Sleep
12(4):354–362, Raven Press, Ltd., New York
© 1989 Association of Professional Sleep Societies

Comparison Between the Results of an Automatic and a Visual Scoring of Sleep EEG Recordings

R. Ferri, P. Ferri, R. M. Colognola, M. A. Petrella, S. A. Musumeci, and *P. Bergonzi

*Oasi Institute for Research and Prevention of Mental Retardation, Troina, and *Cattedra di Neuropatologia e Psicopatologia, University of Cagliari, Cagliari, Italy*

Summary: In this paper, results from the visual scoring of nocturnal polygraphic recordings, carried out by nine different groups of readers from different Italian sleep laboratories, are analyzed; inter- and intragroup variability is shown and statistically discussed. Data are then compared with the results of an automatic scoring of the same recordings, carried out by the Medilog Sleep Stager. The validity of this automatic method of scoring is discussed. Finally, an epoch by epoch analysis is described, with the aim of achieving a more detailed evaluation of the intergroup variability. **Key Words:** Sleep scoring—Automatic analysis—Methodology—Scoring variability.

Downloaded from https://academic.oup.com/sleep/article/12/4/354/2742688 by guest on 21 August 2022

The polygraphic recording of nocturnal sleep is a method of research widely used in neurophysiology laboratories, both for the clinical study of sleep and for the evaluation of the therapeutic effectiveness of drugs acting on sleep.

Two important points must be considered, however:

(a) Rules for visual scoring of sleep recordings to achieve a greater homogeneity of results between different scorers must be better defined.

(b) Costs and time spent for sleep recording and scoring must be decreased.

Devices simpler than those currently used can be helpful in overcoming these problems; they can allow a greater number of recordings to be carried out, in more "physiological" conditions (at home, for example), together with faster analysis by means of automatic methods of scoring.

The Oxford Medilog 9000 System (Oxford Medical Systems, Abingdon, England) has been recently introduced to the market with the aim of overcoming some of the previous problems; several studies have already been completed in order to test its reliability (1–3), but results have frequently been conflicting.

Accepted for publication February 1989.

Address correspondence and reprint requests to Dr. R. Ferri, Dept. of Neurology, Oasi Institute, 94018, Troina (EN), Italy.

Above all, it seems that the subjectiveness of visual scoring has been underevaluated in these studies, even if the visual analysis was used as a reference for statistical comparison with the automatic scoring.

For these reasons we carried out the study presented in this paper, which compared automatic and visual analysis of nine different groups of readers.

MATERIALS AND METHODS

A group of four healthy volunteers, three women (age range 26–32 years) and one man (age 15 years), was used. Sleep was recorded by a Medilog 8 channel recorder, and subjects slept at their usual time until spontaneous awakening.

Three electroencephalogram (EEG) channels (A2-Fp1, A2-C3, and A2-O1), two electrooculogram (EOG) channels (A1, 1 cm vertically upward from outer canthus of left eye; A1, 1 cm vertically downward from outer canthus of right eye), and one electromyogram (EMG) channel (two electrodes placed on the jawbone) were used (Fig. 1).

The signals were reproduced on paper (by a Mingograph EEG 21 polygraph, Siemens, Solna, Sweden) for the visual scoring. Recordings were numbered from 1 to 4 and then divided into epochs of 40 s. Recording 4 was reproduced on paper twice; the second reproduction was numbered recording 5 and used for the evaluation of the repeatability of visual scoring.

The recordings reproduced were of good quality and did not differ significantly from those directly recorded on paper; note that the Medilog 9000 system uses a sampling rate of 128 Hz, which allows a good definition of EEG frequencies.

The sleep recordings were also analyzed by the Oxford Medilog 9000, with its automatic Sleep Stager (Oxford Medical Systems), for comparison with the visual scoring. In this phase we paid special attention to the starting time, because it was technically impossible to automatically control it, both in automatic analysis and in paper reproduction. Cassette 4 was analyzed twice by the Sleep Stager.

The same settings were used for all automatic analyses.

