Journal of Neurology, Neurosurgery, and Psychiatry, 1979, **42**, 1163–1168

Interobserver agreement in assessment of ocular signs in coma

J. H. VAN DEN BERGE, H. J. A. SCHOUTEN, S. BOOMSTRA, S. VAN DRUNEN LITTEL, AND R. BRAAKMAN

From the Departments of Neurosurgery and Biostatistics, Academic Hospital Rotterdam Dijkzigt, Erasmus University, Rotterdam, The Netherlands

SUMMARY There is interobserver agreement in the assessment of various ocular signs found in coma patients. As measure for observer agreement the parameter kappa (K) was determined for (in-)equality of pupils, reaction of pupils, spontaneous eye movements, and oculocephalic responses. The agreement in the assessment of the pupils to light and in the assessment of (in-)equality of pupils appeared to be satisfactory, but more disagreement occurred in assessing spontaneous eye movements and oculocephalic responses.

Spontaneous and induced eye movements are of prognostic value in patients with severe head injuries (Mingrino *et al.*, 1965; Fisher, 1969; Poulsen and Zilstorff, 1972; Teasdale and Smith, 1975; Bricolo *et al.*, 1977).

The prognostic value of various clinical observations is currently being assessed in an international study of prognosis of severe head injury (Jennett *et al.*, 1976, 1977, 1979; Braakman *et al.*, 1979). In this study early features of patients with severe head injuries are recorded prospectively and correlated with outcome, and results have confirmed that eye movements are of prognostic value (Table 1). This is compatible with the value of eye movements as an indication of brainstem function, and also with the concept of a centripetal accumulation of brain damage after head injury, the brainstem being affected only in more severe injuries.

Table 1 *Features with major prognostic value in the first 1000 patients in the international study on prognosis of severe head injury*

Feature
Age in decades
Best motor response, arms
Sum score Glasgow coma scale
Pupil reactions to light
Spontaneous eye movements, oculocephalic response and vestibulo-ocular response
Motor pattern
Presence or absence of apnoea
Change within a certain period of time

Address for reprint requests: Dr R. Braakman, Neurosurgical Department, Academic Hospital Rotterdam Dijkzigt, Erasmus University, Rotterdam, The Netherlands.

Accepted 10 May 1979

The practical prognostic value of all these signs depends on the consistency of assessment. This is determined by both the inter- and intra-observer agreement.

Inter- and intraobserver disagreement may result from the following factors: (1) the patient having undergone a real change in the time that elapsed between observations; (2) different technique of eliciting response; (3) bias or "emperor's clothes syndrome" (Andersen, 1835; White *et al.*, 1969; Gross, 1971; van Gijn and Bonke, 1977); (4) a difference between observers regarding boundary lines of categories, despite similarity in classification system; (5) personality differences, the doubter against the resolute (the doubter is an observer who rates categories like "uncertain" and "other", more frequently than the resolute); (6) mistakes in scoring.

The aim of the present study was to determine the interobserver agreement in rating (in-) equality of pupils, reaction of pupils to light, spontaneous eye movements, and oculocephalic responses.

Patients and methods

PATIENT POPULATION AND OCULAR SIGNS STUDIED

Over a period of three months spontaneous and elicited ocular signs were studied in 30 consecutive patients in whom consciousness was depressed to a level at which the eyes would not open in response to painful stimuli, commands

were not obeyed, and verbal response was, at best, comprehensible—that is, sounds, no words —at the time of the study.

The study included patients with depressed level of consciousness caused either by head injury or by non-traumatic factors.

The following four ocular signs were studied: (in-)equality of pupils; reaction of pupils to light; spontaneous eye movements; oculocephalic responses. The categories of spontaneous eye movements and of oculocephalic responses that were offered to the observer are shown in Table 2. The validity of the rank order with regard to prognosis was confirmed by computer analysis of the data of the first 1000 patients in the international databank of the study on the prognosis of severe head injury (Jennett *et al.*, 1976, 1977; Braakman, 1978; Braakman *et al.*, 1979).

