two of which were the first half of a split-night study) performed at the MUSC Sleep Disorders laboratory were copied onto optical discs and sent to the participating laboratories. All of the PSGs were done for evaluation of OSAHS. All the studies were done using standard polysomnography equipment and montages including: electroencephalography monitoring central and occipital leads; chin electromyography; right and left electrooculography; EKG (lead II); intercostal electromyography; oronasal thermistry; chest and abdominal pneumography; oxygen saturation; anterior tibialis electromyography; and pulse oximetiy. The technologists in the study laboratories, who performed most of their laboratory's scoring, scored each of the eleven studies and returned them for analysis. The laboratories were instructed to use their own scoring rules. We did not obtain information about the individual laboratory's scoring rules. Patient names and any other identifying data as well as any prior scoring done by our laboratory were removed from the polysomnogram. Each technician received $200.00 upon return of scored studies.

Of the original fifteen technologists contacted, eleven completed the scoring. One laboratory had two different scorers for a total of nine different laboratories sampled. Four of the technologists were registered (RPsgT). Technologist experience working in a sleep laboratory ranged from 3–11 years (data on nine technologists available, average 7.7 years). The technologists came from nine different sleep laboratories, seven in the US and two in Canada. Two of the laboratories were University based, seven were in private hospitals.

2.1. Statistical analysis

The studies were analyzed in a variety of ways:

1. We compared how many studies would have given the patient a different 'diagnosis' based on the following: no OSAHS: AHI ≤ 5; mild OSAHS: AHI >5, ≤ 15; moderate OSAHS: AHI > 15, ≤ 40; severe OSAHS: AHI > 40.
2. Apnea–hypopnea index, sleep efficiency, sleep latency and total sleep time were analyzed for variability utilizing the random effects analysis of variance statistical method.
3. Apnea–hypopnea index; total apneas + hypopneas, and sleep efficiency, were also analyzed to determine interrater agreement using the kappa statistic (κ), as described by Cohen and modified by Fleiss for multiple ratings per study [6,7]. There were four categories for each test: The AHI categories were none, mild, moderate and severe; total apneas and hypopneas categories were <30 as none, 30–100 as mild, 101–275 as moderate and >275 as severe; and sleep efficiency categories were 90–100% as excellent, 80–89% as adequate, 70–79% as poor and $<70\%$ as insufficient. κ is equal to 1.00 when there is complete agreement among all the raters, and 0.00 when the agreement is only due to chance. Fleiss also defined the level of agreement according to the following: $\kappa \geq 0.60 =$ good; $0.60 < \kappa \geq 0.40 =$ moderate; $\kappa < 0.40 =$ poor.
4. A coefficient of variation was also calculated for total sleep time and apnea–hypopnea index by dividing the mean by the standard error of the mean.

The study was supported by a grant from Oxford Instruments and approved by the MUSC Institutional Review Board.

3. Results

Figs. 1–3 shows the variability between technologists for total sleep time, sleep efficiency, and sleep latency. The figures show the lowest and highest value for each parameter, and the mean with standard deviation.

Respiratory scoring is similarly depicted in Figs. 4 and 5. A specific analysis of the variation between a respiratory

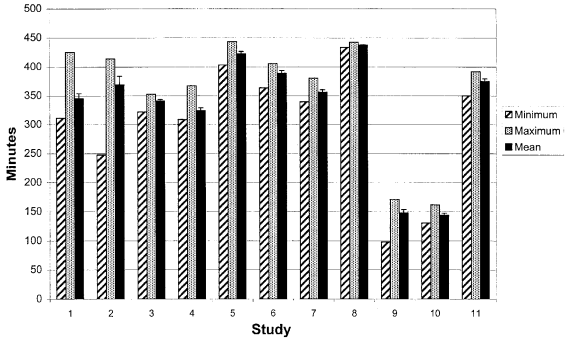

Fig. 1. Total sleep time. This figure depicts the ranges of total sleep time by study. The cross hatched bars represent the minimum any technician scored total sleep time; the dotted bars represent the maximum total sleep time any of the technicians scored total sleep time and; the black bars represent the mean of all scorers with Y-error bars representing the standard deviation.