The five recordings were scored independently by nine different groups, eight composed of two specialists and one of a specialist working alone. These specialists came from different universities in Italy: Drs. Puca and Brancasi from Bari; Drs. Cirignotta and Zucconi from Bologna; Drs. Ferrillo and Franconieri from Genova; Drs. Di Perri and Silvestri from Messina; Drs. Smirne and Ferini-Strambi from Milan; Drs. Terzano and Parrino from Parma; Drs. Murri and Massetani from Pisa; Drs. Guazzelli and Ciapparelli from Pisa; and Dr. Mennuni from Rome.

The groups scored the recordings by the Rechtschaffen and Kales (4) rules, and they were unaware that recordings 4 and 5 were the same. The order of recording analysis was randomly assessed for each group of readers.

Each group developed a printout for each recording, reporting the stage scoring sequentially with the time duration of each stage [S0, S1, S2, S3, S4, and rapid eye

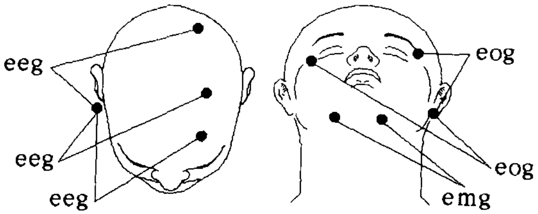

FIG. 1. Electrode placements. See text for details. EEG, electroencephalogram; EOG, electrooculogram; EMG, electromyogram.

Downloaded from https://academic.oup.com/sleep/article/12/4/354/2742689 by guest on 21 August 2022

movement (REM)] in number of 40 s epochs. The results of the 45 visual scorings (nine groups five recordings) and of the five automatic scorings (Sleep Stager) were used as data for further evaluations by a computer program prepared for this purpose, running on an MS-DOS personal computer (Sperry, U.S.A.). The following parameters were considered:

SOL (sleep onset latency): measured from the beginning of 2 min continuous sleep (S1, S2, S3, S4, or REM)
SPT (sleep period time): sleep onset to sleep end
TST (total sleep time): SPT − S0
FRL (first REM latency): sleep onset to 2 continuous min REM stage
SSh (stage shifts): number of stage shifts
Awn (awakenings number): number of awakenings after sleep onset
S0 (stage 0): wake after sleep onset + movements
S1 (stage 1)
S2 (stage 2)
S3 (stage 3)
S4 (stage 4)
REM (stage REM)

Note that these parameters were chosen according to the Sleep Stager manual (5), in order to compare the visual and automatic analyses statistically. The computer program allowed us to evaluate S0, S1, S2, S3, S4, and REM either in number of epochs, in minutes, or in percentage of SPT.

The data resulting from this analysis were organized in a tridimensional array with the following dimensions: parameter, recording, and group.

The computer program allowed us to see data organized in three different types of tables: (a) how the same parameter was scored by the readers and by the Sleep Stager in the different recordings (Table 1); (b) how the same recording, with its different

Downloaded from https://academic.oup.com/sleep/article/12/4/354/2742689 by guest on 21 August 2022

TABLE 1. *One example of visual and automatic sleep scoring data (values of SOL in min)*

	Sleep recording							M vs. A–I	
Group	1	2	3	4	5	Mean	SD	*t* test	*df*
A	4.00	10.67	8.67	12.00	12.00	8.83	3.50		
B	4.00	23.33	9.33	0.67	12.00	9.33	9.99		
C	4.00	22.67	7.33	0.67	0.67	8.67	9.72		
D	4.00	10.00	9.33	0.67	12.00	6.00	4.46		
E	4.00	25.33	9.33	12.00	12.00	12.67	9.08		
F	3.33	4.67	6.67	2.67	2.67	4.33	1.76		
G	4.00	23.33	9.33	12.00	12.00	12.17	8.15		
H	4.00	6.67	7.33	0.67	0.67	4.67	3.03		
I	4.00	6.67	9.33	12.00	12.00	8.00	3.44		
Mean	3.93	14.81	8.52	5.93	8.44	8.30	–		
SD	0.22	8.61	1.09	5.80	5.36	–	6.50		
M[a]	4.00	1.33	8.67	0.67	0.00	3.67	3.63	−2.190	38

Analysis of variance	Sum of sq.	*df*	Variance	
Between groups	279.28	8	34.91	*F*: 0.787
Within groups	1198.44	27	44.39	

[a] Sleep stager scoring.