Table 2 *Categories of spontaneous eye movements and oculocephalic responses offered to observers*

Spontaneous eye movements	*Oculocephalic responses*
1 Normal (fixation)	1 Fixation (normal)
2 Roving conjugate	2 Full
3 Roving disconjugate	3 Minimal
4 Lateral deviation	4 Absent
5 None	
6 Other	

When in doubt the investigators were instructed to rate the higher category and to give the patient the benefit of the doubt on prognosis. For the oculocephalic responses the best response observed was rated.

Oculocephalic responses were considered "minimal" when a response was slight or doubtful, "full" when a clear response was seen to at least one side. Otherwise no strict definitions were given to the observers for rating, in order not to influence pre-existing ideas.

Six observers participated in the study. Each patient was examined by four doctors from this regular group of six, all with several years experience in neurological examination. Two of the physicians were senior staff members (Br and vG). All patients had been in coma for at least six hours, and seemed to be in a steady state at the time. All examinations on a patient were performed consecutively within one hour. It is unlikely that their level of consciousness changed during that period. Two patients were excluded because not all four examinations could be performed within one hour, leaving 28 cases for study. The physicians remained ignorant of each other's results. No information was given on the case histories.

STATISTICAL ANALYSIS

Overall observer agreement

This may be defined as the proportion of instances in which the observers agree as to the presence or absence of a certain feature. It depends on both the number of investigators and the number of categories recognised when scoring an item. However, it is also necessary to take into account the possibility that agreement occurs from "chance". To compensate for these influences a coefficient of pairwise agreement was used.

Coefficient of pairwise agreement between observers

It was assumed that no interobserver bias existed as the group of four physicians was not identical for each patient. In other words, we suppose that no physician gives a particular score more often than another. A raw measure of pairwise agreement is P_o, the number of agreeing pairs divided by the total number of pairs. Suppose a patient receives the four scores 1, 1, 2, 2. Although there is no overall agreement on this patient there are two agreeing pairs, namely 1, 1 and 2, 2 and four disagreeing pairs. P_o therefore expresses the probability that two physicians will give the same score to a patient. On the other hand agreement may also occur by chance. The probability of this is estimated by P_C. Suppose there are two possible scores + and − and the probability of a + equals 0.3. Then the probability that two observers both give a + solely on the basis of chance is $0.3\times0.3=0.09$ and P_C is $0.3\times0.3+0.7\times0.7=0.58$. For P_C a fixed schedule cannot be given, because in all cases different corrections have to be made for the number of times that every single category was chosen.

The estimation procedure of P_C is given by Fleiss (1971). (See also Cohen, 1960, 1968; Spitzer *et al.* 1967; Spitzer and Fleiss, 1974; Koran, 1975.)

Since P_C is the proportion of agreement expected by chance, $1-P_C$ may be regarded as a measure of possible improvement in the proportion of agreement that may be obtained and P_o-P_C expresses the actual improvement. Thus a chance corrected measure of agreement is the coefficient $(P_o-P_C)/(1-P_C)$ termed kappa (K). $K=0$ when there is only chance agreement and $K=1$ when complete agreement exists. A value of kappa between 0 and 1 indicates that there is more agreement than is to be expected from chance alone.

Kappa was calculated for the assessment of

(in-)equality of pupils, reaction of pupils to light, spontaneous eye movements, and oculocephalic responses. The standard error of kappa was computed in order to indicate how accurately the agreement is measured (the computational procedure is explained in the appendix).

The null hypothesis that the physicians only agreed by chance was tested by means of a standard normal value Z (Fleiss, 1971).

The categories are ranked to gravity. *Weighted kappa* ($= K_w$) takes account of the fact that disagreement in choosing two close possibilities is not as serious as choosing two less proximate possibilities. The determination of weighted kappa in this study was not possible because the difference between the categories could not be rendered in exact numerals.

