Inter-Observer Agreement in Assessing Comatose Children

Shashi S. Seshia, Jerome Y. Yager, Bruce Johnston and Philip Haese

ABSTRACT: Inter-observer agreement was evaluated for twelve items used in the neurological assessment of comatose children. Data were obtained prospectively on fifteen patients examined independently by two observers in a double-blind fashion. Observer variability was measured by using the Disagreement Rate and Kappa statistic. The Disagreement Rate ranged from 0.01 to 0.12 for all items. Values for Kappa statistic were generally in accordance with those for Disagreement Rate. The data suggest fair to almost perfect inter-observer agreement for the items used to assess comatose children in this study.

RÉSUMÉ: Concordance inter-observateur dans l'évaluation d'enfants comateux Nous avons évalué la concordance inter-observateur pour douze items utilisés dans l'évaluation neurologique d'enfants comateux. Les données ont été obtenues prospectivement chez quinze patients examinés indépendamment par deux observateurs, en double insu. La variabilité reliée à l'observateur a été mesurée par le Taux de Discordance et la statistique Kappa. Le Taux de Discordance variait de 0.01 à 0.12 pour tous les items. Les valeurs de la statistique Kappa étaient généralement en accord avec les Taux de Discordance. Les données suggèrent une concordance inter-observateur de moyenne à presque parfaite pour les items utilisés pour évaluer les enfants comateux dans cette étude.

Can. J. Neurol. Sci. 1991; 18: 472-475

Clinical data in the comatose patient, using the approach of Plum and Posner,[1,2] provide clues to diagnosis, clinical course and prognosis not only in adults[1-4] but also in children.[5,6] With the exception of a study limited to ocular signs in twenty-eight comatose adults,[7] none has formally tested reliability for all the variables used to assess comatose patients. Reliability is influenced by inter-observer agreement. In this paper, we report on the inter-observer agreement for twelve clinical variables used in the neurological evaluation of comatose children.

Patients and Methods

The study was done prospectively from September 1986 to December 1987.

Clinical

Comatose children admitted to the Pediatric Intensive Care Unit (neonates excluded) of the Children's Hospital, Winnipeg, Canada were included if they could be examined by two clinicians (JYY and SSS) (i) within half-hour of each other to minimize the possibility of clinical change in the interval and (ii) within 24 hours of admission.

The data sheet (Table 1) was designed by one of us (SSS) with the following assumptions, (i) observers would be familiar with the publications[1-6] that had dealt with the variables used in the present study; for example, the term "diffuse" under motor response (item 6) had been used previously to describe bilateral abnormalities of tone and/or deep tendon reflexes other than decorticate or decerebrate patterns,[5] (ii) observers would select only one class for items 1 through 11, (iii) blood pressure (item 9) and temperature (item 10) would be taken at the time the individual observer examined the child, (iv) observers would be aware of the normal values for blood pressure in the pediatric age group, referenced in several readily available texts, (v) information about the seizure time of onset (item 11) and seizure type (item 12) would be obtained by interviewing the guardian or from the hospital chart and (vi) for seizure type, observers would not only record the type but also note if it was prolonged (>15 minutes) or not.

The clinicians (i) filled out separate data sheets, (ii) did not observe each other's examinations and (iii) did not discuss their views on the interpretation or classification of the clinical items either before or during the study. Specifically, no attempt was made by the senior clinician (SSS) to instruct or influence the junior (JYY) on any aspect of the assessment to minimize training bias.

The study was approved by the Faculty Committee on the use of Human Subjects in Research, Faculty of Medicine, University of Manitoba, Winnipeg.