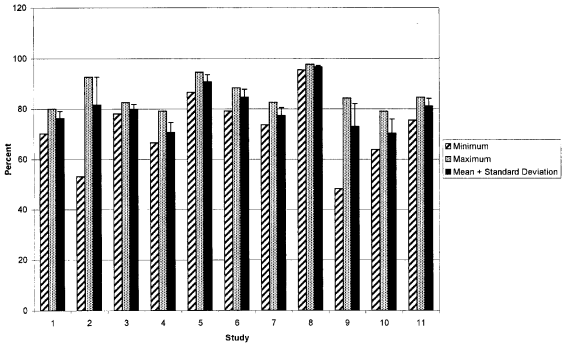

Fig. 2. Sleep efficiency. This figure depicts the ranges of sleep efficiency by study. The cross hatched bars represent the minimum any technician scored sleep efficiency; the dotted bars represent the maximum sleep efficiency any of the technicians scored and; the black bars represent the mean of all scorers with Y-error bars representing the standard deviation.

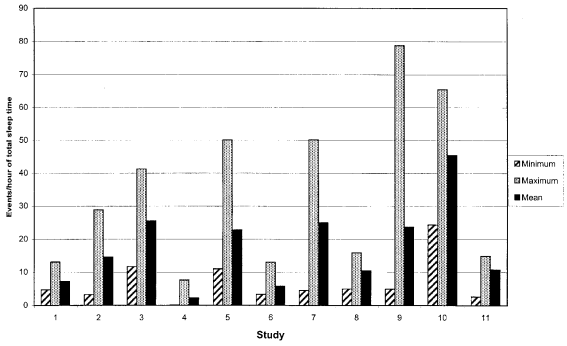

Fig. 4. Apnea–hypopnea indices. This figure depicts the ranges of apnea–hypopnea indices (AHI) by study. The cross hatched bars represent the minimum any technician scored AHI; the dotted bars represent the maximum AHI any of the technicians scored and; the black bars represent the mean of all scorers with Y-error bars representing the standard deviation.

parameter and a sleep staging parameter is seen in Fig. 6 which shows the coefficient of variation for AHI vs total sleep time (TST). As can be seen, there is significantly higher numbers for each study for AHI suggesting increased variability in that parameter.

In Table 1, the severity of OSAHS based on the different AHI determined by each scorer is shown. Table 2 shows the results of the random effects ANOVA. As can be seen, all studies, as expected, had significant variability in all parameters evaluated. In comparing the variability between technologists, there was less variability in the scoring of sleep staging related parameters (sleep efficiency, sleep latency and total sleep time) than there was with the AHI and the sum of apneas and hypopneas. The residual variability (unexplained) is also shown.

Table 3 shows the comparisons utilizing the kappa statistic. AHI, which is a function of both respiratory scoring (apneas + hypopneas) and sleep staging (total sleep time) had a kappa statistic of 0.24 for all studies and scorers. Apneas + hypopneas, a better assessment of respiratory scoring alone, showed a higher kappa statistic ($\kappa = 0.31$) as did sleep efficiency, an assessment of sleep staging alone ($\kappa = 0.44$).

4. Discussion

As in most medical tests, variability is an expected finding. The criteria that Rechtschaffen and Kales [2] set forth back in 1968 for sleep staging has been 'time-tested' and, as can be seen from our data, is relatively reliable from one laboratory to the next. The random effects analysis suggests that among the sleep staging variables, only sleep latency had significant variability between scorers. The sleep latency is the time from lights out to sleep onset. The onset of sleep has variable definitions. One definition is that sleep begins when there is one epoch of Stage 1, i.e. at least 15.5 seconds of that epoch (30 s duration) meet criteria for Stage 1. Another definition commonly used is three consecutive epochs of Stage 1. Some laboratories, define sleep latency at the point when the patient has been asleep for 2–3 consecutive minutes. Again, since our study was to assess variability of the scoring and not the defini-

