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 \( \kappa \) 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)

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 due to acute ischemic stroke 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.

The neurologic deficits were assessed using the single items of an ordinal scale developed by Mathew et al., 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 \( \kappa \) 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. \( \kappa \) 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 \( (P_e = 1) \) and there is some variation in the patients’ neurologic status \( (P_o < 1) \). \( \kappa \) is undefined if \( P_e = 1 \). The significance of \( \kappa \) is tested by dividing it by its asymptotic standard error \( (S_\kappa) \). This ratio is asymptotically distributed as a
standard normal variable, \( S \), 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

\( P \), \( \kappa \), and corresponding significance levels are listed in Figure 4 for all the neurologic items observed. \( \kappa \) showed considerable variation, that is, there was a different \( \kappa \) for each item. Using the criteria defined in the legend of Figure 4, the observed \( \kappa \) 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 \( \kappa \) 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, \( \kappa \) 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.

### Discussion

Our study confirms the results from other studies\(^1\); 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 \( \kappa \) of 0.64, while the assessment of pupillary equality had a \( \kappa \) of 0.61.\(^6\) 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.
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 \( \kappa \) are avoided for future clinical research.

Acknowledgments

The authors would like to thank D. Norbruin, 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


Key Words • cerebral ischemia • cerebrovascular disorders • Netherlands
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

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