The Seoul Journal of Medicine
Vol. 34, No. 4: 295-300, December 1993

The Staging of Sleep Apneics' Sleep: A Comparison of Computerized Analysis with Human Scoring†

Do-Un Jeong

Department of Psychiatry, Seoul National University College of Medicine, Seoul 110-744, Korea

= Abstract = **Computerized sleep staging is maintained to be a possible replacement for human scoring of sleep and is expected to reduce substantially the sleep technologists' time and efforts in scoring sleep into stages and practically to provide paperless polysomnography. So far various computer algorithms have been developed and automatic sleep analyzing systems have been tested mainly on normal human subjects and are alleged to be substantially reliable. However, it still remains to be answered how capable the systems are of reliably diagnosing sleep disordered subjects. The author attempted to review the function of a major automatic sleep analyzer (Oxford Medilog SAC 847 system) in comparison with expert human scoring. In eleven sleep apneic patients, one full night record of nocturnal polysomnography was compared between automatic analysis and human scoring on sleep stages epoch by epoch and on overall sleep architecture including various sleep parameters. The automatic analysis produced fewer stages 1 and REM, and more stage wake. The automatic stager's major difficulty was found to be with identifying wake and REM stages correctly. In conclusion, the present state of sophistication in automatic sleep analysis remains to be tested further in clinical sleep medicine.**

Key Words: *Automatic analysis, Human scoring, Polysomnography, Sleep apnea*

INTRODUCTION

It is always a tedious routine for sleep clinicians and researchers to score overnight polysomnograms epoch by epoch. The idea of having an automatic(computerized) sleep analyzer is very attractive and would be essentially more than an epoch-making development in the field of sleep medicine. Previous studies regarding automatic sleep analysis have concentrated mostly on normal volunteers without sleep disorders, with some studies arguing for the substantial reliability of automatic analysis(Hasan 1983;Kupfer *et al.* 1984; Crawford 1986; Flooh *et al.* 1986; Haustein *et al.* 1986; Höller and Riemer 1986; Matsuoka *et al.* 1986; Stanus *et al.* 1987; Kuwahara *et al.* 1988; Ferri *et al.* 1989; Kubicki *et al.* 1989). However, the findings do not automatically guarantee reliability in the sleep disordered population, because sleep disorders, by definition, affect the sleep structure and possibly add technical sophistication in terms of scoring algorithms. Therefore, automatic scoring systems need to be tested on sleep disordered subjects as well.

The author attempted in this study to clarify the power and the reliability of a major

Received November 1993, and in a final form December 1993.

† This study was supported by a Seoul National University Hospital Research Grant(02-92-140) to Do-Un Jeong.

Author for Correspondence : FAX. 82-2-745-8998
서울대학교 의과대학 정신과학교실 : 정도언

automatic system in analyzing the sleep of patients with sleep apnea syndrome.

SUBJECTS AND METHODS

Overnight polysomnograms(mean ± SE duration of 433.1 ± 10.4 minutes, range 385.5 - 488.5 minutes) of 10 men and 1 woman(mean ± SE 45.8 ± 3.3 years of age, range 31-65) with a clinical diagnosis of sleep apnea syndrome were obtained. According to the international 10-20 system(Jasper 1958), standard EEG electrodes were placed on C3-A2 and O2-A1. Two EOG's, chin EMGs, leg EMGs, 1 ECG, airflow, respiratory efforts(chest and abdomen), oximetry, snoring microphone, and body position sensor were set as indicated in the operator's manual(Oxford Instruments U. K. 1991). Gains and calibration signals were set according to the manual. The data were recorded and analyzed automatically with the Oxford Medilog SAC 847 system(ver. 7.9 Oxford Medical, England). All the raw data stored on hard disk were transferred to optical disks for further analyses.

Each polysomnogram was scored twice, once by automatic analysis algorithms contained in the SAC system and once by human scorers according to Rechtschaffen and Kales' criteria(1968). The human scoring was first done by a research assistant and the optimization of the results was done between the primary human scorer and the author, the board-certified sleep medicine specialist. Statistical analyses of the results mainly consisted of pairwise comparison of automatic analysis and human scoring. SPSS/PC ver. 4.0 was used for data analyses. All the values are expressed as mean ± SE. Significance level of 0.1 was adopted for Wilcoxon matched-pairs signed-ranks test.

