01 Jan 1985

Interobserver Variability in the Assessment of Neurologic History and Examination in the Stroke Data Bank

David Shinar

Cynthia R. Gross

Jay P. Mohr

Louis R. Caplan

et. al. For a complete list of authors, see https://scholarsmine.mst.edu/chem_facwork/3336

Recommended Citation

D. Shinar and C. R. Gross and J. P. Mohr and L. R. Caplan and T. R. Price and P. A. Wolf and D. B. Hier and C. S. Kase and I. G. Fishman and C. L. Wolf and S. C. Kunitz, "Interobserver Variability in the Assessment of Neurologic History and Examination in the Stroke Data Bank," *Archives of Neurology*, vol. 42, no. 6, pp. 557 - 565, JAMA Neurology, Jan 1985.
The definitive version is available at https://doi.org/10.1001/archneur.1985.04060060059010

Interobserver Variability in the Assessment of Neurologic History and Examination in the Stroke Data Bank

David Shinar, PhD; Cynthia R. Gross, PhD; Jay P. Mohr, MD; Louis R. Caplan, MD; Thomas R. Price, MD; Philip A. Wolf, MD; Daniel B. Hier, MD; Carlos S. Kase, MD; Irene G. Fishman, MAT; Christine L. Wolf; Selma C. Kunitz, MS

• **Interobserver reliability in obtaining neurologic histories and examinations was investigated among neurologists collaborating in the Stroke Data Bank (SDB). Seventeen in-hospital stroke patients were examined by six neurologists experienced in stroke over the course of three days. Patients were examined twice a day for two successive days, with each patient seen by four different neurologists. Data were recorded on SDB forms, according to definitions and procedures established for the SDB. Percent agreement and κ coefficients were calculated to assess the levels of agreement for each item. Important differences in levels of agreement were found among items on both neurologic history and examination. Agreement among neurologists was higher for neurologic examination than for history. Patterns of agreement for items with low prevalence or with numerous unknown ratings are discussed. Improvement in interobserver agreement due to data editing for intra-observer consistency was shown.**

(*Arch Neurol* 1985;42:557-565)

The present study was an evaluation of interobserver variation in obtaining the neurologic history and examination information in the Stroke Data Bank (SDB). The centers participating in this study are New York Neurological Institute, and the Departments of Neurology at University of Maryland Hospital, Baltimore, Boston University Medical Center, and Michael Reese Hospital, Chicago. A detailed description of the SDB has been published elsewhere.[1]

The importance of measuring interobserver variability (or inversely, interobserver agreement) is obvious when a research effort involves data collection in several centers with different observers. The key question is, "To what extent do different observers perceive and record the same information when confronted with the same phenomena?" While numerous studies have addressed the problem,[2-8] few have addressed neurologic signs and symptoms in stroke patients.[9,10] A consistent finding is the researchers' surprise to find considerable interobserver variability. Yet Garland almost 25 years ago stated that "a surprising and clinically important degree of inaccuracy is to be expected in the interpretation or evaluation of many clinical and laboratory procedures used in every day practice."[11]

In stroke research, Sisk et al[9] focused on the variation between two staff neurologists who examined 28 patients previously diagnosed as having "probable transient cerebral ischemic attacks or cerebral infarction with minimal residual." They developed a standardized protocol containing 20 neurologic symptoms and 32 items related to the neurologic examination. A manual was written for this study containing the definitions of the terms, the questions, and the response options for both the history and neurologic examination. In the authors' conclusion there were "extreme discrepancies in recording either presence or absence of specific symptoms and of signs on the neurologic examination." These discrepancies existed both for the subjective items such as weakness and the seemingly objective items such as reflex asymmetry.

In the Italian Multi-Center Study on Reversible Cerebral Ischemia, neurologists from eight clinical centers participated in data collection using a uniform protocol developed as a guideline for all examining neurolo-

Accepted for publication June 12, 1984.

From the Ben Gurion University of the Negev, Beer-Sheva, Israel (Dr Shinar); Biometry and Field Studies Branch, NINCDS, National Institutes of Health, Bethesda, Md (Dr Gross and Mss Fishman, Wolf, and Kunitz); New York Neurological Institute, Columbia University, New York (Dr Mohr); Department of Neurology, Michael Reese Hospital and Medical Center, Chicago (Drs Caplan and Hier); Department of Neurology, University of Maryland Hospital and Medical Center, Baltimore (Dr Price); and Department of Neurology, Boston University Medical Center (Drs Wolf and Kase).

Reprint requests to Office of Biometry and Field Studies, Federal Building, 7550 Wisconsin Ave, Bethesda, MD 20205 (Dr Gross).

gists.[10] They found large differences in their index of agreement as a function of the particular item evaluated. The index of crude agreement on neurologic signs varied from 21% (extensor plantar response) to 92% (visual field defects). The general finding was that interobserver discrepancies were a significant problem.

The results of these two studies demonstrate that the development of uniform protocols and common definitions does not automatically assure high interobserver agreement. Each collaborative study must assess its interobserver variability so that data collected in that particular study may be critically evaluated, reviewed to determine potential sources for the interobserver discrepancies, and future efforts can be directed at reducing this variability.

METHODS

Participants

The 17 subjects, ranging in age from 36 to 89, were in-hospital stroke patients at the New York Neurological Institute, who agreed to participate in the study, and were considered to be in a stable state. They had not been identified previously as having an evolving ischemic attack (Table 1).

The observers in the study were the six staff neurologists from the four SDB-participating centers who are directly responsible for data collection in their own centers. All were experienced in clinical neurology with a special interest in stroke.

Procedure and Design

The patient and observer pairings were rotated over the course of three days. Each patient was evaluated four times by four different neurologists as follows: once in the morning and once in the afternoon of two days, in accordance with the counterbalanced design reproduced in Table 1. This design yielded a total of 34 pairs of observations conducted on 17 patients, each pair consisting of two independent observations made by two neurologists on the same patient within the same day, so that day-to-day variations would not be confounded with interobserver differences.