From the Section of Pediatric Neurosciences (S.S., J.Y.Y.), and Department of Statistics (B.J., P.H.), University of Manitoba and Children's Hospital (S.S., J.Y.Y.), Winnipeg

Received December 27, 1990. Accepted in final form May 8, 1991

Presented in part at the XXIII Canadian Congress of Neurological Sciences, June 17, 1988, Quebec City, Canada

Reprint requests to: S.S. Seshia, M.D., Section of Pediatric Neurosciences, University of Manitoba and Children's Hospital, 840 Sherbrook Street, Winnipeg, Canada R3A 1S1

Table 1: Classification of Items

1. FUNDUSCOPY
   - i) Normal
   - ii) Abnormal
2. EXTRAOCULAR MOVEMENTS (including cold caloric and Doll's eye response)
   - i) Orienting
   - ii) Normal (midline at rest and full)
   - iii) Nystagmus (at rest)
   - iv) Spontaneous roving
   - v) Isolated 6th weakness (unilateral or bilateral)
   - vi) Blinking (spontaneous)
   - vii) Impaired lateral gaze
   - viii) Impaired medial gaze
   - ix) Impaired up/down gaze
   - x) Ocular bobbing/other abnormal EOM
   - xi) Combinations (of iii to x)
   - xii) Absent
3. PUPILLARY RESPONSE
   - i) 'Normal' (>2 mm)
   - ii) Small (<2 mm) reactive
   - iii) Unequal reactive
   - iv) Small (<2 mm) non-reactive
   - v) Unequal non-reactive
   - vi Equal (>2 mm) non-reactive
4. CORNEAL REFLEXES
   - i) Present
   - ii) Absent
5. GAG REFLEX
   - i) Present
   - ii) Absent
6. MOTOR RESPONSE
   - i) Normal
   - ii) Monoparesis
   - iii) Hemisyndrome
   - iv) Diffuse signs
   - v) Decortication
   - vi) Decerebration
   - vii) Generalized flaccid & areflexic
7. DEEP TENDON REFLEXES
   - i) Normal
   - ii) Increased
   - iii) Absent
8. RESPIRATORY PATTERN
   - i) Normal
   - ii) Cheyne-Stokes respiration
   - iii) Central hyperventilation
   - iv) Ataxic, apneustic, irregular
   - v) Apnea
   - vi) Assisted
   - vii) Non CNS disturbance
9. BLOOD PRESSURE
   - i) Normal
   - ii) Hypertension
   - iii) Hypotension/unable to maintain
10. TEMPERATURE (RECTAL)
    - i) Normal
    - ii) Febrile (> 38°C)
    - iii) Hypothermia (<35°C)/unable to maintain
11. SEIZURE (TIME OF ONSET)
    - i) Absent
    - ii) Onset of coma
    - iii) After onset of coma
12. SEIZURE (TYPE)
    - i) Absent
    - ii) Generalized tonic clonic
    - iii) Focal
    - iv) Multifocal/myclonic
    - v) Prolonged (>15 minutes)

Statistical

Data were analyzed at the end of the study period by two of us (BJ and PH). Sixteen children met the inclusion criteria. However, one of the two data sheets on one patient was misplaced. Inter-observer agreement was tested on the data from the remaining fifteen using three separate methods: (i) Disagreement Rate as proposed by Teasdale et al.[8] (ii) Kappa statistic,[9] and (iii) weighted Kappa,[10] the weights assigned being those described by Fleiss.[11]

The Disagreement Rate has a range of 0 to 0.5, lower values reflecting higher agreement.[8] Kappa values range from –1 to +1; minus scores for Kappa reflect less than chance agreement, positive values suggest greater than chance agreement and a Kappa of 0 indicates chance agreement.[9] Kappa values were interpreted according to the criteria of Landis and Koch:[12] less than 0, poor agreement; 0 to 0.20, slight agreement; 0.21 to 0.40, fair agreement; 0.41 to 0.60, moderate agreement; 0.61 to 0.80, substantial agreement and 0.81 to 1, almost perfect agreement.