RESULTS

The subjects retired to bed between 09:24 pm and 11:15pm and arose between 04:31am and 06:51am. Average length of sleep recording was 433.1 ± 10.4 minutes(range 385.5-488.5 minutes) with an average sleep period time of 422.5 ± 13.5 minutes(range 347.5-486.5 minutes).

The difference of sleep stage scoring between automatic analysis and human scoring is as shown in Table 1. Stage wake showed the highest difference of 134 epochs out of an average 866 epochs(15.5%) between automatic analysis and human scoring, with the automatic stager producing more stage wake (232.0% of human-scored stage wake)(Wilcoxon, $Z=-2.93$, $p=0.003$). Stage 1 comparison also revealed that, in automatic analysis, 59 epochs of human-scored stage 1 sleep(36.2%) were staged differently(Wilcoxon, $Z=-2.76$, $p=0.006$). Differences in REM sleep scoring were also very significant with automatic analysis producing 43.5 epochs(43.9%) less compared with human

Table 1. Comparison of sleep stages between automatic analysis and human scoring

Stages	Automatic analysis	Human scoring	Z	p
	Epochs(percent) (mean ± SE)	Epochs(percent) (mean ± SE)		
Wake	235.7 ± 47.0(28.1 ± 6.0)	101.6 ± 25.8(12.1 ± 3.1)	-2.93	0.003**
1	104.1 ± 17.1(11.6 ± 1.8)	163.1 ± 18.7(18.8 ± 2.2)	-2.76	0.006**
2	429.5 ± 40.5(49.1 ± 4.1)	467.5 ± 23.5(53.8 ± 2.0)	-1.51	0.131
3	24.9 ± 12.4(2.9 ± 1.4)	29.5 ± 9.9(3.4 ± 1.1)	-0.94	0.345
4	16.4 ± 14.7(1.9 ± 1.7)	5.5 ± 3.8(0.6 ± 0.4)	0.00	1.000
REM	55.6 ± 12.6(6.2 ± 1.4)	99.1 ± 15.5(11.3 ± 1.7)	-2.93	0.003**
Total	866.3 ± 20.8(100.0)	866.3 ± 20.8(100.0)		

Comparison by Wilcoxon matched-pairs signed-ranks test.
** $p<0.05$

Table 2. Comparison of automatic analysis and human scoring on sleep parameters

Sleep parameters	Automatic analysis (mean ± SE)	Human scoring (mean ± SE)	Z	p
SPT	405.4 ± 18.9	422.5 ± 13.5	-2.49	0.013**
TST	315.0 ± 30.3	382.0 ± 19.3	-2.93	0.003**
SL	21.0 ± 10.3	7.3 ± 3.0	-2.49	0.013**
RL	164.7 ± 36.2	124.0 ± 19.3	-0.18	0.859
IW	90.4 ± 20.3	40.5 ± 12.0	-2.93	0.003**
%WT	24.7 ± 5.5	10.6 ± 3.0	-2.93	0.003**
%SE	76.7 ± 5.5	90.2 ± 2.9	-2.93	0.003**
STCHSPT	132.0 ± 10.1	205.5 ± 26.0	-2.58	0.010**

SPT: sleep period time
TST: total sleep time
SL : sleep latency
RL : sleep onset REM sleep latency
IW : intermittent awake
%WT: percentage wake time
%SE: sleep efficiency %
STCHSPT: frequency of sleep stage change during SPT
All values in minutes except %WT, %SE, and STCHSPT.
Comparison by Wilcoxon matched-pairs signed-ranks test.
** $p<0.05$

scoring(Wilcoxon, $Z=-2.93$, $p=0.003$). Overall, the automatic analysis produced more stage wake, and fewer stages 1 and REM.