TABLE 2. *One example of dissimilarity test of Gini data (from the automatic analysis)*

	Sleep recording					Dissimilarity test of Gini
Parameter[a]	1	2	3	4	5	
SOL (min)	4.00	1.33	8.67	0.67	0.00	0.038
SPT (min)	512.67	528.67	514.67	512.00	512.67	
TST (min)	487.33	484.00	491.33	498.00	498.67	
FRL (min)	70.00	0.00	212.00	0.00	0.00	
SSh (no.)	144.00	99.00	145.00	272.00	297.00	
Awn (no.)	28.00	9.00	19.00	15.00	15.00	
S0 (%)	4.94	8.45	4.53	2.73	2.73	
S1 (%)	1.56	5.42	4.27	4.17	3.51	
S2 (%)	71.52	51.58	52.33	29.04	26.79	
S3 (%)	4.16	8.32	9.07	15.10	14.17	
S4 (%)	0.13	2.14	11.14	5.21	6.89	
REM (%)	17.69	24.09	18.65	43.75	45.90	

[a] SOL, sleep onset latency; SPT, sleep period time; TST, total sleep time; FRL, first REM latency; SSh, stage shifts; Awn, awakening number; S0–S4, stage 0–4.

parameters, was scored by the readers; and (c) how the same group of readers scored the five different recordings (Table 2).

Statistical analysis was carried out using Student's *t* test for comparison between the Sleep Stager and readers and analysis of variance for comparison between groups of readers (6). For evaluation of repeatability, the dissimilarity test of Gini (7) was used.

Epoch by epoch analysis

Data from visual scoring, stored on hard disk, were automatically transformed, for each recording, into an epoch sequence. Each epoch could have one of the following values: W (wake), 1 (S1), 2 (S2), S (S3 or S4), and R (REM). We then compared the nine different sequences relating to the same recording, and we obtained a 10th sequence containing the most frequent value for each epoch. When it was not possible to obtain a unique value more frequently scored for an epoch, this epoch was classified as "dubious," and it was indicated by an asterisk (Fig. 2).

The epoch sequence obtained from this analysis was called "consensus," and we used it for automatic compilation of tables describing the percentages of agreement between each group of readers and the consensus. In this way, six tables relating to the five recordings and to their total were obtained for each group.

Table 3 is an example. In the first column the group scoring is indicated, while in the top row the consensus scoring is reported. Crossings between rows and columns indicate the number of epochs correctly classified by the group in comparison with the consensus scoring. Other boxes contain the number of incorrectly scored epochs. The last column indicates the percentage of agreement for each sleep stage and for all sleep stages.

It was not possible to carry out the epoch by epoch comparison between the consensus and the computer scoring because the Sleep Stager does not indicate the starting and end points of each epoch, although it is possible to insert such an indication by a technical modification of the system. In this study the system was not modified because the marketed configuration needed to be tested, and, in addition, since the automatic analysis is carried out by the Sleep Stager with a speed 20 times higher than the real time, a major modification of the writer (Mingograf EEG 21) would have been required.

Downloaded from https://academic.oup.com/sleep/article/12/4/354/2742689 by guest on 21 August 2022

```
148 S S S S 2 S S S S - S
149 S S S S 2 S S S S - S
150 S 2 S S 2 S S S S - S
151 S 2 2 2 2 2 2 S 2 - 2
152 S 2 2 2 2 2 2 S 2 - 2
153 2 2 2 2 W W 2 S 1 - 2
154 2 2 2 2 2 R 2 R 2 - 2
155 2 2 2 2 2 R 2 R 2 - 2
156 2 R R 2 R R R R R - R
157 2 R R 2 R R R R R - R
158 2 R R 2 R R R R R - R
159 R 2 R 2 2 R R R R - R
160 R R R R R R R R R - R
161 R R R R R R R R R - R
162 R R R R R R R R R - R
163 R R R R R R R R R - R
164 R 1 R 1 1 R W R 1 - *
165 1 1 W 1 W 1 W W W - W
166 1 1 2 2 1 1 1 R 1 - 1
167 2 2 2 2 2 2 2 R 2 - 2
168 2 2 2 2 2 2 2 R 2 - 2
169 2 2 2 2 2 2 2 R 2 - 2
170 2 2 2 2 2 2 2 R 2 - 2
```

FIG. 2. Epoch by epoch analysis. The first column shows the epoch number, the nine middle columns the group scorings, and the last column the consensus scoring.