The data sheets (Table 1) did not include either a category titled "unable to examine" or one to reflect the failure of the clinician to make an entry under any of the twelve items. The former omission was an oversight but the latter was deliberate to simulate recording in the clinical setting. When the data sheets were analyzed, we found that one or other observer had failed to make an entry for four items (junior observer in three instances and senior in one) on two patients, whilst the other had done so. In a third case, one observer recorded his inability to examine the fundi, the other marking the fundi as normal. Such differences in entry were treated as disagreements by the statistical methods. On the other hand, observations were considered to be in full agreement if both clinicians recorded "unable to examine" for the same item in the same patient. The information for "funduscopy" (item 1) which is used as a representative example, is summarized in Table 2.

Results

General Clinical Data The ages of the 15 children ranged from 2 months to 17 years. The mean age was 3 1/2 years and the median age was 16 months. Coma was due to traumatic causes in four cases and non-traumatic in eleven.

Statistical Data Disagreement Rates for the twelve items ranged from 0.01 to 0.12. Values for Kappa statistics were generally in accordance with those for Disagreement Rates, a lower Disagreement Rate being associated with a relatively higher Kappa value for all items except respiratory pattern (Table 3). The values for Kappa were not materially different from those for weighted Kappa. Hence, the values for weighted Kappa are not included in the table. The Kappa values suggested (i) fair

Table 2: Funduscopy

Class/Category	Observer 1	Observer 2
Normal	11	10
Abnormal	2	3
Recorded "unable to examine"	1	2
Made no entry	1	0
Total	15	15

https://doi.org/10.1017/S0317167100032182 Published online by Cambridge University Press

agreement for extraocular movements, (ii) moderate agreement for funduscopy, corneal reflexes, gag reflex, deep tendon reflexes, blood pressure and temperature, (iii) substantial agreement for motor response, seizure time and seizure type and (iv) almost perfect agreement for pupillary response.

The Disagreement Rate of 0.03 for respiratory pattern suggested good agreement between the clinicians but the Kappa value of –0.05 suggested less than chance agreement (Table 3). All patients were intubated and on assisted ventilation at the time of assessment.

Discussion

The scaling of clinical items used for evaluating comatose children and adults is based on anatomic, physiologic and clinical principles[1-6] and has been generally accepted in clinical practice. Assessment of inter-observer agreement for these items is desirable, (a) since important judgmental decisions are made on the basis of clinical information and (b) if the data are to be used in clinical research.[13]

The Disagreement Rates of 0.01 to 0.12 in our study suggest that the degree of disagreement was relatively small for most if not all variables, although limits have not been established for the boundaries of clinically acceptable disagreement. Teasdale et al.8 found Disagreement Rates of 0.03 to 0.22 in their assessment of items in the Glasgow Coma Scale and considered these values to reflect more consistent assessments than higher values. Reilly et al.[14] "arbitrarily assumed" (sic) that a Disagreement Rate greater than 0.10 was unacceptable, in their study on observer reliability for the pediatric version of the Glasgow Coma Scale.

Kappa values were in accord with Disagreement Rates and suggested fair to substantial agreement beyond chance for all items, except respiratory pattern. For this variable, the two observers agreed on assisted respiration as the class in twelve of 15 patients (Table 4); for two patients, one observer assigned class vi (assisted) and the other selected class v (apnea); in one patient, one observer chose class iv (ataxic, apneustic, irregular) and the other picked class vi (assisted). Such close clinical agreement is reflected in the relatively low Disagreement Rate (0.03), although the near zero value for Kappa in this situation (–0.05) implies less than chance agreement. We have previously suggested that Disagreement Rate and Kappa statistics may provide different yet complementary information about inter-observer agreement;[15] whereas the former provides a better measure of the degree of disagreement, the latter corrects for chance expected agreement. Our experience with the data in the present study reinforces that impression.

Some of the inter-observer disagreements occurred because one or other physician failed to record his finding or was unable to assess a particular variable, occurrences not uncommon in clinical practice. Disagreements between the actual observations made and recorded by the clinicians in this study may be due to, (a) changes in clinical phenomena during the half hour period between examinations, (b) differences in interpreting clinical observations and (c) differences in defining ambiguous clinical terms. The observers did not discuss their interpretation of the classification of the items before the study so as to simulate the usual clinical setting in which most individuals who assess comatose children do so without prior discussions on their respective methods of examination.

