Assessment of Interobserver Variability in a Dutch Multicenter Study on Acute Ischemic Stroke

Herman J. Gelmers, MD, PhD, Kasper Gorter, MD, PhD,
Cees J. de Weerdt, MD, PhD, and Hans J.A. Wiezer, MD

Quantitative assessment of patient data is a pertinent part of controlled clinical studies. When several centers are involved, the degree of agreement between different observers becomes important. Therefore, in addition to developing a multicenter study on acute ischemic stroke, we have estimated the interobserver agreement expressed in terms of κ statistics. Twelve patients suffering from neurologic deficits due to acute ischemic stroke were examined by four investigators, and the results were assessed using the Mathew scale. Considerable interobserver variability was found. Agreement on items based on subjective information from the patient was low, and it is also possible that this information changes with time. It is advised that in the development of assessment scales, items with low interobserver agreement should be avoided. (*Stroke* 1988;19:709–711)

Downloaded from http://stroke.ahajournals.org/ by guest on July 19, 2018

When research involves data collection in several centers with different observers, it is pertinent to evaluate to what extent different observers perceive and record the same information when confronted with the same phenomena. The investigators are usually very surprised to find considerable interobserver variability[1–4] in what they believed to be consistent findings. After starting a multicenter study on acute ischemic stroke, we decided to determine in a separate study the number of interobserver differences in recording neurologic signs. This article deals with the results of that study.

Subjects and Methods

Twelve patients, aged 62–84 years, who were hospitalized in a nursing home consented to participate in our study. All suffered from neurologic deficits as the result of acute ischemic strokes that had occurred 3–11 months earlier. They were selected randomly from patients available in the wards on one day by an independent and nonparticipating physician. The patients were unknown to the four participating investigators, and all patients were considered to be in a stable state.

The investigators were the four senior neurologists from the participating centers involved in a multicenter study on acute ischemic stroke who were directly responsible for patient care in their own units. All investigators were trained in clinical neurology, had at least 15 years' experience, and were especially interested in cerebrovascular diseases.

From the Department of Neurology, Streekziekenhuis, Almelo (H.J.G.), the Department of Neurology, St. Antonius Ziekenhuis, Sneek (K.G.), the Department of Neurology, Scheper Ziekenhuis, Emmen (C.J. de W.), and the Department of Neurology, St. Elisabeth Ziekenhuis, Venray (H.J.A.W.), the Netherlands.

Address for reprints: H.J. Gelmers, MD, PhD, Department of Neurology, Streekziekenhuis, 7609 PP Almelo, the Netherlands.

Received March 21, 1986; accepted December 1, 1987.

The neurologic deficits were assessed using the single items of an ordinal scale developed by Mathew et al,[5] modified in only minor detail for the assessment of language disturbance.[6] These items are listed in Figure 1 and were considered separately in the statistical analysis because there may be different levels of agreement for the separate parts of the neurologic investigation represented by the items. Definitions of the terms for different items were predetermined and specified in a protocol.

The sequence of the four observers' scoring of each patient was randomized to balance the interaction between investigator and patient (Figure 2). All evaluations were done the same morning to avoid day-to-day effects.

At the beginning of each session, each neurologist was given a case record form containing the scoring system for assessment of neurologic deficits and a list of the patients. The only information that the investigators had was the name and age of the aphasic patients. Each patient was assessed separately by the investigators. The case record form was filled out immediately after examination. Completed forms were not shown or discussed but were collected and put in a sealed envelope.

The level of interobserver variability or, inversely, interobserver agreement, was measured by κ separately for each single item, using the Fleiss method,[7,8] which provides a numerical measure of agreement among multiple investigators on variables scored on a nominal scale. κ is defined as $(P_o - P_e)/(1 - P_e)$, where P_o is the observed proportional agreement, that is, the number of all actual pairwise agreements divided by all possible pairwise agreements, and P_e is the proportion of agreements expected by chance. $\kappa = 1$ only when complete agreement is observed ($P_o = 1$) and there is some variation in the patients' neurologic status ($P_e < 1$). κ is undefined if $P_e = 1$. The significance of κ is tested by dividing it by its asymptotic standard error (S_κ). This ratio is asymptotically distributed as a

Factor	Score
Mentation	
Level of consciousness:	
fully conscious	8
lethargic but mentally intact	6
obtunded	4
stuporous	2
comatose	0
Orientation	
oriented x 3	6
oriented x 2	4
oriented x 1	2
disoriented	0
Speech	
normal	23
incoherent words	15
expressive or impressive words	10
speechless	0
Cranial nerves	
Homonymous hemianopsia	
intact	3
mild	2
moderate	1
severe	0
Conjugate deviation of eyes	
intact	3
mild	2
moderate	1
severe	0
Facial weakness	
intact	3
mild	2
moderate	1
severe	0

