Turkish Archives of Otolaryngology
Türk Otolarengoloji Arşivi

Turk Arch Otolaryngol 2014; 52: 17-21

Comparison of Manual and Automated Scoring Techniques in Polysomnography

Original Investigation

Mehmet Aşık, Aslı Bostancı, Murat Turhan

Department of Otorhinolaryngologic Diseases Faculty of Medicine, Akdeniz University, Antalya, Turkey

Abstract

Objective: Polysomnography (PSG) scoring can be performed manually or with an automated programme. The purpose of this study is to compare two different scoring techniques in PSG.

Methods: The sleep recordings of 120 patients with obstructive sleep apnoea (OSA) suspicion who underwent PSG at ear nose and throat clinic of Akdeniz University Hospital between January and June 2013 were retrospectively analysed. Patients were divided into 4 groups according to the apnoea-hypopnea index (AHI): AHI<5 (normal), AHI 5-15 (mild OSA), AHI 15-30 (moderate OSA) and AHI>30 (severe OSA). There were 30 patients in each group. Manually scored recordings were reanalysed with an automated programme and the results, including sleep stages and respiratory events, were compared.

Results: A total of 86.400 epochs of 120 patients were reanalysed. In all patients, the total sleep time and sleep efficiency were decreased with automated scoring by 29 min and 6%, respectively (p=0.001). The percentage of stage I sleep was higher and REM was lower, respectively (p=0.001 for both parameters). In automated scoring, the number of cases of obstructive and central apnoea were lower (p=0.001), and the number of cases of hypopnoea, mean apnoea duration and hypopnoea duration were higher (p=0.001, p=0.001 and p=0.039, respectively). There were no statistically significant differences in the total AHI and REM AHI between two scoring techniques (p=0.053 and p=0.319, respectively). However, NREM AHI was significantly higher in the automated scoring (p=0.002). Sensitivity, specificity, positive predictive value and negative predictive value of automated scoring were 98.88%, 93.33%, 97.80% and 95.55%, respectively.

Conclusion: Automated scoring is not sufficiently accurate for many sleep parameters. Inconsistency between the two techniques is apparent, especially in patients with mild to moderate forms of OSA.

Key Words: Polysomnography, manual scoring, automated scoring, OSA

Introduction

Obstructive sleep apnea (OSA) has become a public health problem due to its high frequency rate and results. Identification of true patients and accurate interpretation of disease severity directly affect the success of treatment. Polysomnography (PSG) is the common name of the techniques used for recording neurophysiological, cardiorespiratory, and other physiological and physical parameters during sleep at night (1). PSG provides monitoring of sleep stages in detail and also information about the interaction between the functions of various organs and systems and sleep-wake states. It is the gold standard diagnostic method for OSA.

Scoring of PSG records is typically performed by an experienced sleep technician manually (visually). This recording of at least 6-hour sleep is scored by the technician following each epoch on monitor. The length of one epoch is 30 seconds, and about 720 epochs are evaluated separately. This process takes nearly 80-180 minutes, even when performed by the most experienced technician, and it is quite troublesome and time-consuming. In the advent of recently developed software programs, automated scoring can be performed in a shorter time without the need for technician support. Although these programs score according to standard criteria, consistency and reliability between manual and automated scorings are controversial. There may be errors, particularly in recognizing the process of passing from the awake state to Stage I and rapid eye movement (REM) sleep and in distinguishing arousal, epileptic activity, and parasomnia (2).

In the literature, most of the researchers who used these two scoring techniques compared the results using old criteria (3, 4) or analyzed only healthy individuals by automated scoring techniques (5). In this study, PSG recordings of both healthy and unhealthy cases were analyzed with manual and automated scoring techniques, considering the new criteria, and the results of both methods were compared.

Methods

The sleep recordings of the patients who underwent PSG due to a prediagnosis of OSA at the Outpatient Clinic of Otorhinolaryngologic Diseases of Akdeniz University Hospital between the dates of January and June 2013 were retrospectively evaluated. The patients were divided into 4 groups according to the apnea-hypopnea index (AHI): AHI<5 (normal), AHI 5-15 (mild OSA), AHI 15-30 (moderate OSA), and AHI>30 (severe

This study was presented at the 35th Turkish National Otolaryngology and Head and Neck Surgery Congress, 2-6 November 2013, Antalya, Turkey