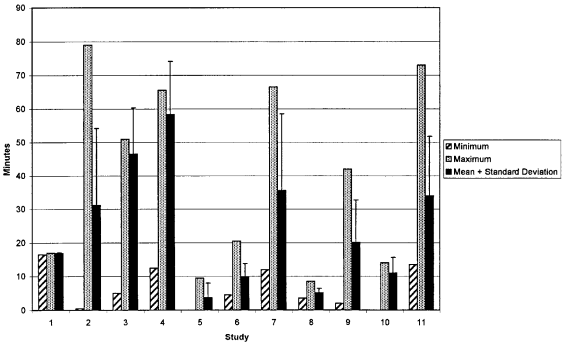

Fig. 3. Sleep latency. This figure depicts the ranges of sleep latency by study. The cross hatched bars represent the minimum any technician scored sleep latency; the dotted bars represent the maximum sleep latency any of the technicians scored and; the black bars represent the mean of all scorers with Y-error bars representing the standard deviation.

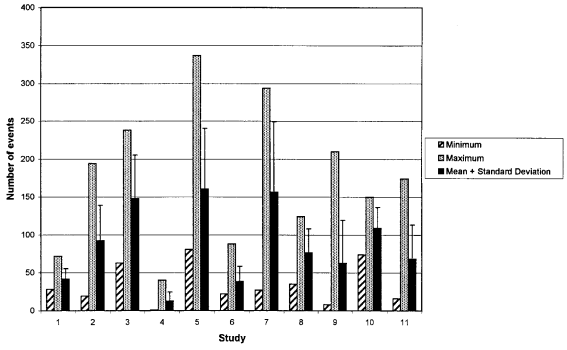

Fig. 5. Sum of apneas and hypopneas This figure depicts the ranges of the total of apneas and hypopneas by study. The cross hatched bars represent the minimum any technician scored the apnea + hypopnea sum; the dotted bars represent the maximum sum of apneas + hypopneas any of the technicians scored and; the black bars represent the mean of all scorers with Y-error bars representing the standard deviation.

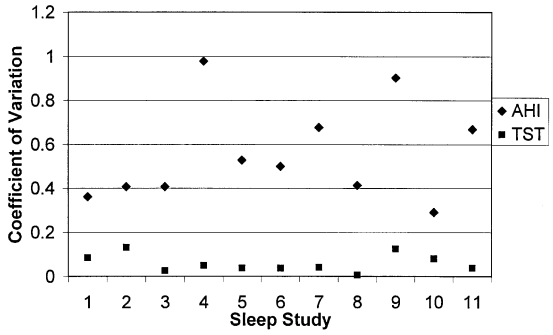

Fig. 6. Coefficient of variation. In this figure, a respiratory parameter (apnea–hypopnea index) is plotted against a sleep staging parameter (total sleep time). The coefficient of variation is the mean divided by the standard error of the mean, the higher the number, the more variability present.

tions, we do not have specific information on how these laboratories and technologists define sleep latency, which likely explains some of the variability in that parameter.

Respiratory scoring was found to be more variable than sleep staging. Fig. 6 comparing total sleep time, a staging variable, to apnea–hypopnea index, a respiratory variable, shows there was much more variability in scoring respiratory events. Fig. 4 shows for one study (#9), one technologist scored an AHI of 4.9 events/h whereas another scored an AHI of 79 events/h. In examining those two PsgT's separately, the PsgT that scored the AHI low (AHI 4.9) consistently scored lower than the average AHI (9/11 below the average) and the PsgT that scored high (AHI 79) consistently scored higher than the average (10/11 above the average). This particular test was the first half of a split night study and there was a wide variability in the AHI scoring with the mean + standard deviation (SD) 23.6 ± 21.4, the highest SD of all the studies. The total number of apneas + hypopneas scored ranged from 8–210 with most of the events scored as hypopneas. Interestingly, the study also had significant variability in the sleep staging, with the percent of wake (Stage 0) ranging from 14.5–51.7%. The reason for these disparities are unclear, and possibly may be due to frequent arousals and movements. This again points out that the scoring of both respiratory events and sleep combine to determine the AHI and errors in both can create disparity.