The neurologists did not discuss these cases with each other. At the beginning of each morning and afternoon session each neurologist was given a folder containing forms for the patients assigned for that session only. Patient forms were filled out immediately after the evaluation and all the completed forms were returned at the end of the session.

The neurologists were provided with limited information on each patient as follows: name, age, hospital admission date, associated medical illnesses, and, in the case of aphasic and comatose patients, the onset and early course of illness.

Table 1.—The Design of Patient-Neurologist Assignments*

Patient	Communication Problems	Wednesday AM	Wednesday PM	Thursday AM	Thursday PM	Friday AM	Friday PM	Total No. of Observations
1	. . .	. . .	. . .	N5	N6	N1	N2	4
2	Severe aphasia	. . .	. . .	N4	N5	N3	N1	4
3	. . .	. . .	. . .	N5	N2	N4	N1	4
4	Spanish speaking, dysarthria	N5	N1	N2	N3	. . .	. . .	4
5	Comatose	N6	N1	N4	N2	. . .	. . .	4
6	Irritable, uncooperative	N2	N3	N6	N1	. . .	. . .	4
7	Dysarthria	N2	N4	N6	N3	. . .	. . .	4
8	Sparse speech, poor comprehension	N3	N5	N4	N1	. . .	. . .	4
9	Severe aphasia	N6	N2	N3	N5	. . .	. . .	4
10	. . .	N3	N4	N2	N6	. . .	. . .	4
11	. . .	N2	N5	N6	N4	. . .	. . .	4
12	Dysarthria	N3	N6	N1	N2	. . .	. . .	4
13	. . .	N4	N5	N1	N3	. . .	. . .	4
14	Mutism	N4	N6	N5	N1	. . .	. . .	4
15	Confused, fluctuating mental status/SAH†	N5	N6	N3	N4	. . .	. . .	4
16	. . .	. . .	. . .	N2	N4	N6	N3	4
17	Slurred speech	. . .	. . .	N3	N6	N2	N1	4

*Neurologists are identified by the letter N.
†SAH indicates subarachnoid hemorrhages.

Forms and Materials

The SDB standard data forms were used for data recording (see Figs 1 to 4). Instructions for data collection and item definitions were provided by the SDB coding and operations manuals. In addition, each observer provided his subjective ratings of the "difficulty of evaluating the patient" on a scale of one to five, and noted subjective comments and observations. These comments were to be used later in a discussion focused on means of reducing interobserver variability.

Statistical Methods

The κ statistic used to measure the level of agreement was based on a formula developed by Fleiss,[12] which provides a numerical measure of agreement among multiple raters on variables that are scored on a nominal scale. The κ statistic is chance-corrected, ie, it measures the observed amount of agreement adjusted for the amount of agreement expected by chance alone. κ approaches -1 for complete disagreement and $+1$ for perfect agreement. When the agreement is that expected by chance, κ equals 0. The significance of κ is tested by dividing it by its standard error.[13] This ratio is distributed as a standard normal variate. It has been suggested that whenever κ is $>.80$ the agreement can be considered excellent; $0.40 < \kappa \leq .80$ indicates moderate to substantial agreement; $0.20 < \kappa \leq .40$ indicates fair agreement; and $\kappa \leq .20$ indicates slight or poor agreement.[7,14]

RESULTS

The results on the neurologic history items and on the neurologic examination items are presented separately.

Neurologic History

During two successive days, four neurologic history assessments were made on each patient. It was assumed that day-to-day variations would have no influence on the neurologic history, unlike potential fluctuations on the neurologic examination. The analysis was conducted on the four sets of data obtained for each patient rather than on the pairs of observations made the same day.

The raw data to assess the extent of agreement on item 8N—"Has Patient Ever Had a Stroke Before This One?" contains a total of 68 observations, 17 patients times four neurologic evaluations (Table 2). In 45 (66%) of 68 evaluations, no previous stroke was reported. In all cases in which all four neurologists agreed, it was in the "no" category. Most of the disagreements were due to the inability of one or more neurologists to determine the answer to this question, reflected by the "unknown" responses (Table 2). In 35% of the cases there was perfect agreement among the four neurologists (six of 17 cases) and in 82% of the cases at least three of four neurologists agreed. The κ value of 0.31 ($P < .001$) indicated a fair amount of agreement. The extent of agreement would not be improved by collapsing the "yes" categories into one category since all the discrepancies were due to differences in opinion about prior stroke rather than disagreements in the time elapsed since the last stroke. However, if the question is rephrased

Table 2.—The Response Pattern for Variable 8N: Has Patient Ever Had a Stroke Before This One?

Patient	No	Yes 1-7 Days Ago	Yes 8-30 Days Ago	Yes 1-6 Months Ago	Yes Over 6 Months Ago	Unknown	Total
1	4	...	...	...	...	...	4
2	3	...	...	...	...	1	4
3	2	...	...	2	...	...	4
4	4	...	...	...	...	...	4
5	2	...	...	...	...	2	4
6	4	...	...	...	...	...	4
7	1	...	...	...	3	...	4
8	3	...	...	...	...	1	4
9	1	...	...	...	...	3	4
10	4	...	...	...	...	...	4
11	4	...	...	...	...	...	4
12	4	...	...	...	...	...	4
13	1	...	...	...	3	...	4
14	2	...	...	...	...	2	4
15	3	...	...	...	...	1	4
16	3	...	...	1	...	...	4
17	...	...	...	...	1	3	4
Total	**45**	**0**	**0**	**3**	**7**	**13**	**68**

Table 3.—The Response Pattern for Variable 4N: Has Patient Ever Had a TIA?*

Patient	No, Never	Yes 1-7 Days Ago	Yes 8-30 Days Ago	Yes 1-6 Months Ago	Yes Over 6 Months Ago	Unknown	Total
1	4	...	...	...	...	...	4
2	2	...	...	...	...	2	4
3	2	...	...	2	...	...	4
4	1	2	1	...	...	...	4
5	1	1	1	...	...	1	4
6	4	...	...	...	...	...	4
7	2	1	...	...	...	1	4
8	2	...	...	...	...	2	4
9	1	...	...	...	...	3	4
10	2	...	...	...	...	2	4
11	4	...	...	...	...	...	4
12	4	...	...	...	...	...	4
13	4	...	...	...	...	...	4
14	2	...	...	...	...	2	4
15	2	...	...	...	...	2	4
16	3	...	...	...	1	...	4
17	2	...	...	...	...	2	4
Total	**42**	**4**	**2**	**2**	**1**	**17**	**68**

*TIA indicates transient ischemic attack.

to detect the ability to elicit a history of prior stroke, and "unknown" and "no" responses combined, then complete agreements would be 71% and the κ coefficient 0.40 ($P < .001$).