Address for Correspondence:
Aslı Bostancı, Department of Otorhinolaryngologic Diseases Faculty of Medicine, Akdeniz University, Antalya, Turkey
Phone: +90 242 249 68 43
E-mail: draslibostanci@gmail.com

Received Date: 31.01.2013
Accepted Date: 13.02.2014

DOI:10.5152/tao.2014.422

OSA). Each group included 30 patients who were randomly selected (total patient number: 120). Manually scored data obtained from all patients were reanalyzed using the Profusion® sleep scoring program. Parameters, including the recordings of sleep stages and respiratory events, were compared. Patients diagnosed with central sleep apnea syndrome and obesity-hypoventilation syndrome in manual scoring and patients found to have inadequate sleep efficiency in the PSG were excluded from the study.

PSG was performed using a 44-channel E-series device with electrodes placed according to the international 10-20 system (Compumedics Profusion). Many parameters, including electroencephalography (C3-A2, C4-A1), left and right electrooculography, electromyography (maxilla and anterior tibial muscle), electrocardiography, thoracoabdominal movements, oronasal airflow, and oxygen saturation, were recorded. PSG was scored manually by an otorhinolaryngologist who was experienced in sleep disorders. Cessation of airflow for at least 10 seconds with increased ventilatory effort (obstructive) or no associated ventilatory effort (central) is defined as apnea. Hypopnea was defined a as 50% (or greater) reduction in airflow or decreased respiratory amplitude less than 50% accompanied by a 3% reduction in oxygen saturation and arousal and also obstruction lasting for at least 10 seconds. AHI was defined as the number of apneas and hypopneas per hour during sleep. Sleep stages were scored according to the 2012 American Academy of Sleep Medicine criteria (6). This study was approved by the ethics committee of Akdeniz University (07.06.2013 / Document no: 009448). Data were used after getting written informed consent from the patients who participated in the study.

Statistical analysis

Statistical Package for Social Sciences (SPSS) for Windows 17.0 was used for statistical analyses of data. Descriptive statistical data were calculated initially (mean, standard deviation, frequency). Then, Wilcoxon signed-ranks test was employed to compare quantitative data. Additionally, Bland and Altman analyses were performed for the selected comparisons. The value of $p<0.05$ was considered statistically significant.

Results

In our study, a total of 86,400 epochs of 120 patients were evaluated. Of the patients, 24 were female and 96 were male. Their age range was from 21 to 74 years, and the mean age was 46.75±11.37. The mean body mass index was 29.74±4.58 kg/m^2. A 29-min decrease in total sleep time and 6% reduction in sleep efficiency were determined through automated scoring in all patients (for both parameters p=0.001). The percentage of Stage 1 was found to be significantly higher in automated scoring (p=0.001), while there was no statistically significant difference between automated and manual scoring systems for Stage II and Stage III (p=0.26 and p=0.19, respectively). The percentage of REM stage was significantly lower for all patients and all OSA subgroups in automated scoring (Table 1).

Table 1. Comparison of results according to sleep stages

	Manual Scoring	Automated scoring	p
Total sleep time. min.	421.85±55.22	392.85±57.78	0.001
Sleep efficiency. %	90.34±9.62	84.32±11.04	0.001
Stage I. %	2.54±1.51	5.82±4.05	0.001
AHI<5	2.10±1.14	5.17±4.11	0.001
AHI 5-15	2.50±1.48	5.03±2.84	0.001
AHI 15-30	2.97±1.66	6.18±3.98	0.001
AHI>30	2.60±1.63	6.91±4.89	0.001
Stage II. %	59.57±11.96	58.99±9.53	0.261
AHI<5	55.43±9.52	56.06±7.84	0.003
AHI 5-15	55.95±10.63	56.90±9.72	0.716
AHI 15-30	60.23±10.23	61.33±9.53	0.018
AHI>30	63.69±14.68	61.68±9.97	0.229
Stage III. %	23.91±12.78	23.40±11.06	0.193
AHI<5	26.87±8.69	25.32±7.3	0.087
AHI 5-15	26.50±11.32	25.46±10.84	0.379
AHI 15-30	20.27±11.51	20.92±11.65	0.168
AHI>30	22.01±17.35	21.92±13.23	0.942
REM. %	13.97±6.64	11.7±6.06	0.001
AHI<5	15.62±6.60	13.44±4.95	0.004
AHI 5-15	15.06±7.06	12.59±6.83	0.001
AHI 15-30	13.51±6.32	11.56±5.77	0.001
AHI>30	11.68±6.16	9.46±6.11	0.001

AHI: apnea-hypopnea index; REM: rapid eye movements

Regarding respiratory events, in automated scoring, the numbers of cases with obstructive and central apnea were significantly lower, while the numbers of cases with hypopnea, mean apnea duration, and mean hypopnea duration were higher (p=0.001, p=0.001, and p=0.039, respectively) (Table 2).