RESULTS

Table 1 gives an example of the data output of the computer program. This table refers to the SOL parameter. In the columns it is possible to see how the different groups evaluated the same recording, while in the rows it is possible to observe how the same group evaluated the different recordings. Columns 7 and 8 indicate the row means and standard deviations of scorings from recordings 1 to 4, 5 being a repetition of 4.

Column means and standard deviations indicate the mean scoring of groups. The mean and SD rows give the total means and standard deviations. These last values are

TABLE 3. *One example of the agreement between group scoring and consensus (recording 1, scoring of group A)*

	Wake	Stage 1	Stage 2	St. 3/4	REM	Dubious	Total	Agreement (%)
Wake	100	1	0	0	0	0	101	90.9
Stage 1	7	45	9	0	0	3	64	88.2
Stage 2	3	5	305	0	4	2	319	96.8
Stage 3/4	0	0	1	119	0	0	120	100.0
REM	0	0	0	0	152	1	153	97.4
Total	110	51	315	119	156	6	757	96.0

Downloaded from https://academic.oup.com/sleep/article/12/4/354/2746688 by guest on 21 August 2022

needed for the statistical analysis between visual and automatic scoring (row M) by Student's *t* test.

The last three rows show the analysis of variance between visual scorings for the statistical evaluation of intergroup differences.

Sleep onset latency

Although the analysis of variance did not show significant differences among groups, it was possible to observe in recordings 4 and 5 a bimodal distribution of SOL scorings. In effect, some groups indicated 0.67 min (1 epoch) and others 12 min. Furthermore, two groups indicated 0.67 min in recording 4 and 12 min in recording 5. Of course, this indicates a true difficulty in scoring. The Sleep Stager agreed with 0.67 min. Generally the Sleep Stager scored SOL with values lower than those of readers ($p < 0.025$).

Downloaded from https://academic.oup.com/sleep/article/12/4/354/2742689 by guest on 21 August 2022

Sleep period time and total sleep time

There were no significant differences among groups, or between them and the Sleep Stager.

First REM latency

Statistical analysis did not show significant differences, but it was possible to note a bimodal scoring of FRL in recording 3, in which some groups indicated a value of about 45 min while others indicated about 208 min; it was not possible to establish which is the correct value, but a value of 45 min was more frequently scored (five groups out of nine). The Sleep Stager agreed with the value of 208 min. The Sleep Stager scored FRL of recordings 2, 4, and 5 as 0 min; this means that the Sleep Stager presented discrimination problems between wakefulness, S1, and REM.

Stage shifts

This parameter was evaluated differently by groups ($p < 0.01$) with a variability range up to 100%. The Sleep Stager showed the highest value, significantly different from the mean of groups ($p < 0.025$).

Awakenings number

The evaluation of this parameter produced results significantly different among groups ($p < 0.025$). The Sleep Stager showed no statistically different results from those of the groups; in fact, they were within one standard deviation from the mean of groups.

Stage 0

No differences were found among groups, nor between groups and the Sleep Stager.

Stage 1

Groups did not show significant differences in S1 scoring; the Sleep Stager showed a lower mean value ($p < 0.05$).

Stage 2

The S2 scoring was significantly different in different groups ($p < 0.01$), with a wide range of variation. No significant differences were found between group and Sleep Stager means.

Stage 3

S3 was differently scored by different groups ($p < 0.01$), while the Sleep Stager provided results not significantly different from the mean of groups.

Stage 4

The S4 scoring was also different among groups ($p < 0.05$). The Sleep Stager evaluation was lower than the group mean (4.66% versus 12.55%) ($p < 0.001$).

Stage REM

This parameter was homogeneously scored by groups with a narrow range of variation. There were no statistically significant differences between groups and Sleep Stager, although there was a great difference in recording 4 and, of course, 5.

