Original Research

Exploration of Reliable Parameters Scored by Automated Analysis in Polysomnography

Nahid Nikoee[1], Mohammad Seyed Hoseini[2], Arezu Najafi[3], Amin Amali[4], Behrouz Amirzargar[4], Reihaneh Heidari[4]*

[1.] *Department of Management, School of Humanities, Electronic Branch, Islamic Azad University, Tehran, Iran*
[2.] *Department of Industrial Engineering, School of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran*
[3.] *Occupational Sleep Research Center, Baharloo Hospital, Tehran University of Medical Sciences, Tehran, Iran*
[4.] *Otorhinolaryngology Research Center, Tehran University of Medical Sciences, Tehran, Iran; Occupational Sleep Research Center, Tehran University of Medical Sciences, Tehran, Iran; Department of Otorhinolaryngology, Imam Khomeini Educational Complex Hospital, Valiasr Hospital, Tehran University of Medical Sciences, Tehran, Iran*

Received: 03 Sep. 2018 Accepted: 20 Oct. 2018

Abstract

Background and Objective: Polysomnography (PSG) is the gold standard for diagnosis of sleep disorders. Several software programs are available to analyze sleep tests according to available guidelines and decrease the time and cost of PSG analysis. This study aimed to compare the parameters of automated analyzer software with analysis of trained technician (manual analysis).

Materials and Methods: Twenty patients who underwent full-night PSG were randomly selected. A sleep technologist who was blind to the study, scored sleep stages and respiratory events according to recommended criteria of American Academy of Sleep Medicine (AASM) 2013, then an auto analysis was done using N-7000 amplifier. Results of auto analysis and manual analysis were compared. Descriptive statistics and paired t-test were used for data analysis.

Results: Total sleep time (TST) and sleep efficiency (SE) calculated by auto analysis was significantly more than manual analysis (511.82 ± 35.34 vs. 396.85 ± 75.97 for TST and 95.47 ± 3.74 vs. 74.14 ± 35.34 for SE, respectively). Furthermore, there was no concordance for sum of apneas and hypopneas during TST. However, calculated number of hypopneas in non-rapid eye movement (NREM) stage in auto analysis and manual analysis was quite similar. The least precision was observed in scoring of stages 3 and REM for auto analysis scoring and the most similarity for scoring of stage N2.

Conclusion: Detecting hypopneas in NREM stage by auto analysis maybe the reliable parameter that could help the technicians during analysis of sleep test. There is a need for more advanced automated algorithms. Furthermore, manual analysis is superior to automated one in PSG analysis according to the current results.

Keywords: Sleep apnea; Polysomnography; Sleep

Citation: Nikoee N, Seyed Hoseini M, Najafi A, Amali A, Amirzargar B, Heidari R. **Exploration of Reliable Parameters Scored by Automated Analysis in Polysomnography.** J Sleep Sci 2018; 3(3-4): 90-94.

Introduction

Obstructive sleep apnea (OSA) is one of the most common sleep disorders characterized by repeated collapse of upper airway during sleep (1, 2). Many efforts are made to improve and facilitate diagnosis of OSA (1). Attended full polysomnography (PSG) using electroencephalography (EEG) is the gold standard of diagnosis (1). Each PSG test requires manual scoring that is time and cost-consuming. Strategies to decrease the cost and time of the test, especially scoring, are considered one of the important challenges in the field of sleep medicine (3). One strategy to reduce cost and time of test is using automated scoring (3). Several studies have raised the problem of inter-

***Corresponding author**: R. Heidari, Otorhinolaryngology Research Center, Tehran University of Medical Sciences, Tehran, Iran; Occupational Sleep Research Center, Tehran University of Medical Sciences, Tehran, Iran; Department of Otorhinolaryngology, Imam Khomeini Educational Complex Hospital, Valiasr Hospital, Tehran University of Medical Sciences, Tehran, Iran*
Tel: +98 21 55460184, Fax: +98 21 55648189
Email: rhn_heidari@yahoo.com

scorer variability in manual scoring that may make its reliability and validity questionable. Furthermore, different companies have introduced various automated scoring on their devices (4). However, manual scoring performed by sleep technologist and edited by sleep specialist is still superior and more accurate than the automated one (1, 3).

Automated analyses are shown to overestimate respiratory events and underestimate sleep efficiency (SE) (1, 5). Computer-generated report uses fast Fourier transform (FFT) algorithms; thus, they are too sensitive and may falsely score stages that have error in recording signals or artifacts such as popping electrodes and sweat artifacts (1). This may lead to misjudging of sleep stages and respiratory events (1, 5, 6).