No difference was found between the two scoring techniques in terms of total AHI and REM AHI (p=0.053 and p=0.319, respectively), but NREM AHI was significantly higher in automated scoring (p=0.002) (Table 3).

In the diagnostic comparison of the two scoring techniques, 2 of 30 cases found to be healthy in manual scoring were evaluated as sick (mild OSA) in automated scoring. Of 90 patients who were found to be sick in manual scoring, one was accepted as healthy and 89 were evaluated as sick in automated scoring. In terms of OSA subgroups, of 30 patients who were considered to have mild OSA in manual scoring, one was normal, 27 had mild OSA, and 2 had moderate OSA in automated scoring. Of 30 patients considered to have moderate OSA in manual scoring, one was found to have mild OSA, 26 had moderate OSA, and 3 had severe OSA in automated scoring. All 30 patients who were found to have severe OSA in manual

Table 2. Comparison of respiratory events

	Manual Scoring	Automated scoring	p
Number of OA	96.22±121.96	92.28±118.65	0.001
AHI<5	5.87±6.56	5.87±6.09	0.916
AHI 5-15	35.33±20.67	36.00±19.89	0.246
AHI 15-30	75.27±33.71	69.00±30.97	0.001
AHI>30	268.40±126.60	258.57±127.74	0.001
OA duration. sec.	19.91±7.58	21.03±7.64	0.001
AHI<5	13.26±6.07	15.39±7.79	0.079
AHI 5-15	18.65±4.16	20.07±4.38	0.001
AHI 15-30	20.65±5.20	21.14±5.33	0.001
AHI>30	27.08±7.30	27.53±7.33	0.001
Number of CA	3.68±8.04	2.73±5.84	0.001
AHI<5	0.83±1.41	0.80±1.29	0.729
AHI 5-15	1.87±3.04	1.37±1.81	0.259
AHI 15-30	7.27±13.3	5.13±10.18	0.003
AHI>30	4.77±7.13	3.63±4.32	0.122
Number of hypopnea	28.36±34.17	35.52±38.67	0.001
AHI<5	6.13±6.02	10.33±9.02	0.001
AHI 5-15	18.40±15.29	26.13±25.05	0.002
AHI 15-30	52.40±29.97	60.93±32.07	0.023
AHI>30	36.50±48.42	44.67±53.64	0.006
Hypopnea duration, sec.	27.81±9.54	28.93±9.38	0.039
AHI<5	23.68±11.12	25.07±11.01	0.126
AHI 5-15	29.04±8.36	30.73±6.39	0.110
AHI 15-30	30.38±6.24	31.30±6.48	0.409
AHI>30	28.14±10.72	28.62±11.51	0.775

AHI: apnea-hypopnea index; OA: obstructive apnea; CA: central apnea

Table 3. Comparison of results according to AHI values

	Manual Scoring	Automated scoring	p
Total AHI	21.90±22.02	22.63±21.87	0.053
AHI<5	2.04 ±1.35	2.69±1.63	0.018
AHI 5-15	8.68±2.91	9.23±4.31	0.028
AHI 15-30	21.92±4.29	22.75±5.63	0.168
AHI>30	54.96±15.53	55.15±15.39	0.845
NREM AHI	21.70±22.53	22.65±22.35	0.002
AHI<5	1.59±1.30	2.25±1.54	0.002
AHI 5-15	8.24±3.75	9.68±4.72	0.017
AHI 15-30	21.16±4.71	22.54±6.02	0.025
AHI>30	55.81±15.35	56.14±14.97	0.721
REM AHI	22.81±23.60	21.04±21.96	0.319
AHI<5	4.10±6.12	5.24±6.37	0.053
AHI 5-15	12.55±13.33	13.16±13.23	0.611
AHI 15-30	27.32±19.74	25.10±19.69	0.230
AHI>30	47.29±23.86	40.64±25.28	0.026

AHI: apnea-hypopnea index; REM: rapid eye movement; NREM: nonrapid eye movement

scoring were also evaluated as severe OSA patients in automated scoring (Table 4).