Sleep parameters characterizing sleep architecture also revealed significant differences between automatic analysis and human scoring(Table 2). Total sleep time was significantly less in the automatic analysis(Wilcoxon, $Z=-2.93$, $p=0.003$). Sleep latency was longer in the automatic analysis(Wilcoxon, $Z=-2.49$, $p=0.013$). Sleep efficiency was profoundly less in the automatic analysis(Wilcoxon, $Z=-2.93$, $p=0.003$). Frequency of sleep stage changes during sleep period was more conservatively counted in the automatic analysis(Wilcoxon, $Z=-2.58$, $p=0.010$).

Further analysis of sleep architecture with special reference to various latencies, as shown in Table 3, revealed that latencies to stages 1 and REM from record start time were significantly longer in automatic analysis compared with human scoring(Wilcoxon, $Z=-2.80$, $p=0.005$; $Z=-2.20$, $p=0.028$). Latencies to stages 1 and 2 from sleep onset were also found to be significantly longer in automatic analysis compared with human scoring(Wilcoxon, $Z=-2.52$, $p=0.012$; $Z=-1.78$, $p=0.075$).

The discrepant allocation of sleep stages between automatic analysis and human scoring is as shown in Table 4. Generally, more than 50% of stages 1 and REM epochs were allocated differently to other stages by the automatic analysis. Almost all of the stage 1 epochs scored differently(97.3%) were mistaken as stage wake or 2. Automatic analysis tended to have stage REM scored as stage wake or 1. The majority, 54.2 epochs out of 88. 0 differently scored stage 2 epochs, were allocated to stage wake. Allocation difference of delta sleep, particularly stage 4, was not remarkable. Discordantly scored stage wake made up 54.8% of the total sum of incorrectly scored epochs. Gamma values computed for overall concordance of sleep stages between automatic analysis and visual scoring ranged from 0.074 to 0.990 with means (± SE) of 0.680 (± 0.084).

Correlation of the total number of differently staged epochs for human scored stage 1 sleep with the total number of respiratory disturbances(RD) and RD index(defined as the number of RD per hour of sleep period time) revealed significant correlations($p=0.021$, $p=0.016$, respectively). Other correlations were not significant.

Table 3. Sleep architecture detected differently by automatic analysis vs. human scoring with special reference to various latencies

Latencies	Automatic analysis (mean ± SE)	Human scoring (mean ± SE)	Z	p
LS1	41.3 ± 13.6	6.1 ± 2.8	-2.80	0.005*
LS2	9.3 ± 3.1	10.4 ± 3.0	-0.93	0.353
LS3	117.9 ± 56.1	45.1 ± 12.8	-1.21	0.225
LS4	21.2 ± 4.8	23.0 ± 3.0	-1.00	0.317
LS5	186.4 ± 34.0	131.4 ± 20.0	-2.20	0.028*
SL	21.0 ± 10.3	7.3 ± 3.0	-2.49	0.013*
SOLS1	21.5 ± 12.0	0.1 ± 0.1	-2.52	0.012*
SOLS2	1.9 ± 1.7	3.0 ± 1.2	-1.78	0.075*
SOLS3	110.1 ± 56.2	40.9 ± 14.1	-0.94	0.345
SOLS4	12.3 ± 1.8	15.7 ± 0.9	-1.60	0.109
SOLS5	164.7 ± 36.2	124.0 ± 19.3	-0.18	0.859

LS1: latency to stage 1 LS2: latency to stage 2
LS3: latency to stage 3 LS4: latency to stage 4
LS5: latency to stage REM
Latency to stages 1-5 means the amount of time in minutes between record start time and the first epoch of each stage.
SL: sleep latency
SOLS1: sleep onset latency to stage 1 SOLS2: sleep onset latency to stage 2
SOLS3: sleep onset latency to stage 3 SOLS4: sleep onset latency to stage 4
SOLS5: sleep onset latency to stage REM
Sleep onset latency to stages 1-5 means the amount of time in minutes between sleep onset and the first epoch of each stage.
Comparison by Wilcoxon matched-pairs signed-ranks test.
* $p<0.1$

Table 4. Differently scored sleep stages in automatic analysis matched with human-scored epochs