Barreiro *et al.* also indicated limited concordance between automated analysis and manual scoring, especially in terms of rapid eye movement (REM) stage and hypopneas (7). However, evidence is available that automated scoring may have good concordance with the manual one and these reports recommend automated analysis to save time and also cost of the test (8). Younes and Hanly reported the role of digitally-obtained information such as *k* complexes and spindles in reducing inter-rater variability during manual analysis (9).

Several automated analyses are recommended to reduce the time of analysis for sleep technologists. In this regard, along with limited and conflicted results for usefulness of automated versus manual scoring, reliability and reproducibility of automated scoring needs to be validated against manual scoring. Up to our knowledge, no study in Iran has evaluated the issue in population of patients with OSA.

Accordingly, this study aimed to evaluate the accuracy of Embla system auto analysis against manual scoring in terms of sleep stages and respiratory events. We sought to determine the agreement of scoring respiratory events and sleep stages. The results would help sleep technicians and physicians in terms of areas that automated scoring can be more reliable and valid.

Materials and Methods

In the current study, 20 previously-recorded in-laboratory attended PSGs were chosen. The studies were randomly selected from a database of 20 patients who underwent full attended PSG for OSA confirmation. The current study was conducted at sleep clinic of Baharloo Hospital, Tehran, Iran, in 2017 and the tests were selected randomly among recorded available PSGs. The split-night and titration studies were excluded. All patients underwent full-night PSG (class I) using EEG, electrooculography (EOG), electromyography (EMG) (legs and chin), movements of thoracic cage and abdomen, electrocardiography (EKG), and o_2 saturation. The recorded tests were analyzed according to American Academy of Sleep Medicine (AASM) 2013 manual scoring guideline. A sleep technologist scored epoch by epoch of sleep stages and respiratory events according to recommended criteria of AASM and a sleep specialist revised the manual scoring performed by sleep technician; both were blinded to study aims.

To perform a comparison between manual and auto analysis, an auto analysis was done using N-7000 amplifier (Embla Systems, LLC, Ontario, Canada). Results of auto analysis were compared with the ones scored by trained sleep technologist in terms of total sleep time (TST), sleep stages, and respiratory events including hypopnea and apnea.

To describe the data, mean and standard deviation (SD) were used for qualitative variables and frequency and percentage for continuous variables. Paired t-test (Wilcoxon test for non-normal variables) was used to calculate mean difference and P-value for the comparison of auto and manual scoring. P-value less than 0.0500 was considered significant.

Results

This was a cross-sectional study. Mean and SD of age and body mass index (BMI) of study participants were 40.85 ± 9.33 years and 29.90 ± 6.10 kg/m^2, respectively. 85% of the patients were men.

Participants' TST and SE by manual scoring was more than those by auto analysis (511.82 ± 35.34 vs. 396.85 ± 75.97 for TST and 95.47 ± 3.74 vs. 74.14 ± 35.34 for SE) (Table 1).

Table 1. Demographic characteristics of study participants

Variable		Mean ± SD
Age (year)		40.85 ± 9.33
Height (cm)		168.39 ± 13.66
Weight (kg)		89.02 ± 19.92
BMI (kg/m^2)		29.94 ± 6.13
		Frequency (%)
Age group (year)	21-40	10 (50)
	41-60	10 (50)
Sex	Male	17 (85)
	Female	3 (15)

SD: Standard deviation; BMI: Body mass index

Table 2. Respiratory events and sleep stages by manual analysis versus autoanalysis

Category	Auto analysis	Manual analysis	Mean difference	P-value
	Mean ± SD	Mean ± SD		
TST (minute)	511.82 ± 35.34	396.85 ± 75.97	114.97	< 0.0001
Sleep period (minute)	528.29 ± 30.73	476.51 ± 47.60		
SE (%)	74.14 ± 13.93	95.47 ± 3.74	21.33	< 0.0001
Apnea/Hypopnea index (AHI)	122.55 ± 160.60	237.80 ± 219.15	115.25	< 0.0001
Hypopnea	122.55 ± 160.60	129.95 ± 83.26	7.40	0.1230
N1	0.15 ± 0.46	106.52 ± 51.01	106.37	< 0.0001
N2	511.39 ± 35.61	214.67 ± 63.57	296.72	< 0.0001
N3	0.27 ± 0.69	9.80 ± 14.84	9.52	0.0110
REM stage	0	66.22 ± 27.01	66.22	< 0.0001
Waking stage	24.34 ± 20.52	139.31 ± 77.14	114.97	< 0.0001
AHI	14.87 ± 20.15	35.36 ± 31.22	20.49	< 0.0001
Hypopnea (REM)	0	23.82 ± 17.85	23.82	< 0.0001
Hypopnea (NREM)	14.49 ± 19.86	17.94 ± 13.79	3.45	0.1540
Apnea (REM)	0	18.04 ± 27.15	18.04	0.0080
Apnea (NREM)	0	15.25 ± 25.26	15.25	0.2310