The sensitivity, specificity, positive predictive value, and negative predictive value of automated scoring were found to be 98.88%, 93.33%, 97.80%, and 95.55%, respectively. On the other hand, the values of sensitivity for the subgroups of normal, mild OSA, moderate OSA, and severe OSA were determined as 93.33%, 90%, 86.66%, and 100%, respectively.

Discussion

Evaluation of sleep scoring and PSG recordings has been performed according to the Rechtschaffen and Kales (R&K) criteria until recently (7). The American Academy of Sleep Medicine first published a new manual on scoring of sleep and associated events in 2007 (8) and then updated it in 2012 (6). Today, scoring of sleep is performed, based on these rules.

In this study, manually scored recordings of 120 patients who underwent PSG due to a prediagnosis of OSA were reanalyzed with automated scoring, and the results of both techniques were evaluated. Various parameters, including sleep stages and respiratory events, were compared, but only 40% of them were found to be consistent with each other. Considering the comparison of the two scoring techniques, the rate of consistency was reported to range from 60% to 90% in the literature (9-11). Öztürk et al. (12) compared patients diagnosed with OSA and found the rate of consistency to be 58%, which emphasized that automated systems were less reliable than assumed. Possible causes of this inconsistency between the two techniques may be the performance of automated scoring with different devices in different studies, the restricted number of samples in some studies, or inhomogeneous cases of the sampling in terms of the severity of OSA.

In our study, total sleep time was found to be shorter, and Stage I sleep was found to be longer in automated scoring. After re-evaluation of the recordings, it was detected that automated scoring assessed some part of sleep as awake time, and thus, a decrease was observed in the rates of sleep time and efficiency. Moreover, it was found that awake state in manual scoring was evaluated as Stage I in automated scoring. Passing from the awake state to Stage I is usually characterized with retardation of the EEG, and a decrease is observed in the amplitude and frequency of alpha activity (1). As seen, automated scoring can not perceive the changes that occur in the wave pattern and evaluates awake state as Stage I. In automated scoring, the REM stage was found to be lower for all patients, because it was mostly evaluated as Stage II. However, the REM stage has specific saw tooth waves (1). This can cause the detected apneas to shift to the NREM stage and the diagnosis of REM-based OSA to be overlooked.

Table 4. Diagnostic comparison of results according to OSA subgroups

			Manual scoring			
			Normal (n=30)	OSA (n=90)		
				Mild OSA (n=30)	Moderate OSA (n=30)	Severe OSA (n=30)
Automated scoring	Normal		28	1	0	0
	OSA	Mild OSA	2	27	1	0
		Moderate OSA	0	2	26	0
		Severe OSA	0	0	3	30

OSA: obstructive sleep apnea

The numbers of cases with obstructive apnea, mixed apnea and central apnea were found to be low in automated scoring, while the number of the cases with hypopnea was found to be higher. The apnea that was marked in manual scoring was marked as another type of apnea in automated scoring, or present apnea was not marked as apnea. All of these situations lead to errors in diagnosing and misinterpreting the cause of OSA and thus can directly affect treatment.

In the study, it was revealed that some apneas recorded with automated scoring lasted longer, which was contrary to normal physiology. Therefore, the recordings were reanalyzed, and it was found that the beginning and end of apnea were not accurately marked in automated scoring. For example, 2 apneas recorded as 12 min in some epochs of manual scoring were recorded as 20- and 26-min apneas in automated scoring. On the contrary, the number of hypopneas was found to be higher in automated scoring. This also resulted from the fact that the beginning and end of hypopnea were marked as different, and in addition, longer hypopneas occurred in automated scoring. The observed differences affect the number of apneas and also the OSA classification by leading to changes in total AHI value.

In our study, it was seen that automated scoring displayed quite high rates of sensitivity and specificity (98.88% and 93.33%, respectively) for AHI values in all patients. However, the rate of specificity decreased in the subgroups of mild and especially moderate OSA apparently (90.0% and 86.6%, respectively). The fact that diagnostic AHI intervals do not have an upper limit for severe OSA (AHI>30) and that they are relatively narrow for mild (AHI, 5-15) and moderate OSA (AHI, 15-30) can explain the decreased specificity. Similarly, Pittman et al. emphasized that inconsistency between the two scoring techniques was clear for the subgroup of moderate OSA (3). As stated above, any categorical change in OSA subgroups would affect the treatment approach. However, PSG alone is not sufficient for deciding on the treatment method. It can be more beneficial in company with physical examination findings, concurrent diseases of the patients, and the patient's decision.