Results

OVERALL AGREEMENT

The scores of 28 patients, given by four physicians per patient, are shown in Table 3. Not only did small differences occur, but sometimes even conflicting scores were given. In case 23, for example, two of the four physicians considered the right pupil the largest and two the left. In another case (19) three of the four physicians did not observe any spontaneous eye movement, while the fourth rated roving conjugated eye movements.

The overall agreement is shown in Table 4. It appears that the highest agreement occurred in the assessment of the reaction of pupils to light.

Table 4 *Overall agreement between observers*

Feature	*Agreement*	*%*
Inequality or equality of pupils	16/28	57
Reaction of right pupil to light	19/28	68
Reaction of left pupil to light	20/28	71
Spontaneous eye movements	13/28	46
Oculocephalic responses	12/28	43

PAIRWISE AGREEMENT

Table 5 contains the computed values of P_o, P_c, kappa and its standard error, and Z for each investigated feature. Each Z value corresponds with a tail probability smaller than 0.02, meaning that each kappa is significantly greater than 0. In other words: the observed proportion of agreement P_o is significantly greater than the chance proportion of agreement P_c.

The agreement in the assessment of pupils to light and in the assessment of (in-)equality of pupils appeared to be satisfactory. More disagreement between the investigators occurred in assessing spontaneous eye movements and oculocephalic responses.

Discussion

Any clinical sign is subject to different interpretation by various observers.

In recent years there has been increasing interest in studies of clinical observer disagreement (Houfek and Ellingson, 1959; Bull *et al.*, 1960; Abbassioun *et al.*, 1966; McCance *et al.*, 1968; Woody, 1968; White *et al.*, 1969; van Gijn, 1977; Teasdale *et al.*, 1978; a review of observer disagreement studies is given by Koran, 1975).

The prognosis of patients comatose as a result of head injury, includes the level of consciousness as represented in the three aspects of the Glasgow Coma Scale and also on brainstem signs like pupil reactions and eye movements.

The first reference to observer disagreement in comatose patients is by Teasdale and Jennett (1976). Braakman *et al.* (1977) demonstrated a satisfactory agreement in the assessment of the motor response of the Glasgow Coma Scale. This was confirmed by Teasdale *et al.* (1978), who also found a satisfactory agreement in the assessment of eye opening and verbal response in comparison with other features. They observed a lower rate of agreement in assessing the (in-)equality of the pupils than in their reaction to light. We were able to show that the observer agreement in assessing ocular signs is more than may be expected from chance. Yet the calculated values of kappa are not high. Especially in the interpretation of spontaneous eye movements and oculocephalic responses, only a moderate interobserver agreement was seen. Studies on

Table 5 *Pairwise agreement between observers*

Feature	*Number of categories*	P_o	P_c	K	$SE(K)$	Z
Inequality or equality of pupils	3	0.76	0.39	0.61	0.09	8.71
Reaction right pupil	2	0.82	0.51	0.62	0.46	2.10
Reaction left pupil	2	0.83	0.51	0.65	0.45	2.25
Spontaneous eye movements	6	0.63	0.31	0.46	0.10	7.08
Oculocephalic responses	4	0.67	0.35	0.49	0.09	8.07

Table 3 *Scores of equality of pupils, reaction of pupils to light, spontaneous eye movements and oculocephalic responses, in 28 patients, given by four physicians per patient. The four physicians belong to a regular group of six*