N1: Stage 1 of sleep; N2: Stage 2 of sleep; N3: Stage 3 of sleep; REM: Rapid eye movement; NREM: Non-rapid eye movement; TST: Total sleep time; SE: Sleep efficiency; AHI: Apnea hypopnea index; SD: Standard deviation

Hypopnea was defined as 30% decrease of nasal flow plus 3 and 4 percent desaturation. Respiratory effort-related arousals (RERA) were also included in the number of hypopneas.

The difference was significant and no concordance was observed between the two approaches of analysis in this term. The sum of apneas and hypopneas during TST was much higher by manual analysis (median 150 for manual analysis and 43 for auto analysis). However, calculated number of hypopneas in non-REM (NREM) stage in auto and manual analysis was quite similar and the difference was not statistically significant (Table 2). This finding indicated significant difference of auto analysis and manual scoring in terms of calculating number of apneas.

Regarding scoring of different sleep stages using two approaches (automated against manual scoring), the least precision was observed in scoring of stages N3 and REM for automated scoring and the most similarity between the two approaches was observed for scoring of stage N2 (Table 2). The only concordant parameter between manual and auto analysis was hypopneas in NREM stage as depicted in table 2.

Discussion

As research about sleep disorders and its related disease progresses, demands for questionnaires and devices for screening and diagnosis of sleep disorders become more warranted.

The most useful technique for diagnosis of sleep disorders and especially sleep-disordered breathing (SDB) (i.e., sleep apnea) is PSG, which its results require to be analyzed and interpreted [e.g., sleep stages based on recorded EEG during overnight PSG (10), other diseases such as seizure which affect EEG waves (11)]. There is also a need to analyze respiratory events such as apnea, hypopnea, and respiratory effort-related arousals (RERAs) using nasal flow and thermistor signals (10).

In our sleep lab, trained sleep technicians with more than 6000 experience of PSG analysis perform scoring of sleep tests and then trained sleep specialists edit and interpret the results; because of the need for early report of tests and better diagnosis of SDB to prevent adverse outcomes, faster and cheaper analysis is required in many sleep labs; thus, many companies design and produce PSG software programs to offer algorithms for automated analysis. In this study, we compared 20 patients' sleep studies which were analyzed with both auto analysis and manual analysis to find out how much auto analysis differs and whether it can assist or even replace technicians' manual scoring.

Most of our patients, which were selected randomly, were men (85%) with high mean BMI that brings into mind the probability of association between male gender and high BMI with increase of possible SDB that requires further medical workup such as PSG. The auto analysis report indicated a significant higher TST (563 vs. 510 minutes, 28% more with auto) and N2 and N3 stages than manual scoring, but it could not detect N1 and REM stages. Accordingly, we cannot rely

on aforementioned outputs of automated analysis, especially SE that is calculated based on measured TST of the patient. As automated REM stage scoring was not similar to manual analysis, therefore, REM stage calculation by auto analysis also cannot be reliable especially for patients whose REM stages and events on REM stages like REM-related apnea or hypopnea we want to survey. Furthermore, auto analysis revealed lower wakefulness in comparison with technician's analysis and this may be the cause that calculated SE in auto analysis was more than the manual one.

Regarding scoring of respiratory events during sleep, auto analysis could not score the apnea events as accurate as manual analysis and because of non-accurate scoring of REM stage, auto analysis did not score REM stage-related hypopnea or apnea events appropriately. Automated analysis was just better at detecting hypopneas in NREM stage and accordingly total number of apneas and hypopneas was not acceptable. However, in Pittman *et al.* study about the accuracy of Morpheus I Sleep Scoring System, the results of Respiratory Disturbance Index (RDI) which included Apnea-Hypopnea Index (AHI) had agreement in both manual and auto analysis scoring among patients with moderate sleep apnea (12).