The response patterns for two more Neurologic History variables, 4N—"Has Patient Ever Had a TIA?" and 13N—"Was There Severe Headache at the Time of Onset?," are presented in Tables 3 and 4, respectively. For five (ie, 29%) of 17 patients there was perfect agreement on the occurrence of a transient ischemic attack (TIA); and in 41% of the cases three or more neurologists agreed on the occurrence of a TIA. Again, all of the perfect agreements were obtained for the null category. The κ coefficient for this variable was 0.11, indicating a level of agreement barely greater than chance. As in the case of a previous stroke, when the "unknown" and "no" responses were pooled into one category and all the "yes" responses were pooled into another, the κ coefficient rose to 0.19 ($P < .01$). A similar pattern can be observed in Table 4 for the variable of presence or absence of severe headache at stroke onset. Perfect agreement was obtained for seven of the 17 patients, ie, 41% agreement and $\kappa = 0.36$, ie, fair agreement. When "unknown" and "no" responses were combined, the κ coefficient rose to 0.52 ($P < .001$).

Table 4.—The Response Pattern for Variable 13N: Was There a Severe Headache at the Time of Onset?

Patient	No	Yes	Unknown	Total
1	4	...	...	4
2	2	...	2	4
3	4	...	...	4
4	3	1	...	4
5	1	...	3	4
6	4	...	...	4
7	4	...	...	4
8	1	3	...	4
9	1	...	3	4
10	1	1	2	4
11	4	...	...	4
12	4	...	...	4
13	2	...	2	4
14	1	3	...	4
15	...	4	...	4
16	2	1	1	4
17	2	...	2	4
Total	**40**	**13**	**15**	**68**

The history variables (Table 5) are ordered by their κ values, from the most consistently to the least consistently scored. Because the κ coefficient is chance-corrected, it is a more conservative measure of agreement than the percent of complete agreements. Thus, an agreement on presence of a particular deficit is typically weighted more heavily than an agreement in the null category since the null category is the most common response. It can be seen that there are large variations in interobserver reliability among the variables studied; moderate agreement was achieved for only two variables (alcohol intake within 24 hours of onset and the last glucogenic intake), fair agreement for an additional nine variables, and only chance agreement for six of the variables studied.

Neurologic Examination

To minimize the influence of true day-to-day changes in the patients' conditions on interobserver agreement, the comparisons reported are based on observations made within the same day. Thus, instead of looking at 17 patients each observed four times, the analysis was based on 34 pairs of observations, each pair of observations obtained on the same day.

Systematic changes in the patients' conditions either from morning to afternoon or from the first to the second day might confound the study of interobserver agreement. Such a trend either in improvement or worsening would have caused a spurious reduction in the level of agreement among observers, a reduction due to changes in actual patient condition

Table 5.—Level of Agreement on Neurologic History Variables

Variable No.	Variable Name	Before Consistency Checks: κ	Before Consistency Checks: % Complete Agreement
30N	Alcohol within 24 hr of onset	0.65*	65
31N	Last glucogenic intake	0.46†	35
15N	Seizures at onset	0.39*	47
17N	Decreased consciousness at onset	0.36*	41
13N	Severe headache at onset	0.36*	53
14N	Vomiting at onset	0.35*	35
21N	Course of illness 1-12 hr of onset	0.34*	29
18N	Coma at onset	0.32*	47
8N	Previous stroke	0.31*	35
22N	Course of illness 12-24 hr of onset	0.30*	35
20N	Course of illness 11-60 min of onset	0.22*	18
19N	Course of illness 1-10 min of onset	0.17‡	18
16N	Focal deficit at onset	0.15§	35
29N	Documented hypotension	0.12	24
4N	Previous transient ischemic attacks	0.11	29
12N	Deficit present on awakening	0.11	24
27N	Antiplatelet/anticoagulant used	0.08	29

*Significant at $P < .001$.
†Intraclass correlation coefficient based on only eight patients, $P < .05$.
‡Significant at $P < .01$.
§Significant at $P < .05$.

rather than to interobserver variations. To test for this, a two-way analysis of variance was conducted on the mean performance levels of the ordinal variables (19x, 26x, 49x, 56x, 62x, 68x, and 72x). There was no morning-afternoon or day 1-day 2 effect for these variables. Thus, no systematic improvements or deteriorations in neurologic status across patients were seen from morning to afternoon or from day 1 to day 2. Accordingly, the data were analyzed using four observations per patient. The results indicated no significant difference between the two approaches (average κ being 0.36 and 0.37 for pairs and foursomes, respectively), and the results presented below are for pairs of observations made on the same day.

Three data patterns emerged from the analysis. The first is illustrated in Table 6 with each pair of observations on a patient on the same day represented by the single-cell entry. The morning responses are plotted against the afternoon responses. Whenever the two neurologists agree, their observation appears on the main diagonal. For the variable X5—verbal response, there are a total of 21 agreements, of which 12 were on category five which is the "no deficit" category. The extent of disagreements is represented by the distance from the diagonal. Thus, excluding the situations in which one of the neurologists coded that item as "unknown," all the disagreements were by one level only, ie, one cell away from the diagonal. In summary, when the unknown responses are considered as a legitimate code, the agreement is 62% and $\kappa = 0.47$; when unknown responses are viewed as missing data, the agreement is 69% (18 of 26 pairs) and $\kappa = 0.49$.