Table 2 shows how the same group scored the five different recordings and the statistical test used for the evaluation of repeatability. In the lower right corner the dissimilarity index of Gini (7) is shown. This index can vary from 0 (in the case of identical distributions) up to 1 (in the case of different distributions) and has been applied to the distributions of stage percentages of recording 4 and 5 scored by the same group. The Sleep Stager showed an index of 0.038, while group indexes varied from 0.035 to 0.164, with a median of 0.084. Only one group index (0.035) was lower than that of the Sleep Stager.

Results of the epoch by epoch analysis

Because of the high number and size of all tables we report here only the main results of the epoch by epoch analysis (Fig. 2), relating to the group agreement percentages of all the recordings (Table 3 gives one example).

The agreement percentages relating to stage 0 varied from 47.5% to 100%; from 33.3% to 96.3% for stage 1; and from 39.9% to 99.1% for stage 2. The stage 3/4 (S) variation range was very wide, from 3% to 100%. However, it was due only to a scoring greatly differing from the mean; if we do not consider this scoring, the minimum value is 32.3%, that is, always the lower minimum observed value of agreement. The agreement percentages varied from 62.9% to 100% for REM sleep. Total values varied from 60.9% to 96%.

In conclusion, the lowest values were observed in slow sleep scoring, while the highest values were obtained in REM sleep scoring.

DISCUSSION

In this study, in spite of the use of a standardized method of scoring (4), different groups of readers show in some cases very different scorings of the same recording. On the other hand, the intragroup variability seems to be good and reliable.

The causes of the intergroup variability could be due both to a different application of standardized rules of scoring by different readers and to a real difficulty in scoring some sleep recordings by the means of the above rules; this is demonstrated by the bimodal distribution of the visual scoring of SOL and FRL. In these cases it is very difficult, if not impossible, to indicate the correct scoring. These difficulties have been discussed by Kubicki et al. (8).

Downloaded from https://academic.oup.com/sleep/article/12/4/354/2742689 by guest on 21 August 2022

We think that it is possible to indicate as acceptable an agreement of 80% with the consensus for all stages; in this work, seven groups of nine show values greater than 84%. Naturally this "threshold" is arbitrary and can be further discussed.

In this study the Sleep Stager shows difficulties in discriminating rapid stages of sleep from wakefulness and between them (S0, S1, and REM), while the scoring of the slow wave sleep (S3 and S4) could be affected by the fact that the automatic analysis recognized delta waves of amplitude higher than that of delta waves recognized by the readers, who pay more attention to the frequency than to the amplitude of waves. In fact, in a previous study (2), in which the readers rigidly applied a threshold of 75 μV for delta wave recognition, a lower amount of S3 and S4 was found in the visual analysis when compared with the automatic scoring. In the same study, SOL was also found to be shorter in the automatic analysis, but a different definition of it was used by the reader; on the other hand, the same authors found a lower amount of stage REM in the automatic analysis.

In the present study, the amount of REM sleep does not present any statistically significant difference in the comparison between automatic and visual analysis.

Different results were obtained in other studies: Kurosaki et al. (9) found a lower amount of stage 1 and of stage 3/4 in the automatic analysis; Kubicki et al. (3) found decreased stage REM, increased stage 1, reduced stage 2, and prolonged SOL in the Sleep Stager's analysis.

These different results point out the necessity of comparing the automatic analysis with results from a group of different readers and not only from 1 or 2 because of the possibility of different visual scoring.

The small differences seen in repeated automatic evaluation of the same recording are mainly due to some practical factors influencing the application of the method:

(a) When repeating the analysis, it was impossible to indicate to the Sleep Stager exactly the same starting and end points, because of technical reasons, and a difference of several seconds was possible.

(b) It was impossible to make a precise calibration of the amplitude of the signal using the available interface. (A new, more reliable interface is now available.)

This variability has also been described by Marsh and Erwin (10).