Factor	Score
Motor power*	
normal strength	5
contracts against resistance	4
elevates against gravity	3
gravity eliminated	2
flicker	1
no movements	0
Performance or disability status scale	
normal	28
mild impairment	21
moderate impairment	14
severe impairment	7
death	0
Reflexes	
normal	3
asymmetrical or pathological reflexes	2
clonus	1
no reflexes elicited	0
Sensation	
normal	3
mild sensory abnormality	2
severe sensory abnormality	1
no response to pain	0

*Each limb separately.

FIGURE 1. *Single items of Mathew scale for assessment of neurologic deficit in patients with acute ischemic stroke,[5] modified in minor detail.[6]*

standard normal variable. S_κ is only approximate because of the small sample size. It was not possible to perform this investigation on more than 12 patients.

The total score from the Mathew scale is listed in Figure 3.

Results

P_o, κ, and corresponding significance levels are listed in Figure 4 for all the neurologic items observed. κ showed considerable variation, that is, there was a different κ for each item. Using the criteria defined in the legend of Figure 4, the observed κ indicates excellent interobserver agreement for one item, moderate to substantial agreement for four items, and fair agreement for two items. Five items showed only poor interobserver agreement, and for one item κ was not assessable. On the basis of unadjusted *p* values, eight items showed a significantly higher agreement than would be expected by chance.

To look for order (e.g., learning) effects, κ was calculated again excluding the first, second, third, and fourth scorings, each of which was performed three times by the same observer. The differences between the results were small, and thus it can be concluded that the data did not show any order effect.

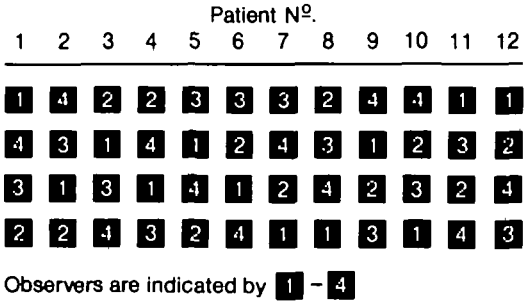

FIGURE 2. *Design of patient–observer assignments.*

Discussion

Our study confirms the results from other studies[1-4]; there is considerable interobserver variability in cooperative studies, even when performed by skilled and experienced observers. In a study of head injury patients, the assessment of reaction of the pupils to light gave an average κ of 0.64, while the assessment of pupillary equality had a κ of 0.61.[9] It would seem reasonable to expect the assessment of other important features in neurologic practice to be at least as consistent as those well-established clinical signs.

Patient	Observer 1	2	3	4
1	76	77	79	85
2	81	75	77	88
3	81	76	90	79
4	74	73	73	76
5	53	50	62	51
6	52	52	50	51
7	38	43	37	43
8	87	80	90	90
9	88	86	90	86
10	84	75	84	85
11	74	66	73	72
12	39	40	37	51

FIGURE 3. *Total Mathew score for all patients by observer.*

Downloaded from http://stroke.ahajournals.org/ by guest on July 19, 2018

item	P_o	P_e	Kappa*	S_{Kappa}	p≤
level of consciousness	0.92	0.92	0**	0.118	0.99
orientation	0.79	0.74	0.189	0.118	0.11
speech, aphasia	0.88	0.48	0.758	0.081	0.0001
homonymous hemianopsia	0.39	0.27	0.159	0.074	0.032
conjugate deviation of eyes	0.79	0.79	0**	0.093	0.099
facial weakness	0.47	0.40	0.126	0.100	0.21
motor power:					
right arm	0.96	0.54	0.909	0.111	0.0001
right leg	0.82	0.50	0.637	0.092	0.0001
left arm	0.58	0.24	0.455	0.061	0.0001
left leg	0.53	0.21	0.399	0.061	0.0001
performance or disability	0.71	0.33	0.563	0.084	0.0001
reflexes	1.00	1.00	–***	–	–
sensation	0.63	0.49	0.265	0.100	0.0081

FIGURE 4. *Observed (P_o), expected (P_e), and interobserver (Kappa) agreement, asymptotic standard error (S_{Kappa}), and level of significance (p). *It has been suggested in the literature[10,11] that when $\kappa > 0.80$ agreement can be considered excellent; $0.40 < \kappa < 0.80$ indicates moderate to substantial agreement, $0.20 < \kappa < 0.40$ indicates fair agreement, and $\kappa < 0.20$ indicates poor or slight agreement. **All investigators except one score all patients equally. ***All patients are scored identically by all investigators.*

In our study, the single items of the Mathew scale (a first attempt to describe neurologic status in acute brain ischemia in numerical terms) were used to assess neurologic deficits in stroke patients. Although this scale has not been validated in any way, it has been used for many years despite some shortcomings (arbitrary weighting of items, inclusion of items of dubious functional significance). However, our objective was not to validate the Mathew scale but rather to assess interobserver variability. Despite the shortcomings of the Mathew scale, we still believe that the assessment of interobserver agreement on the components of the Mathew scale by κ statistics is useful.