Patient	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*	*21*	*22*	*23*	*24*	*25*	*26*	*27*	*28*	
Equality of pupils																													
Br	<	=		=			=	>		=	=		=	=						>	<					=	>	=	= =R=L
vG	<	=	=		=	=	<	>	=	=					>	=	<	=	=			=	>	<	<			=	> =R>L
Bo			>		>	=					<	>	>	=	>	=	<	=	=	>	<	>	>	<	<	=	=		
Fr	<	=	>	=	=	<			=	<	<	>												<	<				< =R<L
Be				=	=	=	=	>	=	=		>	=	=	>	=	<	=	=	>	<	=	<			=	=	=	
vdBe	<	=	=	=			<	>	=		<	>	=	=	>	=	<	<	=	=	<	>	<	<	<	=	=	=	
Reaction right pupil																													
Br	–	–		–			+	+		+	–		+	+						+	–					+	–	+	
vG	–	–	–		+	+	–	–	+	+					–	+	–	–	–			–	–	–	–			+	
Bo			–		+	–					–	–	+	+	–	+	–	+	+	–	–	–	–	–	–	+	+		
Fr	–	–	–	+	+	–			+	+	–	–												–	–				
Be				+	+	–	+	+	+	+		–	+	–	–	+	–	–	+	+	–	–	–			+	+	+	
vdBe	–	–	–	–			+	–	+		–	–	+	–	–	+	–	–	+	+	–	–	–	–	–	+	–	+	
Reaction left pupil																													
Br	–	–		–			+	+		+	–		+	+						+	–					+	–	+	
vG	–	–	–		+	+	+	–	+	+					–	+	–	–	+			–	–	–	–			+	
Bo			–		+	–					–	–	+	+	–	+	–	+	+	–	–	–	–	–	–	+	+		
Fr	–	+	–	–	+	–			+	+	–	–												–	–				
Be				+	+	–	+	+	+	+		–	+	–	–	+	–	+	+	+	–	–	–			+	+	+	
vdBe	–	–	–	–			+	–	+		–	–	+	–	–	+	–	–	+	–	–	–	–	–	–	+	–	+	
Spontaneous eye movements																													
Br	5	6		5			2	5		6	5		2	6						5	5					2	5	6	1 normal (fixation)
vG	3	2	5		4	5	2	5	2	3					3	3	3	3	5			5	5	5	5			2	2 roving conjugate
Bo			5		4	5					5	5	3	2	6	5	5	5	5	5	5	5	5	5	5	4	5		3 roving disconjugate
Fr	1	3	5	4	4	5			2	2	5	5												5	5				4 lateral deviation
Be				3	4	3	2	5	2	6		5	2	6	2	6	3	3	5	5	5	6	5			2	5	2	5 none
vdBe	3	2	5	2			2	5	2		5	5	3	6	3	3	3	2	2	5	5	5	5	5	5	3	5	2	6 other
Oculocephalic reactions																													
Br	4	2		2			3	4		2	4		2	4						4	4					2	4	2	1 fixation (normal)
vG	3	2	4		3	2	4	4	3	2					2	3	2	2	3			4	4	4	2			2	2 full
Bo			4		4	2					4	4	3	4	2	3	2	2	3	2	4	4	4	4	2	2	4		3 minimal
Fr	4	2	3	2	3	3			3	2	4	4												4	3				4 absent
Be				3	3	2	2	4	3	2		4	2	3	2	2	2	3	2	3	4	4	4			3	4	2	
vdBe	3	2	3	2			2	4	2		4	4	2	4	2	2	2	2	2	2	4	4	4	4	2	2	4	1	

observer agreement in vestibulo-ocular responses are still in progress, as it was difficult to assess kappa for these responses without audiovisual methods.

In our study junior doctors did not disagree more often than their more senior colleagues.

A reason for the relatively low agreement rates may be that no attempt was made to influence pre-existing ideas on the classification of the responses, as it was our intention to approach the actual clinical situation as closely as possible. More exact definitions of the various types of spontaneous eye movements and oculocephalic responses might have improved the agreement rates.

P_C in kappa $(P_O-P_C/(1-P_C)$ is corrected for the number of categories. The lowest kappa values in our study were observed in the features with the largest number of categories. It is more difficult to obtain total agreement in a classification of many categories than in case of just two. This means a classification should not contain more categories than necessary for the purpose; the number of classes should be limited.