In a second kind of data pattern the distance from the diagonal is meaningless since the categories are on a nominal scale as in the distribution of response pairs for the variable 14X—"Weakness" (Table 7). For identification of the location of weakness, there is 79% agreement, $\kappa = 0.67$. No unknown responses were coded. All but two of the disagreements occurred when one neurologist identified a weakness on one side and the other neurologist identified the weakness on that same side but also on the other side (ie, bilateral hemiparesis). There were no disagreements as to the major side of weakness. A similar pattern of responses, but with a lower level of agreement, was obtained for variable 45X—"Sensory Deficits." Here agreement was 50%, $\kappa = 0.32$. When the "Untestable" responses were excluded, agreement increased to 55%, and $\kappa = 0.35$. None of the discrepancies were due to confusions of laterality, but rather to the identification of the location of the sensory deficit by one neurologist and the code of "none" or "Untestable" by the other neurologist.

In the third pattern the data are again nominal (Table 8 [71X—"Language Abnormalities"]). However, here most of the disagreements are not due to "unknown" but to disagreements as to the specific language abnormality. Here improvement in consistency among observers requires either some merging of overlapping or ambiguous categories, or providing mutually exhaustive operational definitions for each term. Thus, when "Language" was collapsed into four categories, normal, global aphasia, other, and unknown, interobserver agreement rose from $\kappa = 0.54$ to $\kappa = 0.68$.

Tables 6 through 8 also illustrate the statistical observation that there was no consistent bias between morning and afternoon evaluations, manifested by the similar distributions of disagreements above and below the diagonals. If, for example, patients tended to perform better in the afternoon, then most of the disagreements would have been below the diagonal in Table 6, and below or above the diagonal in a consistent manner for some disagreements in Tables 7 and 8. "Folding" the table at the diagonal provides a more direct appreciation of the number and type of disagreements. To illustrate, in Table 8 by folding the table it could be observed that Wernicke and Broca aphasia were paired twice and global aphasia and Broca aphasia were also paired twice.

The extent of agreement and κ coefficients on all the neurologic examination variables on which the interobserver consistency can be compared is shown in Table 9. This table cannot provide insights to improve interobserver reliability. For this, one must examine the actual distribution of responses for each variable as they were presented (Tables 6 through 8). Nonetheless, the intuitive meaning of the level of agreement on all of the variables listed in Table 9 can be grasped by comparing the percent agreement and κ statistics of these variables against the ones displayed in Tables 6 through 8, ie, verbal response, weakness, and language. Table 9 shows that, in general, perfor-

Table 6.—Distribution of Response Pairs for Variable 5X: Verbal Response

AM	PM None	Incomprehensible Sounds	Inappropriate Words	Disoriented	Oriented and Converses	Untestable	Total
None	...	2	...	...	...	...	2
Incomprehensible sounds	...	2*	...	...	...	1	3
Inappropriate words	...	...	...	1	...	1	2
Disoriented	...	...	...	4*	3	...	7
Oriented and converses	...	...	...	2	12*	...	14
Untestable	...	1	1	...	1	3*	6
Total	**0**	**5**	**1**	**7**	**16**	**5**	**34**

*Agreeing pairs.

Table 7.—Distribution of Response Pairs for Variable 14X: Weakness

AM	PM Normal	Left Hemiparesis	Right Hemiparesis	Bilateral Hemiparesis	Paraparesis	Unknown	Total
Normal	...	1	...	...	...	...	1
Left hemiparesis	...	12*	...	...	...	...	12
Right hemiparesis	1	...	13*	2	...	...	16
Bilateral hemiparesis	...	2	...	2*	...	...	4
Paraparesis	...	...	1	...	...	...	1
Unknown	...	...	...	...	...	...	0
Total	**1**	**15**	**14**	**4**	**0**	**0**	**34**

*Agreeing pairs.

Table 8.—Distribution of Response Pairs for Variable 71X: Language

AM	PM Normal	Broca	Wernicke	Global	Anomic	Other	Unknown	Total
Normal	15*	...	...	...	...	1	1	17
Broca	...	...	1	1	...	1	...	3
Wernicke	...	1	...	...	...	1	...	2
Global	...	1	...	4*	...	...	...	5
Anomic	...	...	...	...	2*	...	...	2
Other	1	...	...	...	1	...	...	2
Unknown	1	...	...	...	...	...	2*	3
Total	**17**	**2**	**1**	**5**	**3**	**3**	**3**	**34**

*Agreeing pairs.

mance was better on the neurologic examination variables than on the neurologic history variables. This could have been expected since the neurologic history was based solely on the interview with the patient (many of whom were difficult to communicate with, as indicated in Table 1), whereas the neurologic examination was based on actual observations and testing. As was the case with the neurologic history variables, the variation in the level of agreement is impressive. For more than half of the variables evaluated the level of agreement was moderate to substantial, but for ten of 47 variables in Table 9 it was no better than chance.

Some variables were of particular interest. On the weakness scale the average κ for the left side of the body equaled that of the right side of the body (0.45 and 0.46). In contrast, for the sensory scale the average κ on the left side was much poorer than on the right side (0.33 v 0.52). Furthermore, there were some consistent patterns with respect to the κ values for different body parts. Evaluation of weakness of the tongue was significantly lower than that of the other body parts and agreement on the evaluations of the face, hand, and foot were consistently higher than the agreement on the hip or shoulder.

A note of caution is appropriate in interpreting the meaning of "agreement" for some of the variables at the bottom of the list in Table 9. As may be observed, the percent agreement is extremely high for these variables while the κ value is essentially zero (eg, nuchal rigidity, neurologic symptoms, and cervical bruit). This is due to the distribution of the responses. For each of these items in most of the 68 examinations the evaluation was "absence of deficit," and in the very few cases where a deficit was observed by one or more of the neurologists it was not corroborated by the paired examiner. This indicates that when such a deficit exists infrequently its recording is unreliable. Since for most neurologic deficits, the probability of any specific deficit occurring for any given patient is relatively low, a statistic such as "percent agreement" should be interpreted very cautiously. To illustrate, in the case of nuchal rigidity an even greater appearance of interobserver reliability would have been obtained had none of the neurologists coded its presence. In that case, κ would have been undefined and the percent agreement would have been 100. However, when one or two of the examiners do code its presence, if they do not agree, then the percent agreement remains high but that observed agreement may not be above chance as indicated by the κ coefficient.