Stages	No. Epochs (Human)	No. Discordant epochs(±SE) (Human - SAC)	No. Epochs allocated differently to each sleep stage					
			Wake	1	2	3	4	REM
Wake	101.6	3.2(± 1.0)	-	1.3	1.4	0	0	0.5
1	163.1	94.2(±15.8)	52.1	-	39.6	0.4	0	2.1
2	467.5	88.0(±28.7)	54.2	16.5	-	10.9	3.6	2.8
3	29.5	16.1(± 7.5)	3.3	0.1	5.2	-	7.5	0
4	5.5	0.3(± 0.2)	0.1	0	0	0.2	-	0
REM	99.1	48.9(±11.9)	27.6	17.4	3.9	0	0	-
Total	866.3	250.6(±42.9)	137.3	35.3	50.1	11.5	11.1	5.4

All values are mean values from 11 subjects.
SAC: sleep analyzing computer
SE : standard error

DISCUSSION

It is very time-consuming to score overnight polysomnograms epoch by epoch according to the Rechtschaffen and Kales' criteria(1968). Especially considering the repetitive pattern of sleep and arousal/awakening observed in sleep apnea patients(Jeong 1993), it would be very helpful if an automatic sleep analyzer could detect the sleep stages correctly in the midst of the repetitive influence of respiratory events on sleep architecture progression.

However, the present level of technical sophistication does not seem to meet with expectations. The automatic stagers use relatively simple algorithms in sleep scoring in comparison with human eyes. They do not permit strict adherence to the criteria by Rechtschaffen and Kales(1968) based on pattern recognition. Having much less capacity for recognizing patterns even in the systems maintained to have some capacity for it, subtle changes such as miniarousals cannot be detected. Especially in pathologic sleep states such as sleep apnea syndrome, the algorithms adopted may be too simple and rigid for scoring epochs from the point of pattern recognition.

Comparison of the automatic analysis with expert human scoring in this study made it clear that scoring of wake stages(before falling asleep as well as during the night) was one of the main problems for automatic analysis. The much higher production of stage wake and percentage wake time(see Table 2) by automatic stager seem to be mainly due to the difficulty in distinguishing with automatic scoring algorithms among stages wake, 1, and REM and in scoring epochs during stage 2 without sleep spindles and K-complexes. Miniarousals induced by sleep apneas repetitively arousing the subject can be a great problem in sleep staging. The automatic stager does not seem to look at the whole picture as human eyes do.

Reliable detection of REM sleep is another issue to be discussed. The majority of stage REM epochs was scored as stage wake, possibly because of alpha background activity in stage REM. Considerable epochs were scored as stage 1, probably due to missing REMS. Some were mistaken as stage 2 epochs, possibly because of interspersed sleep spindles or relatively stable fast alpha activity(Kubicki *et al.* 1989).

Delayed identification of REM sleep onset by the automatic stager is a problem. Especially in diagnosing sleep disorders such as narcolepsy and major depression, for which sleep-onset REM period is a critical diagnostic clue (Guilleminault 1989; Walsh and Sugerman 1989), false diagnoses could be reached if clinicians relied only on automatic analyses.

Besides nocturnal polysomnography, the assessment of daytime sleepiness using multiple sleep latency test(MSLT)(Carskadon 1989) or maintenance of wakefulness test(MWT)(Mitler *et al.* 1982) is now a common practice in sleep clinics. Considering that sleep and REM latencies are counted in the procedures, it is obvious that the assessment cannot be made relying exclusively on automatic analyzers.

Comparing the results of this study with the study by Kubicki *et al.* (1989), it is interesting to see that there are some differences. In the study by Kubicki *et al.* (1989), the automatic stager produced a decrease in stages wake, REM, and 2 and an increase in stages 1, 3, and 4, while in the author's study fewer stages 1 and REM and more stage wake were scored by automatic analysis. The discrepancies seem to be at least partially attributable to the difference in study subjects between the two studies - normal volunteers and sleep disordered patients(Hasan 1983; Kupfer *et al.* 1984).