The methodological biases and limitations of our approach have been discussed recently[15] and include (i) the use of only two observers, (ii) the background of the observers, one a pediatric neurologist and the other a pediatric neurology fellow, and (iii) the small sample size. Ethical considerations limited the number of "observers" who could examine a critically ill child within a short time of each other. Pediatric intensivists, pediatric residents and nurses could not participate because of their extended clinical commitment and the rotational system of patient coverage. Although a deliberate attempt was made to minimize training bias, such a factor may have influenced the results because of the relatively long duration of the study. Koran[16,17] has suggested that studies of clinical reliability should be done with two or three physicians to mimic clinical practice and has drawn attention to the factors that influence

Table 3: Disagreement Rate and Kappa Statistics

Items in the Neurological assessment	Disagreement Rate	Kappa Statistics			
		Kappa	S.E.	C.I.	p-value
Funduscopy	0.10	0.57*	0.19	(0.19,0.95)	<.001
Extraocular movements	0.12	0.39*	0.13	(0.14,0.64)	.001
Pupillary response	0.01	0.81*	0.16	(0.50,1.0)	<.001
Corneal reflexes	0.09	0.50*	0.23	(0.05,0.95)	.014
Gag reflex	0.06	0.59*	0.26	(0.08,1.0)	.010
Motor response	0.09	0.62*	0.14	(0.34,0.89)	<.001
Deep tendon reflexes	0.12	0.58*	0.18	(0.22,0.94)	.001
Respiratory pattern	0.03	–0.05	0.11	(–0.28,0.18)	.658
Blood pressure	0.12	0.44*	0.21	(0.02,0.87)	.018
Temperature	0.08	0.52*	0.19	(0.15,0.90)	.003
Seizure (time of onset)	0.05	0.65*	0.18	(0.29,1.0)	<.001
Seizure (type)	0.05	0.66*	0.16	(0.35,0.98)	<.001

*Significantly different from zero at 95% level, one-tailed test.
S.E.: Standard error.
C.I.: Confidence interval.
Minor discrepancies in C.I. are due to rounding of Kappa and S.E.

agreement. For example, pairs of physicians with more training to the test task will agree more than pairs with less training.[17] Sample sizes have generally been limited in studies of this type because it is difficult to get the same set of examiners to assess patients within a short time of each other.[13] Teasdale et al.[8] had 16 patients and Reilly et al.[14] had 15 patients in their respective studies.

Some of the short-comings of our study may be avoided by using video recordings of comatose patients, a method employed by Reilly et al.[14] provided (i) permission to do so is obtained from parents and Ethics Committees and (ii) the recordings precisely and consistently capture all the aspects of the assessment. However, such an approach will not test variability created by differences in (a) examination methods, (b) recording of findings and (c) clinical performance, factors that are important in clinical practice, particularly in an intensive care setting.

We would suggest three modifications to the classification of two of the 12 items used in this study. First, for respiration (item 8), classes v and vi should be collapsed into one class of "assisted" since those who are apneic will be intubated and their ventilation assisted. Second and also under respiration (item 8), class vii of "Non-CNS disturbance" should become a distinct item, for example 8B, to unambiguously distinguish between respiratory disturbances not of neurological origin and other classes (item 8, classes i to vi) which have been scaled neurologically.[1-6] "Non-CNS disturbance" such as that caused by pulmonary or cardiac involvement was included in the data sheet to differentiate neurological from non-neurological causes of respiratory disturbance, recognizing that either both could co-exist or it might be clinically difficult to distinguish them. Finally, under seizure type (item 12), duration of the seizure (class v — prolonged) should be separated from type of seizure since it is an entirely different variable. These steps may improve inter-observer agreement.