Table 1
Severity of OSAHS based on technician AHI score[a]

Sleep study	No Disease	Mild	Moderate	Severe
1	2	9	0	0
2	1	5	5	0
3	0	1	8	2
4	10	1	0	0
5	0	4	6	1
6	6	5	0	0
7	1	3	4	3
8	1	7	3	0
9	1	4	5	1
10	0	0	4	7
11	3	7	1	0

[a] No disease = AHI ≤ 5; mild AHI > 5, ≤15; moderate AHI > 15, ≤40; severe = AHI > 40

Table 2
Random effects analysis of variance

Variable	Studies	Technicians	Residual
Apnea–hypopnea index	12.19[a]	7.62[a]	7.84
Sleep efficiency	8.03[a]	1.53[b]	5.50
Sleep latency	17.32[a]	0	13.73
Sum of apneas + hypopneas	49.80[a]	35.39[a]	37.05
Total sleep time	97.84[a]	3.97	20.63

[a] $P = 0.05$.
[b] $P < 0.001$.

Other analysis also pointed out the differences in respiratory scoring. The random effects analysis shows for both respiratory variables examined, AHI and sum of apneas plus hypopneas, there was variability between scorers ($P < 0.0001$).

Utilizing the kappa statistic, it was shown there is significant variability in scoring both sleep variables and respiratory events. Although AHI had the lowest kappa, both apneas + hypopneas and sleep efficiency when categorized into degree of severity, showed variability. According to Fleiss's definition, only sleep staging exhibited moderate agreement [7].

The importance of the differences between technologist's scoring is underscored when one looks at how it may change a diagnostic category. In four of the studies, the diagnosis was relatively consistent, varying by only one diagnostic group for all scorers. However, for four other studies the diagnosis ranged from no disease to moderate OSAHS depending on which laboratory it was scored in (#2, #8, #9, #11). For two of the studies, the diagnosis could have ranged from mild OSAHS to severe OSAHS (#3, #5) and in one study it could have varied anywhere from no disease to severe OSAHS (#7). As AHI is the most common parameter looked at regarding defining OSAHS, this points out the problems with its use to determine whether treatment should be initiated and/or paid for by third party payers.

One explanation for this variability is possibly related to our sleep disorders laboratory's methods of monitoring. Nasal thermistry in recent years has come under significant criticism as a way of monitoring airflow [8–10]. The method is non-quantitative, though the criteria frequently utilized to

Table 3
Variability in AH1, sum of apneas and hypopneas, and sleep efficiency

Variable	κ statistic
Apnea–hypopnea index	0.24
Sum of apneas + hypopneas	0.31
Sleep efficiency	0.44

define an obstructive event such as an hypopnea (e.g. 50% reduction in airflow) is quantitative. More quantitative measurement techniques such as a full-face mask with an attached pneumotachometer, or an esophageal pressure measurement, are perceived by most laboratories to be too bulky or invasive and may disrupt patient's sleep patterns [12]. Use of these methods, or nasal cannula pressure transducers, however, would likely lead to diminished variability [9,11,13]. By improving the airflow signal and making it more quantitative, it is possible that the scoring of hypopneas might be less ambiguous.

Another problem with PSG scoring is related to definitions. Definitions of obstructive events vary from laboratory to laboratory, especially for hypopneas [3]. This, however, does not explain all the variability in this study, because there were differences between technicians related to the scoring of both apneas and hypopneas. It has been shown that the scoring of polysomnography can be more standardized. Establishment and documentation of clearer scoring rules such as was done in the Sleep Heart Health Study, can result in excellent interscorer variability [14]. Alternatively, use of computer scoring has also been suggested as a possible method to improve problems associated with manual scoring [15–17]. Recommendations for both techniques and scoring of respiratory events have been proposed for research purposes and perhaps similar recommendations should be proposed for clinical polysomnography [18].

Ways to decrease variability in the scoring of sleep studies should include: better standardization of monitoring equipment to improve the signals obtained from the physiologic response which will decrease variability. Additionally, reports generated from any sleep laboratory should include their individual scoring methods and definitions. This would at least let the reader of the report know which definitions were used to score the study. Finally, each laboratory should have quality control to assure that all scorers in the laboratory are comparable to each other. It may also be useful to set up centralized scoring such that intermittently technologists could be sent studies to score that can be returned to a 'scoring bureau' to assess their accuracy. All these measures would likely improve standardization between laboratories. Until measures such as these are taken, it will be difficult for physicians to know if the results reported on sleep studies performed in one laboratory are equivalent to those performed in another.