Automatic sleep analysis could be a promising and time-saving tool for sleep quality and quantity evaluation in the future. The Medilog SAC system seems to be a possibility, but it remains to be tested further like other systems on the market. There are still issues to be solved(Hasan 1985) as well as too high a number of discrepancies from human scoring for it to be acceptable clinically(Roffwarg 1990).

Updated and improved versions of the system need to be tested in the future.

REFERENCES

Carskadon M. Measuring daytime sleepiness. In Kryger MH, Roth T, Dement WC(Eds). Principles and Practice of Sleep Medicine, WB Saunders Co, Philadelphia, 1989: pp. 684-8

Crawford C. Sleep recording in the home with automatic analysis of results. Eur Neurol 1986; 25(suppl 2):30-5

Ferri R, Ferri P, Colognola RM, Petrella MA, Musumeci SA, Bergonzi P. Comparison between the results of an automatic and a visual scoring of sleep EEG recordings. Sleep 1989; 12:354-62

Flooh E, Korner E, Lechner H. Computer evaluation of sleep. Eur Neurol 1986; 25(suppl 2) :46-52

Guilleminault C. Narcolepsy syndrome. In Kryger MH, Roth T, Dement WC(Eds). Principles and Practice of Sleep Medicine, WB Saunders Co, Philadelphia, 1989: pp. 340-1

Hasan J. Differentiation of normal and disturbed sleep by automatic analysis. Acta Physiol Scand Suppl 1983; 526:1-103

Hasan J. Automatic analysis of sleep recordings: a critical review. Ann Clin Res 1985; 17:280-7

Haustein W, Pilcher J, Klink J, Schulz H. Automatic analysis overcomes limitations of sleep stage scoring. Electroencephalogr Clin Neurophysiol 1986; 64:364-74

Höller L, Riemer H. Comparison of visual analysis and automatic sleep stage scoring (Oxford Medilog 9000 System). Eur Neurol 1986; 25(suppl 2):36-45

Jeong DU. Sleep architecture analysis in obstructive sleep apnea syndrome. Seoul J Psychiatry 1993; 18:1-11

Jasper HH(Committee Chairman). The ten twenty electrode system of the International Federation. Electroencephalogr Clin Neurophysiol 1958; 10:371-5

Kubicki St, 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; 12:140-9

Kupfer DJ, Ulrich RF, Coble PA, Jarrett DB, Grochocinski V, Doman J, Matthews G, Borbely AA. Application of automated REM and slow wave sleep analysis: I. Normal and depressed subjects. Psychiatry Res 1984; 13:325-34

Kuwahara H, Higashi H, Mizuki Y, Matsunari S, Tanaka M, Inanaga K. Automatic real-time analysis of human sleep stages by an interval histogram method. Electroencephalogr Clin Neurophysiol 1988; 70:220-9

Matsuoka S, Ishikawak T, Inoue K, Hatashi A. Automatic determination system of human sleep stages on an experimental basis. Sangyo Ika Daigaku Zasshi 1986; 8(suppl) :169-71

Mitler MM, Gujavarty KS, Sampson MG, Brownman CP. Multiple daytime nap approaches to evaluating the sleepy patients. Sleep 1982; 5(suppl 2): 5119-27

Oxford Instruments plc. Medilog SAC operator's manual. Oxon, England, 1991

Rechtschaffen A, Kales A, eds. A manual of standardized terminology, technique and scoring system for sleep stages of human subjects. Los Angeles: Brain Information Service/Brain Research Institute, University of California at Los Angeles, 1968

Roffwarg HP. ASDA position statement: automatic scoring. Sleep 1990; 13:284-5

Stanus E, Lacroix B, Kerkhofs M, Mendlewicz J. Automated sleep scoring: a comparative reliability study of two algorithms. Electroencephalogr Clin Neurophysiol 1987; 66:448-56

Walsh JK, Sugerman JL. Disorders of initiating and maintaining sleep in adult psychiatric disorders. In Kryger MH, Roth T, Dement WC (Eds). Principles and Practice of Sleep Medicine, WB Saunders Co, Philadelphia, 1989: p. 449