We also examined data from the automatic analysis of the same recordings using another software version (version 3.2), and we observed a partial correction of previous

TABLE 4. *Summary of results of visual scoring (Vis.), first automatic analysis (Old), and last automatic analysis by the new software (New)*

	Recording 1			Recording 2			Recording 3			Recording 4		
Parameter[a]	Vis.	Old	New	Vis.	Old	New	Vis.	Old	New	Vis.	Old	New
SOL (min)	3.93	4.00	4.0	14.81	1.33	2.0	8.52	8.67	7.5	5.93	0.67	0.5
SPT (min)	493.11	512.67	515.0	575.78	528.67	529.5	491.48	514.67	516.0	492.74	512.00	514.0
TST (min)	430.59	487.33	490.0	557.19	484.00	482.0	471.70	491.33	494.0	478.96	498.00	499.0
FRL (min)	68.44	70.00	69.5	88.74	0.00	0.0	128.89	212.00	51.0	65.41	0.00	21.5
SSh (no.)	88.56	144		99.56	99		87.22	145		81.67	272	
Awn (no.)	18	28	34	15.67	9	15	13.22	19	27	10.67	15	23
S0 (%)	12.68	4.94	4.9	3.23	8.45	9.0	4.02	4.53	4.3	2.79	2.73	2.9
S1 (%)	7.88	1.56	14.1	5.67	5.42	16.6	4.66	4.27	8.0	3.84	4.17	14.8
S2 (%)	42.85	71.52	59.0	51.91	51.58	48.0	46.46	52.33	49.7	44.09	29.04	33.3
S3 (%)	6.76	4.16	3.2	9.30	8.32	8.7	12.43	9.07	9.7	12.45	15.10	14.4
S4 (%)	9.56	0.13	0.0	8.38	2.14	3.1	16.06	11.14	11.4	16.21	5.21	9.1
REM (%)	20.27	17.69	18.8	21.50	24.09	14.6	16.36	18.65	16.9	20.21	43.75	25.6

[a] For abbreviations, see footnote *a*, Table 2.

Downloaded from https://academic.oup.com/sleep/article/12/4/354/2742689 by guest on 21 August 2022

mistakes, mainly in FRL measurement and REM scoring (Table 4). On the other hand, the previous software seems to be more reliable than the new one in S1 scoring (Table 4). The new software causes an increase in S1; it was updated according to studies in which a comparison between automatic and visual analysis of no more than one or two readers (of the same group) was carried out (1,2). Nevertheless, the new software could be considered more appropriate. The intergroup variability, already described in this paper, allows us to affirm that it is not possible to accept as significant comparisons between automatic analysis and only one or two readers.

Finally, we suggest that the Sleep Stager could be used for analyzing large numbers of recordings in research in which only an evaluation of the amount of the sleep stages is needed (taking into account the problems already discussed); it does not seem appropriate when data on sleep organization are important. This is clearly evidenced by the higher number of SSh and by the different evaluation of SOL and FRL in the automatic analysis.

Acknowledgments: We thank Ms. Pina Siciliano, Mr. Mariano Maggio, and Mr. Giuseppe Miano for their kind help. This work was supported by Novamedica S.p.A., Rome, Italy.

REFERENCES

1. Crawford C. Sleep recording in the home with automatic analysis of results. *Eur Neurol* 1986;25 (suppl 2):30–5.
2. 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.
3. Kubicki St, Höller L, Berg I, Dorow R. Validation of an automatic sleep analysis system in normal sleepers. Chart presented at the 5th International Congress of Sleep Research, Copenhagen, June 28–July 3, 1987 [Abstract], 650.
4. Rechtschaffen A, Kales A. *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.
5. *Sleep stager manual*. Oxford: Oxford Instrument Group, 1985.
6. Armitage P. *Statistical methods in medical research*. Oxford: Blackwell Scientific Publications, 1971.
7. Gini C. La dissomiglianza. *Metron* 1965;24(1–4).
8. Kubicki St, Hermann WM, Höller L. Critical comments on the rules by Rechtschaffen and Kales concerning the visual evaluation of EEG sleep records. In: Kubicki St, Hermann WM, eds. *Methods of sleep research*. Stuttgart: Gustav Fisher, 1985:19–35.
9. Kurosaki Y, Sasaki M, Okudaira N, Spinweber CL, Graeber RC. Comparison of automatic sleep stage scoring and visual analysis. Chart presented at the 5th International Congress of Sleep Research, Copenhagen, June 28–July 3, 1987 [Abstract], 651.
10. Marsh GR, Erwin CW. The Oxford Sleep Stager: assessment of variability. *J Clin Neurophysiol* 1987;4:291–2.

Downloaded from https://academic.oup.com/sleep/article/12/4/354/2742689 by guest on 21 August 2022