Consistency Checks

As part of the standard SDB procedures, all data are computer-checked for their consistency before they are entered as part of the data bank. The purpose of these checks is to verify that the responses to different items are not inconsistent with each other. To illustrate, one consistency check specifies that if a neurologist codes weakness (14X) as being only on the

Table 9.—Level of Agreement on Neurologic Examination Variables

Variable No.	Variable Name	Pairwise Agreements: Before Consistency Checks κ	Before Consistency Checks Agreement	After Consistency Checks* κ	After Consistency Checks* Agreement
80X	Neurologic signs	Undefined	100	...	...
34X	Extraocular movements	0.77†	94	...	...
33X	Swallowing	0.74†	88	...	...
32X	Articulation	0.68†	79	0.62†	76
14X	Weakness (general location)	0.67†	79	0.68†	79
17X	Weakness scale-left, face	0.66†	79	0.70†	82
55X	Sensory scale-right, shoulder	0.60†	79	...	...
26X	Weakness scale-right, hand	0.58†	74	0.59†	74
54X	Sensory scale-right, face	0.58†	79	...	...
28X	Weakness scale-right, foot	0.54†	71	0.60†	74
71X	Language	0.54†	68	...	...
72X	Dysarthria	0.53†	74	0.59†	76
21X	Weakness scale-left, foot	0.52†	71	0.56†	74
24X	Weakness scale-right, face	0.51†	74	0.62†	79
57X	Sensory scale-right, hip	0.50†	76	...	...
59X	Sensory scale-right, trunk	0.50†	76	...	...
70X	Speech content	0.50†	68	...	...
56X	Sensory scale-right, hand	0.50†	74	...	...
19X	Weakness scale-left, hand	0.49†	68	0.53†	71
20X	Weakness scale-left, hip	0.49†	68	0.52†	71
25X	Weakness scale-right, shoulder	0.47†	65	0.48†	65
5X	Verbal response	0.47†	62	0.50†	65
27X	Weakness scale-right, hip	0.46†	65	0.52†	68
68X	Other cognitive functions	0.46†	68	0.41‡	65
31X	Ataxia	0.45†	76	...	...
58X	Sensory scale-right, foot	0.44†	71	...	...
62X	Visual fields	0.40‡	68	...	...
9X	Degree of alertness	0.38†	85	...	...
18X	Weakness scale-left, shoulder	0.36‡	59	0.40†	62
52X	Sensory scale-left, trunk	0.36‡	68	0.40†	71
47X	Sensory scale-left, face	0.36‡	62	0.40†	65
83X	Examiner believes patient is demented	0.34‡	68	...	...
50X	Sensory scale-left, hip	0.34‡	65	0.38‡	68
49X	Sensory scale-left, hand	0.32‡	59	0.36†	62
51X	Sensory scale-left, foot	0.32‡	65	0.36‡	68
45X	Sensory deficits	0.32‡	50	0.36†	53
48X	Sensory scale-left, shoulder	0.28‡	59	0.31‡	62
16X	Weakness scale-left, tongue	0.20	59	0.23	62
23X	Weakness scale-right, tongue	0.17	71	0.14	68
82X	Examiner believes patient is depressed	0.15	56	0.18	59
7X	Motor response	0.08	74	...	...
6X	Eye opening	0.02	76	0.05	79
73X	Nuchal rigidity	−0.01	97	...	...
81X	Neurologic symptoms	−0.01	97	Undefined	100
74X	Cervical bruits	−0.05	88	−0.05	91
79X	Pure motor syndrome	−0.08	85	0.10	79
11X	Remainder of neurologic examination	−0.15	74	...	...

*Variables without entries in these columns were unchanged by the consistency checks.
†$P < .001$.
‡$P < .01$.

right side (14X = 2) then there should be no weakness coded for any parts of the left side of the body (16X − 21X = 0) and at least one of the parts on the right side of the body should be coded as having a weakness (one of the variables 23X − 28X must be >0). These consistency checks are useful in detecting data entry, coding, and transmission errors. It was of interest, therefore, to assess the value of these checks as a putative means of raising consistency among observers. Of the numerous consistency checks that are applied to the neurologic history and neurologic examination variables in the SDB, 27 were found to be relevant to the present study. One consistency check involved the neurologic history variables, and 26 consistency checks involved the neurologic examination variables.

Table 9 demonstrates the degree of improvement in interobserver variability that is achieved by the routine application of the consistency checks to the neurologic examination raw data prior to inclusion in the data bank. Only the variables affected by the consistency checks are noted in the last two columns of this table. As can be readily observed, these intraobserver consistency checks and the resultant corrections tended to improve the interobserver consistency, yielding a slight increase in the κ coefficient in 24 of 29 neurologic examination variables affected. In the neurologic history, the κ coefficient for the one variable affected, 16N—"Focal Deficit at Onset," increased dramatically from 0.15 to 0.35 ($P < .001$).