In conclusion, variability exists between polysomnography technologists in the scoring of sleep studies particularly regarding respiratory events. This variability may result not only in differences in the severity but even of the diagnosis of an individual patient's sleep apnea. Physicians should be aware this variability exists and sleep laboratories and other related organizations should work on decreasing this variability to improve the overall quality and comparability of polysomnography.

Acknowledgements

The author would like to acknowledge Ms. Julie Percy, RPsgT for her assistance in collecting the technical data; Oxford Instruments for their generous grant to assist in paying the participants; and Drs Phil Rust and Marcy Petrini for their statistical assistance.

References

[1] Young T, Palta M, Dempsey J, Skatrud J, Weber S, Badr S. The occurrence of sleep-disordered breathing among middle-aged adults. N Engl J Med 1993;328:1230–1235.
[2] Rechtschaffen A, Kales A, editors. A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. Washington, DC: Public Health Service, US Government Printing Office, 1968.
[3] Moser N, Phillips B, Berry D, Harbison L. What is hypopnea, anyway? Chest 1994;105:426–428.
[4] Whitney C, Gottlieb D, Redline S, Norman R, Dodge R, Shahar E, Surovec S, Nieto F. Reliability of scoring respiratory disturbance indices and sleep staging. Sleep 1998;21(7):749–757.
[5] Drinnan M, Murray A, Griffiths C, Gibson G. Interobserver variability in recognizing arousal in respiratory sleep disorders. Am J Respir Crit Care Med 1998;158:358–362.
[6] Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Measure 1960;XX:37–46.
[7] Fleiss J. Measurement of interrater agreement. In: Fleiss J, editor. Statistical methods for rates and proportions. New York: Wiley and Sons, 1981:212–236.
[8] Berg S, Haight J, Yap V, Hoffstein V, Cole P. Comparison of direct and indirect measurements of respiratory airflow: Implications for hypopneas. Sleep 1997;20(1):60–64.
[9] Montserrat J, Farré R, Ballester E, Felez M, Pastó M, Navajas D. Evaluation of nasal prongs for estimating nasal flow. Am J Respir Crit Care Med 1997;155:211–215.
[10] Hosselet J, Norman R, Ayappa I, Rapoport D. Detection of flow limitation with a nasal cannula pressure transducer system. Am J Respir Crit Care Med 1998;157:1461–1467.
[11] Ballester E, Badia J, Hernández L, Farré R, Navjas D, Monserrat J. Nasal prongs in the detection of sleep-related disordered breathing in the sleep apnoea-hypopnoea syndrome. Eur Respir J 1998;11:880–883.
[12] Phillipson E, Remmers J. Indications and standards for cardiopulmonary sleep studies. Am Rev Respir Dis 1989;139:559–568.
[13] Norman R, Ahmed M, Waisleben J, Rapoport D. Detection of respiratory events during NPSG: nasal cannula/pressure sensor vs. thermistor. Sleep 1997;20:1175–1184.
[14] Whitney C, Gottleib D, Redline S, Norman R, Dodge R, Shahar E, Surovec S, Nieto F. Reliability of scoring respiratory disturbance indices and sleep staging. Sleep 1998;21(7):749–757.
[15] White D, Gibb T. Evaluation of a computerized polysomnographic system. Sleep 1998;21(2):188–196.
[16] Carrasco O, Montserrat J, Lioberes P, Ascasco C, Ballester E, Fomas C, Rodriguez-Roisin R. Visual and different automatic scoring profiles of respiratory variables in the diagnosis of sleep apnoea–hypopnoea syndrome. Eur Respir J 1996;9:125–130.
[17] Lord S, Sawyer B, Pond D, O'Connell D, Eyland A, Mant A, Hensley M, Saunders N. Interrater reliability of computer-assisted scoring of breathing during sleep. Sleep 1989;12(6):550–558.
[18] AASM Task Force. Sleep-related breathing disorders in adults: Recommendations for syndrome definition and measurement techniques in clinical research. Sleep 1999;5:667–689.