κ indicates low interobserver agreement for those items for which the investigators must rely on subjective information. This is particularly true for the items orientation, homonymous hemianopsia, and sensation. On the other hand, for eight items interobserver agreement was significantly higher than that expected by chance. Although ours was a study with few patients, the results indicate that, in this area, quantitative assessment is very difficult. Although the multicenter study on acute ischemic stroke has already been finished with the Mathew scale as an assessment tool, it seems worthwhile to modify the scale in such a way that items having a low κ are avoided for future clinical research.

Acknowledgments

The authors would like to thank D. Norbruis, MD, PhD, head of the Department of Rehabilitation and Long-term Care of the "Meulenbelt" nursing home, Almelo, The Netherlands, for recruiting the patients in this study. We also would like to thank P.M. North, PhD, University of Kent at Canterbury, for his helpful discussions about the statistical aspects and his comments on the text of this article.

References

1. Sisk C, Ziegler DK, Zileli T: Discrepancies in recorded results from duplicate neurological history and examination in patients studied for the prognosis in cerebral vascular disease. *Stroke* 1970;1:14–18
2. Tomasello F, Mariani F, Fieschi C, Argentino C, Bono G, De Zanche L, Inzitari D, Martini A, Perrone P, Sangiovanni G: Assessment of inter-observer differences in the Italian Multicenter Study on reversible ischemia. *Stroke* 1982;13:32–35
3. Lindsay KW, Teasdale GM, Knill-Jones RP: Observer variability in assessing the clinical features of subarachnoid hemorrhage. *J Neurosurg* 1983;58:57–62
4. Shinar D, Gross CR, Mohr JP, Caplan LR, Price TR, Wolf PA, Hier DB, Kase CS, Fishman JG, Wolf CL, Kunitz SC: Interobserver variability in the assessment of neurologic history and examination in the Stroke Data Bank. *Arch Neurol* 1985;42:557–565
5. Mathew NT, Meyer JS, Rivera VH: Double blind evaluation of glycerol in acute cerebral infarction. *Lancet* 1972;2:1327–1333
6. Gelmers HJ: Effect of glycerol treatment on the natural history of acute cerebral infarction. *Clin Neurol Neurosurg* 1975; 78:277–282
7. Fleiss JL: Measuring nominal scale agreement among many raters. *Psychol Bull* 1971;76:378–382
8. Fleiss JL, Nee JCM, Landis JR: Large sample variance of kappa in the case of different sets of raters. *Psychol Bull* 1979; 86:974–977
9. van den Berge JH, Schouten HJA, Boomstra S, van Drunen Littel S, Braakman R: Interobserver agreement in the assessment of ocular signs in coma. *J Neurol Neurosurg Psychiatry* 1979;42:1163–1168
10. Theodossi A, Skene AM, Portmann B, Knill-Jones RP, Patrick RS, Tate RA, Kealey W, Jarvis KJ, O'Brian DJ, Williams R: Observer variation in assessment of liver biopsies including analysis by kappa statistics. *Gastroenterology* 1980;79: 232–241
11. Landis JR, Koch GG: The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–174

KEY WORDS • cerebral ischemia • cerebrovascular disorders • Netherlands

Downloaded from http://stroke.ahajournals.org/ by guest on July 19, 2018

Stroke
Journal of the American Heart Association

Assessment of interobserver variability in a Dutch multicenter study on acute ischemic stroke.

H J Gelmers, K Gorter, C J de Weerdt and H J Wiezer

Stroke. 1988;19:709-711
doi: 10.1161/01.STR.19.6.709
Stroke is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 1988 American Heart Association, Inc. All rights reserved.
Print ISSN: 0039-2499. Online ISSN: 1524-4628

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://stroke.ahajournals.org/content/19/6/709

Downloaded from http://stroke.ahajournals.org/ by guest on July 19, 2018

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in *Stroke* can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to *Stroke* is online at:
http://stroke.ahajournals.org//subscriptions/