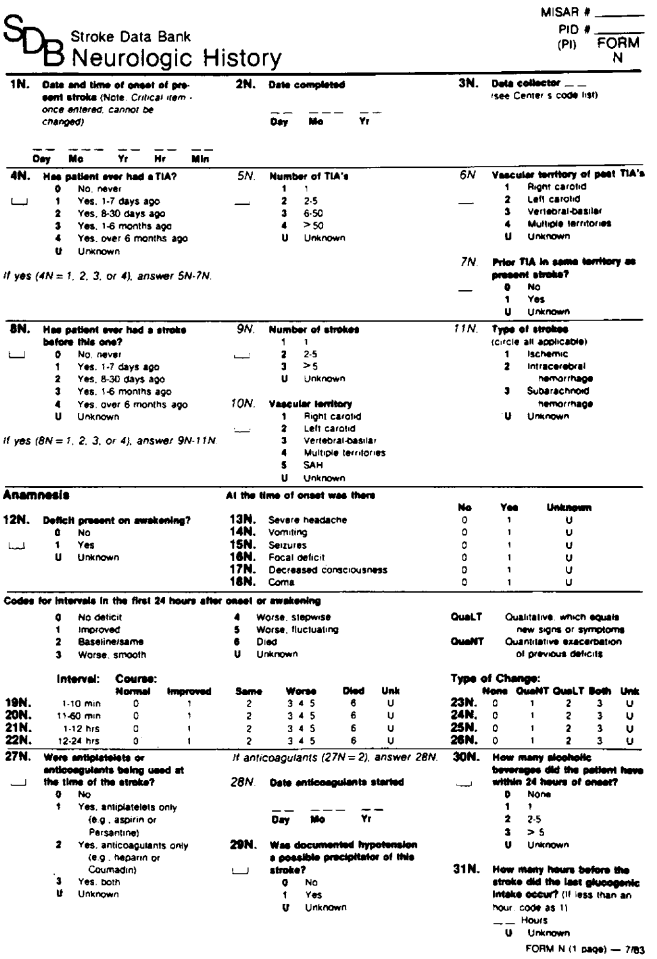
Agreements and Perceived Difficulty

It was hypothesized that the level of interobserver agreement would correspond to the neurologists' perceived difficulty in making their assessments. After each examination, the neurologists rated the examination and history separately on a scale of one to five from easy to difficult. Accordingly, for each patient the mean difficulty rating (among the four neurologists) was correlated with the observed percent agreement (based on the average of the same four examining neurologists across all items). This was done separately for the neurologic history and the neurologic examination. As expected, significant correlations were obtained for both assessments as follows: Spearman $\rho = -0.45$ ($P < .05$), for the neurologic history, and −0.69 ($P < .01$) for the neurologic examination. Thus, as the perceived difficulty of assessment decreased, the level of agreement among the observed rose.

COMMENT

Interobserver Agreement and Variability in the Data Bank

Perhaps the first qualifications that should be noted in interpreting these results are some basic differences that exist between the data collection effort used in the present study and the nature of the activities involved in assessment of the neurologic history and neurologic examina-



SDB Stroke Data Bank
Neurologic History

MISAR # _____
PID # _____
(PI) FORM N

1N. Date and time of onset of present stroke (Note: Critical item - once entered, cannot be changed)
Day Mo Yr Hr Min

2N. Date completed
Day Mo Yr

3N. Data collector __ (see Center's code list)

4N. Has patient ever had a TIA?
0 No, never
1 Yes, 1-7 days ago
2 Yes, 8-30 days ago
3 Yes, 1-6 months ago
4 Yes, over 6 months ago
U Unknown

If yes (4N = 1, 2, 3, or 4), answer 5N-7N.

5N. **Number of TIA's**
1 1
2 2-5
3 6-50
4 >50
U Unknown

6N. **Vascular territory of past TIA's**
1 Right carotid
2 Left carotid
3 Vertebral-basilar
4 Multiple territories
U Unknown

7N. **Prior TIA in same territory as present stroke?**
0 No
1 Yes
U Unknown

8N. Has patient ever had a stroke before this one?
0 No, never
1 Yes, 1-7 days ago
2 Yes, 8-30 days ago
3 Yes, 1-6 months ago
4 Yes, over 6 months ago
U Unknown

If yes (8N = 1, 2, 3, or 4), answer 9N-11N.

9N. **Number of strokes**
1 1
2 2-5
3 >5
U Unknown

10N. **Vascular territory**
1 Right carotid
2 Left carotid
3 Vertebral-basilar
4 Multiple territories
5 SAH
U Unknown

11N. **Type of strokes** (circle all applicable)
1 Ischemic
2 Intracerebral hemorrhage
3 Subarachnoid hemorrhage
U Unknown

Anamnesis

12N. Deficit present on awakening?
0 No
1 Yes
U Unknown

At the time of onset was there

		No	Yes	Unknown
13N.	Severe headache	0	1	U
14N.	Vomiting	0	1	U
15N.	Seizures	0	1	U
16N.	Focal deficit	0	1	U
17N.	Decreased consciousness	0	1	U
18N.	Coma	0	1	U

Codes for intervals in the first 24 hours after onset or awakening
0 No deficit
1 Improved
2 Baseline/same
3 Worse, smooth
4 Worse, stepwise
5 Worse, fluctuating
6 Died
U Unknown

QualT Qualitative, which equals new signs or symptoms
QuaNT Quantitative exacerbation of previous deficits

	Interval:	Course: Normal	Improved	Same	Worse	Died	Unk
19N.	1-10 min	0	1	2	3 4 5	6	U
20N.	11-60 min	0	1	2	3 4 5	6	U
21N.	1-12 hrs	0	1	2	3 4 5	6	U
22N.	12-24 hrs	0	1	2	3 4 5	6	U

Type of Change:

	None	QuaNT	QuaLT	Both	Unk
23N.	0	1	2	3	U
24N.	0	1	2	3	U
25N.	0	1	2	3	U
26N.	0	1	2	3	U

27N. Were antiplatelets or anticoagulants being used at the time of the stroke?
0 No
1 Yes, antiplatelets only (e.g., aspirin or Persantine)
2 Yes, anticoagulants only (e.g., heparin or Coumadin)
3 Yes, both
U Unknown

If anticoagulants (27N = 2), answer 28N.