In spite of the adequate number of patients in our study, there were some limitations, such as the restricted number of cases in each subgroup and the possibility of potential bias for the selection of cases, which is a nature of retrospective studies. On the other hand, some features, including "use of current criteria in scoring," "scoring performed by a specialist physician experienced on respiratory disorders in sleep," "the study group consisted of both healthy and sick individuals," and "unlike most other studies, involving OSA patients with different weights," make our study valuable.

Conclusion

Automated scoring misevaluates many PSG parameters. Inconsistency between the two scoring techniques is more apparent, especially in the subgroups of mild and moderate OSA. Automated scoring technique may lead to errors in diagnosis and in the treatment decision. Manual staging is superior to automated staging in terms of diagnosing, although it is more troublesome, and its results are obtained in a longer time.

Ethics Committe Approval: Ethics committee approval was received for this study from the ethics committee of Akdeniz University (07.06.2013 / Document no: 009448).

Informed Consent: Written informed consent was obtained from patients who participated in this study.

Peer-review: Externally peer-reviewed.

Author Contributions: Concept - A.B., M.T.; Design - M.A., M.T.; Supervision - A.B., M.T.; Funding - M.A., A.B.; Materials - M.A., A.B.; Data Collection and/or Processing - M.A., A.B.; Analysis and/or Interpretation - A.B., M.T.; Literature Review - M.A.; Writing - M.A., A.B.; Critical Review - M.T.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study has received no financial support.

References

1. Köktürk O. Diagnostic methods and polysomnography for respiratory disorders during sleep. In: Respiratory system and diseases. Özlü T, Metintaş M, Karadağ M, Kaya A (Eds). İstanbul: İstanbul Tıp Kitapevi; 2010.p.2109-25.
2. Köktürk O. Scoring sleep recordings. Solunum 2013; 15: 14-29.
3. Pittman SD, MacDonald MM, Fogel RB, Malhotra A, Todros K, Levy B, et al. Assessment of automated scoring of polysomnographic recordings in a population with suspected sleep-disordered breathing. Sleep 2004; 27: 1394-403.
4. Caffarel J, Gibson GJ, Harrison JP, Griffiths CJ, Drinnan MJ. Comparison of manual sleep staging with automated neural network-based analysis in clinical practice. Med Biol Eng Comput 2006; 44: 105-10. **[CrossRef]**
5. Pardey J, Roberts S, Tarassenko L, Stradling J. A new approach to the analysis of the human sleep/wakefulness continuum. J Sleep Res 1996; 5: 201-10. **[CrossRef]**
6. Berry RB, Brooks R, Gamaldo CE, Harding SM, Marcus CL and Vaughn BV for the American Academy of Sleep Medicine. The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, Darien, Illinois: American Academy of Sleep Medicine, 2012; www.aasmnet.org.
7. Rechtschaffen A, Kales A. A manual of standardized, techniques and scoring system for sleep stages in human subjects. Washington DC, US Government Printing Office. NIH Publication 1968: 204.
8. Iber C, Ancoli-Israel S, Chesson A, Quan SF. The AASM Manual for the Scoring of Sleep and associated Events. Rules Terminology and Technical Specifications. 1st Ed. Westchester, Illinois: American Academy of Sleep Medicine, 2007.
9. Schaltenbrand N, Lengelle R, Toussaint M, Luthringer R, Carelli G, Jacqmin A, et al. Sleep stage scoring using the neural network model: comparison between visual and automatic analysis in normal subjects and patients. Sleep 1996; 19: 26-35.
10. Malhotra A, Younes M, Kuna ST, Benca R, Kushida CA, Walsh J, et al. Performance of an automated polysomnography scoring system versus computer-assisted manual scoring. Sleep 2013; 36: 573-82.
11. Redline S, Budhiraja R, Kapur V, Marcus CL, Mateika JH, Mehra R, et al. The scoring of respiratory events in sleep: reliability and validity. J Clin Sleep Med 2007; 3: 169-200.
12. Oztürk O, Mutlu LC, Sağcan G, Deniz Y, Cuhadaroğlu C. The concordance of manuel (visual) scoring and automatic analysis in sleep staging. Tuberk Toraks 2009; 57: 306-13.