Furthermore, in Sangal et al. study on SANDMAN system, percentages of agreement and Cohen's kappa coefficients between all scorers for sleep staging, arousals, premature ventricular contractions (PVCs), and respiratory events in sleep were statistically significant. The ratios of computer-human agreement descriptors to human-human agreement descriptors indicated that computerized analysis of abnormal human sleep offers reasonable results with savings in technician's time and work, but not in physician's time and work (13).

Another study in 2013 on comparing the results between expert technicians in different labs and automated scoring system demonstrated that the agreement between the results of the current automated algorithm and the average of 10 expert scorers was comparable to the agreement between two expert scorers in the same site and similar to or better than the agreement between expert scorers across sites. This finding applies to sleep staging and scoring of arousals and respiratory events (3).

This study has several limitations. Using larger sample size may bring more reliable results regarding superiority of manual analysis. Furthermore, we used old version of the automated analysis software. More updated algorithms may have more concordance with manual analysis. Investigation of different automated analysis software programs may also lead to more exploration of reliable parameters of auto analysis.

Conclusion

It seems that in our study, auto analysis had significant differences to manual analysis, especially for staging EEG and respiratory events including the ones occurring in REM stage.

Auto analysis is not recommended to be used as the only method for PSG analysis; however, this technique may be helpful to determine hypopnea events in NREM stage and then be completed with manual scoring. Furthermore, updated algorithms according to available AASM guidelines also would be helpful and are recommended to be more elucidated and investigated. Such different results between studies may come from the variability of different software programs designed for auto analysis of PSG which warrants development of updated and more intelligent algorithms. Manual analysis by the trained sleep technicians and sleep specialists is still the best way for analysis of sleep tests recorded by PSG.

Conflict of Interests

Authors have no conflict of interests.

Acknowledgments

Authors would like to thank Dr. Samaneh Akbarpour, epidemiologist of Occupational Sleep Research Center, for revising data analysis results.

References

1. Rao VSK, Narender M, Kumar AS, et al. Comparison of manual scoring with auto generated polysomnographic scoring in obstructive sleep apnoea (OSA) patients. J Evol Med Dent Sci 2015; 4: 13070-6.
2. Magalang UJ, Chen NH, Cistulli PA, et al. Agreement in the scoring of respiratory events and sleep among international sleep centers. Sleep 2013; 36: 591-6.
3. Malhotra A, Younes M, Kuna ST, et al. Performance of an automated polysomnography scoring system versus computer-assisted manual scoring. Sleep 2013; 36: 573-82.
4. Penzel T, Hirshkowitz M, Harsh J, et al. Digital analysis and technical specifications. J Clin Sleep Med 2007; 3: 109-20.

5. Ozturk O, Mutlu LC, Sagcan G, et al. The concordance of manuel (visual) scoring and automatic analysis in sleep staging. Tuberk Toraks 2009; 57: 306-13. [In Turkish].
6. Stege G, Vos PJ, Dekhuijzen PN, et al. Manual vs. automated analysis of polysomnographic recordings in patients with chronic obstructive pulmonary disease. Sleep Breath 2013; 17: 533-9.
7. Barreiro B, Badosa G, Quintana S, et al. Comparison between automatic and manual analysis in the diagnosis of obstructive sleep apnea-hypopnea syndrome. Arch Bronconeumol 2003; 39: 544-8. [In Spanish].
8. Stepnowsky C, Levendowski D, Popovic D, et al. Scoring accuracy of automated sleep staging from a bipolar electroocular recording compared to manual scoring by multiple raters. Sleep Med 2013; 14: 1199-207.
9. Younes M, Hanly PJ. Minimizing interrater variability in staging sleep by use of computer-derived features. J Clin Sleep Med 2016; 12: 1347-56.
10. Berry RB, Budhiraja R, Gottlieb DJ, et al. Rules for scoring respiratory events in sleep: Update of the 2007 AASM Manual for the Scoring of Sleep and Associated Events. Deliberations of the Sleep Apnea Definitions Task Force of the American Academy of Sleep Medicine. J Clin Sleep Med 2012; 8: 597-619.
11. Bazil CW, Walczak TS. Effects of sleep and sleep stage on epileptic and nonepileptic seizures. Epilepsia 1997; 38: 56-62.
12. 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: 1394-403.
13. Sangal RB, Semery JP, Belisle CL. Computerized scoring of abnormal human sleep: A validation. Clin Electroencephalogr 1997; 28: 64-7.