28N. **Date anticoagulants started**
Day Mo Yr

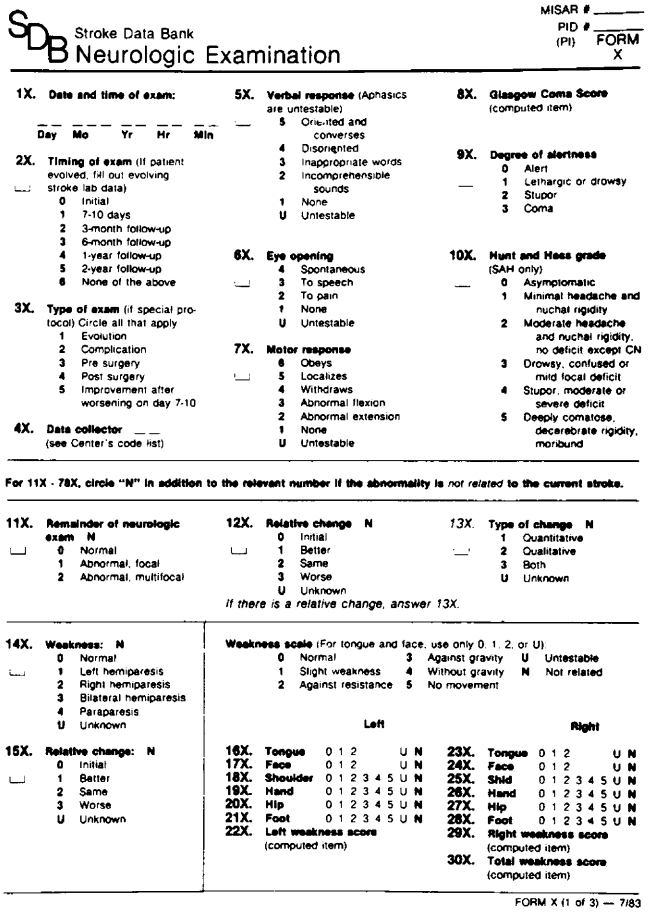
29N. Was documented hypotension a possible precipitator of this stroke?
0 No
1 Yes
U Unknown

30N. How many alcoholic beverages did the patient have within 24 hours of onset?
0 None
1 1
2 2-5
3 > 5
U Unknown

31N. How many hours before the stroke did the last glucogenic intake occur? (If less than an hour, code as 1)
__ Hours
U Unknown

FORM N (1 page) — 7/83

Fig 1.—Stroke Data Bank Neurologic History form.



SDB Stroke Data Bank
Neurologic Examination

MISAR # _____
PID # _____
(PI) FORM X

1X. Date and time of exam:
Day Mo Yr Hr Min

2X. Timing of exam (If patient evolved, fill out evolving stroke lab data)
0 Initial
1 7-10 days
2 3-month follow-up
3 6-month follow-up
4 1-year follow-up
5 2-year follow-up
6 None of the above

3X. Type of exam (If special protocol) Circle all that apply
1 Evolution
2 Complication
3 Pre surgery
4 Post surgery
5 Improvement after worsening on day 7-10

4X. Data collector __ (see Center's code list)

5X. Verbal response (Aphasics are untestable)
5 Oriented and converses
4 Disoriented
3 Inappropriate words
2 Incomprehensible sounds
1 None
U Untestable

6X. Eye opening
4 Spontaneous
3 To speech
2 To pain
1 None
U Untestable

7X. Motor response
6 Obeys
5 Localizes
4 Withdraws
3 Abnormal flexion
2 Abnormal extension
1 None
U Untestable

8X. Glasgow Coma Score (computed item)

9X. Degree of alertness
0 Alert
1 Lethargic or drowsy
2 Stupor
3 Coma

10X. Hunt and Hess grade (SAH only)
0 Asymptomatic
1 Minimal headache and nuchal rigidity
2 Moderate headache and nuchal rigidity, no deficit except CN
3 Drowsy, confused or mild focal deficit
4 Stupor, moderate or severe deficit
5 Deeply comatose, decerebrate rigidity, moribund

For 11X - 78X, circle "N" in addition to the relevant number if the abnormality is *not related* to the current stroke.

11X. Remainder of neurologic exam N
0 Normal
1 Abnormal, focal
2 Abnormal, multifocal

12X. Relative change N
0 Initial
1 Better
2 Same
3 Worse
U Unknown
If there is a relative change, answer 13X.

13X. **Type of change N**
1 Quantitative
2 Qualitative
3 Both
U Unknown

14X. Weakness: N
0 Normal
1 Left hemiparesis
2 Right hemiparesis
3 Bilateral hemiparesis
4 Paraparesis
U Unknown

15X. Relative change: N
0 Initial
1 Better
2 Same
3 Worse
U Unknown

Weakness scale (For tongue and face, use only 0, 1, 2, or U)
0 Normal
1 Slight weakness
2 Against resistance
3 Against gravity
4 Without gravity
5 No movement
U Untestable
N Not related

	Left		Right	
Tongue	16X. 0 1 2 U N		23X. 0 1 2 U N	
Face	17X. 0 1 2 U N		24X. 0 1 2 U N	
Shoulder	18X. 0 1 2 3 4 5 U N		25X. 0 1 2 3 4 5 U N	
Hand	19X. 0 1 2 3 4 5 U N		26X. 0 1 2 3 4 5 U N	
Hip	20X. 0 1 2 3 4 5 U N		27X. 0 1 2 3 4 5 U N	
Foot	21X. 0 1 2 3 4 5 U N		28X. 0 1 2 3 4 5 U N	
Left weakness score (computed item)	22X.		Right weakness score (computed item)	29X.
			Total weakness score (computed item)	30X.

FORM X (1 of 3) — 7/83

Fig 2.—Stroke Data Bank Neurologic History and Neurologic Examination form, page 1.