Clinical agreement can be enhanced by adopting the approaches suggested by Koran[16,17] and the McMaster group.[18,19] These include, (a) clear definition of and agreement on terms, criteria and examination items listed in data recording forms, (b) observation (either directly or by video recording) of interview and physical examination techniques so that deficient skills can be identified and corrected, an inherent assumption in clinical training in which one or more physicians serve as a "gold standard" and (c) reporting evidence rather than inference.

Table 4: Respiratory Pattern

Observer 1 Classification of Respiratory Pattern	Observer 2 Classification of Respiratory Pattern i	ii	iii	iv	v	vi	vii	Total
i Normal	-	-	-	-	-	-	-	-
ii Cheyne-Stokes	-	-	-	-	-	-	-	-
iii Central hyperventilation	-	-	-	-	-	-	-	-
iv Ataxic, apneustic, irregular	-	-	-	-	-	-	-	-
v Apnea	-	-	-	-	-	2	-	2
vi Assisted	-	-	-	1	-	12	-	13
vi Non CNS disturbance	-	-	-	-	-	-	-	-
Total	-	-	-	1	-	14	-	15

Acknowledgements

The study was supported by Health & Welfare Canada. We thank the staff of the Pediatric Intensive Care Unit for their co-operation and Drs. Victor Chernick and Richard Stanwick for reviewing the manuscript.

References

1. Plum F, Posner JB. The diagnosis of stupor and coma. Contemporary Neurology Series, Philadelphia: F.A. Davis Co. 1966.
2. Plum F, Posner JB. The diagnosis of stupor and coma; Edition 3: Contemporary Neurology Series, Philadelphia: F.A. Davis Co. 1980.
3. Bates D, Caronna JJ, Cartlidge NEF, et al. A prospective study of non-traumatic coma: Methods and results in 310 patients. Ann Neurol 1977; 2: 211-220.
4. Levy DE, Bates D, Caronna JJ, et al. Prognosis in nontraumatic coma. Ann Intern Med 1981; 94: 293-301.
5. Seshia SS, Seshia MMK, Sachdeva RK. Coma in childhood. Dev Med Child Neurol 1977; 19: 614-628.
6. Seshia SS, Johnston B, Kasian G. Non-traumatic coma in childhood: clinical variables in prediction of outcome. Dev Med Child Neurol 1983; 25: 493-501.
7. Van Den Berge JH, Schouten HJA, Boomstra S, et al. Interobserver agreement in assessment of ocular signs in coma. J Neurol Neurosurg Psychiatry 1979; 42: 1163-1168.
8. Teasdale G. Knill-Jones R, Van der Sande J. Observer variability in assessing impaired consciousness and coma. J Neurol Neurosurg Psychiatry 1978; 41: 603-610.
9. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960; 20: 37-46.
10. Cohen J. Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968; 70: 213-220.
11. Fleiss JL. Statistical methods for rates and proportions. New York: John Wiley and Sons, 1981, 212-236.
12. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics, 1977; 33: 159-174.
13. Feinstein AR. Clinimetrics. Yale University Press: New Haven and London, 1987:188, 215.
14. Reilly PL, Simpson DA, Sprod R, et al. Assessing the conscious level in infants and young children: a pediatric version of the Glasgow Coma Scale. Child's Nerv Syst 1988; 4: 30-33.
15. Yager JY, Johnston B, Seshia SS. Coma scales in pediatric practice. Am J Dis Child 1990; 144: 1088-1091.
16. Koran LM. The reliability of clinical methods, data and judgments. N Engl J Med 1975; 293: 642-246.
17. Koran LM. The reliability of clinical methods, data and judgments. N Engl J Med 1975; 293: 695-701.
18. Department of clinical epidemiology and biostatistics, McMaster University, Hamilton, Ont. Clinical disagreement: I. How often it occurs and why? Can Med Assoc J 1980; 123: 499-504.
19. Department of clinical epidemiology and biostatistics, McMaster University, Hamilton, Ont. Clinical disagreement: II. How to avoid it and how to learn from one's mistakes? Can Med Assoc J 1980; 123: 613-617.