Apart from a study on the motor response of the Glasgow coma scale (Braakman *et al.*, 1977), too few studies on observer agreement analysed by this method have yet been performed to allow comparison of the observer values of kappa for ocular signs with those observed for other features.

In clinical practice the moderate observer agreement rate will decrease the value of separate ocular signs as prognostic features in patients with impaired consciousness. Possibly a higher agreement rate may be achieved by combining the ocular signs to form an index as a total eye score consisting of spontaneous eye movements plus oculocephalic responses, perhaps in combination with the vestibulo-ocular responses.

When features with a high discriminating value are used to assess prognosis in individual patients, the interobserver agreement should be taken into account. A feature, even if generally considered to be of prognostic significance, will be of little value in making predictions in individual patients when the interobserver agreement is very small. Prognostic features should have not only a large discriminating value, but also a high interobserver agreement rate.

Our results confirm that doctors may disagree not only on the best method of treatment for a particular patient, but even on the interpretation of basic clinical signs. Important management decisions should not rely on one single observation by one physician, but preferably on repeated examinations by different physicians.

We would like to express our gratitude to M. W. Berfelo, A. Boesten, K. Franke, and J. van Gijn, for scoring the symptoms. This paper was completed with the aid of a grant of the Fund of the KNAC (Royal Dutch Automobile Club) for juvenile victims of traffic accidents.

References

Abbassioun, K., Walker, A. E., Udvarhelyi, G. B., and Fueger, G. F. (1966). Critical evaluation of brain scan. *Neurology (Minneapolis)*, **16**, 746–748.

Andersen, H. C. (1853). *Fairy Tales.* Copenhagen.

Braakman, R. (1978). Data bank of head injuries in three countries. *Scottish Medical Journal*, **23**, 107.

Braakman, R., Avezaat, C. J. J., Maas, A. I. R., Roel, M., and Schouten, H. J. A. (1977). Interobserver agreement in the assessment of the motor response of the Glasgow "coma" scale. *Clinical Neurology and Neurosurgery*, **80**, 100–106.

Braakman, R., Habbema, J. D. F., Gelpke, G. J., and Minderhoud, J. M. (1979). Prognose van patienten met zwaar hersenletsel. *Nederlands Tÿdschrift van Geneeskunde.* In press.

Bricolo, A., Turazzi, S., Alexandre, A., and Rizzuto, Nl. (1977). Decerebrate rigidity in acute head injury. *Journal of Neurosurgery*, **47**, 680–698.

Bull, J. W. D., Couch, R. S. C., Joyce, D., Marshall, J., Potts, D. G., and Shaw, D. A. (1960 . Observer variation in cerebral angiography: an assessment of the value of minor angiographic changes in the radiological diagnosis of cerebrovascular disease. *British Journal of Radiology*, **33**, 165–170.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Education and Psychological Measurement*, **20**, 37–46.

Cohen, J. (1968). Weighted Kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, **70**, 213–220.

Fisher, C. M. (1969). Neurological examination of the comatose patient. *Acta Neurologica Scandinavica*, **45**, supplement 36.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, **76**, 378–382.

Gijn, J. van (1977). *The plantar reflex. A historical, clinical and electromyographic study*, pp. 53–63. Thesis Rotterdam. Krips Repro: Meppel.

Gijn, J. van, and Bonke, B. (1977). Interpretation of plantar reflexes: biasing effect of other signs and symptoms. *Journal of Neurology, Neurosurgery, and Psychiatry*, **40**, 787–789.

Gross, F. (1971). The emperor's clothes syndrome. *New England Journal of Medicine,* **285,** 863.

Houfek, E. E., and Ellingson, R. J. (1959). On the reliability of clinical EEG interpretation. *Journal of Nervous and Mental Diseases,* **128,** 425–437.

Jennett, B., Teasdale, G., Braakman, R., Minderhoud, J. M., Heiden, J., and Kurze, T. (1979). Prognosis of patients with severe head injury. *Neurosurgery,* **4,** 283–289.

Jennett, B., Teasdale, G., Braakman, R., Minderhoud, J., and Knill-Jones, R. (1976). Predicting outcome in individual patients after severe head injury. *Lancet,* **1,** 1031–1034.

Jennett, B., Teasdale, G., Galbraith, S., Pickard, J., Grant, H., Braakman, R., Avezaat, C. J. J., Maas, A. I. R., Minderhoud, J. M., Vecht, C. J., Heiden, J. S., Small, R., Caton, W., and Kurze, T. (1977). Severe head injuries in three countries. *Journal of Neurology, Neurosurgery, and Psychiatry,* **40,** 291–298.

Koran, L. M. (1975). The reliability of clinical methods, data and judgments. *New England Journal of Medicine,* **293,** 642–646 and 695–701.

McCance, C., Watt, J. A., and Hall, D. J. (1968). An evaluation of the reliability and validity of the plantar response in a psychogeriatric population. *Journal of Chronic Diseases,* **21,** 369–374.

Mingrino, S., Molinari, G., Andrioli, G., and Frugoni, P. (1965). Some observations upon vestibular reactions in acute head injury. In *Proceedings of the Third International Congress of Neurological Surgery.* Excerpta Medica Foundation International Congress Series, 110.

Poulsen, J., and Zilstorff, K. (1972). Prognostic value of the caloric-vestibular test in the unconscious patient with cranial trauma. *Acta Neurologica Scandinavica,* **48,** 282–292.

Spitzer, R. L., and Fleiss, J. L. (1974). A reanalysis of psychiatric diagnosis. *British Journal of Psychiatry,* **125,** 341–347.

Spitzer, R. L., Cohen, J., Fleiss, J. L., and Endicott, J. (1967). Quantification of agreement in psychiatric diagnosis. *Archives of General Psychiatry,* **17,** 83–87.

Teasdale, G., and Jennett, B. (1976). Assessment and prognosis of coma after head injury. *Acta Neurochirurgica (Vienna),* **34,** 45–55.

Teasdale, G., and Smith, J. (1975). Eye movements and brainstem dysfunction after head injury. *Journal of Neurology, Neurosurgery, and Psychiatry,* **38,** 822–823.

Teasdale, G., Jennett, B., and Knill-Jones, R. (1976). Assessment and prognosis of severe head injury. *Acta Neurochirurgica,* **34,** 45–55.

Teasdale, G., Knill-Jones, R., and Sande, J. van de (1978). Observer variability in assessing impaired consciousness and coma. *Journal of Neurology, Neurosurgery, and Psychiatry,* **41,** 603–610.

White, D. N., Kraus, A. S., Clark, J. M., and Campbell, J. K. (1969). Interpreter error in echoencephalography. *Neurology (Minneapolis),* **19,** 775–784.

Woody, R. H. (1968). Interjudge reliability in clinical electro-encephalography. *Journal of Clinical Psychology,* **24,** 251–256.

Appendix

In computing the standard error of kappa, the chance proportion of agreement P_C is treated as a constant. Since the notation of Fleiss (1971) is used below, the following can only be understood after reading his paper.

The estimated variance of kappa may be written as

$$\text{var}(K) = \text{var}(\Sigma_i s_i)/\{ Nn(n-1)\,(1-\Sigma_j p_j^2)\ \}^2,$$

where $s_i = \Sigma_j n^2_{ij}$,

and $\text{var}(\Sigma_i s_i) = \Sigma_i s_i^2 - (\Sigma_i s_i)^2/N$.

The standard error of kappa SE(K) equals the square root of var(K). In the case of the binary variable pupil reaction, formulae for agreement on a particular category are used.