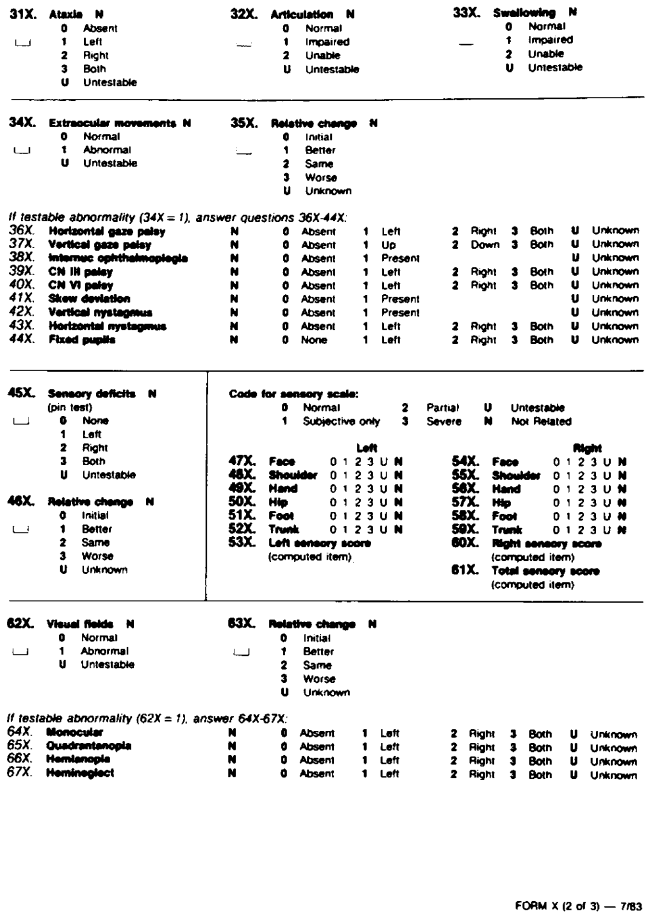
tion in the SDB. There are three significant differences as follows: (1) in the SDB the patient is seen on several occasions and on daily rounds so that a momentary impression gained in the course of the neurologic assessment is either supported or refuted by additional observations available to the neurologist; (2) the neurologist is exposed to more than the patient since he also has access to the chart, and talks to nurses and medical residents who see the patient, as well as friends and family of the patient. This is particularly important in the assessment of the neurologic history items in cases where the patient is not coherent. In the present study, the neurologists were not permitted to obtain data from anything other than a brief summary of the medical history written for this study and whatever information they could gain through their interview with the patient; (3) in the SDB many of the decisions are reached on the basis of "group" consensus since the patient is typically seen by more than just the SDB neurologist. Other participants include the stroke fellow(s) and the stroke research nurse.

The particular nature of the sample of patients studied was also important. For practical reasons an attempt was made to observe as many patients as possible at one center. This resulted in a sample with many patients (11 of 17) having communication difficulties and some who were, on occasion, uncooperative. It should be stressed, however, that these differences would not have necessarily affected the interobserver reliability in any consistent or predictable manner. They only relate to the validity of the data and to the extent that it is more valid in the SDB it would reflect better data quality. Nonetheless, the observed correlation between the interobserver level of agreement and the rated difficulty of assessment of each patient strongly supports the notion that a more communicative sample of patients would yield a higher level of agreement. Unfortunately, such a sample would probably be less representative of the in-hospital stroke patient population.

In summary, given the nature of the patient sample and the variations in their levels of cooperation, the levels of reliability obtained in this study are conservative: the reliability of the SDB can be assumed to be at least as high or higher.

Interpretation of the Discrepancies

The magnitude of interobserver variability should be defined in qualitative terms as well as quantitative terms such as the κ coefficient. The first point to note is that the percent agreement is consistently and markedly higher than the κ coefficient. However, percent agreement is a relatively noninformative measure of interobserver reliability for research purposes. κ coefficients essentially indicate the amount of agreement among neurologists that exists beyond chance. In the particular case of the SDB items, this is the extent of agreement when the neurologic deficit may be present. Furthermore, the κ coefficients should not be confused with a correlation coefficient (even though they range over the same val-



31X. Ataxia N: 0 Absent; 1 Left; 2 Right; 3 Both; U Untestable

32X. Articulation N: 0 Normal; 1 Impaired; 2 Unable; U Untestable

33X. Swallowing N: 0 Normal; 1 Impaired; 2 Unable; U Untestable

34X. Extraocular movements N: 0 Normal; 1 Abnormal; U Untestable

35X. Relative change N: 0 Initial; 1 Better; 2 Same; 3 Worse; U Unknown

If testable abnormality (34X = 1), answer questions 36X-44X:

36X. Horizontal gaze palsy	N	0 Absent	1 Left	2 Right	3 Both	U Unknown
37X. Vertical gaze palsy	N	0 Absent	1 Up	2 Down	3 Both	U Unknown
38X. Internuc ophthalmoplegia	N	0 Absent	1 Present			U Unknown
39X. CN III palsy	N	0 Absent	1 Left	2 Right	3 Both	U Unknown
40X. CN VI palsy	N	0 Absent	1 Left	2 Right	3 Both	U Unknown
41X. Skew deviation	N	0 Absent	1 Present			U Unknown
42X. Vertical nystagmus	N	0 Absent	1 Present			U Unknown
43X. Horizontal nystagmus	N	0 Absent	1 Left	2 Right	3 Both	U Unknown
44X. Fixed pupils	N	0 None	1 Left	2 Right	3 Both	U Unknown

45X. Sensory deficits N (pin test): 0 None; 1 Left; 2 Right; 3 Both; U Untestable

46X. Relative change N: 0 Initial; 1 Better; 2 Same; 3 Worse; U Unknown

Code for sensory scale: 0 Normal; 1 Subjective only; 2 Partial; 3 Severe; U Untestable; N Not Related

	Left		Right	
47X. Face	0 1 2 3 U N		54X. Face	0 1 2 3 U N
48X. Shoulder	0 1 2 3 U N		55X. Shoulder	0 1 2 3 U N
49X. Hand	0 1 2 3 U N		56X. Hand	0 1 2 3 U N
50X. Hip	0 1 2 3 U N		57X. Hip	0 1 2 3 U N
51X. Foot	0 1 2 3 U N		58X. Foot	0 1 2 3 U N
52X. Trunk	0 1 2 3 U N		59X. Trunk	0 1 2 3 U N
53X. Left sensory score (computed item)			60X. Right sensory score (computed item)	
			61X. Total sensory score (computed item)	

62X. Visual fields N: 0 Normal; 1 Abnormal; U Untestable

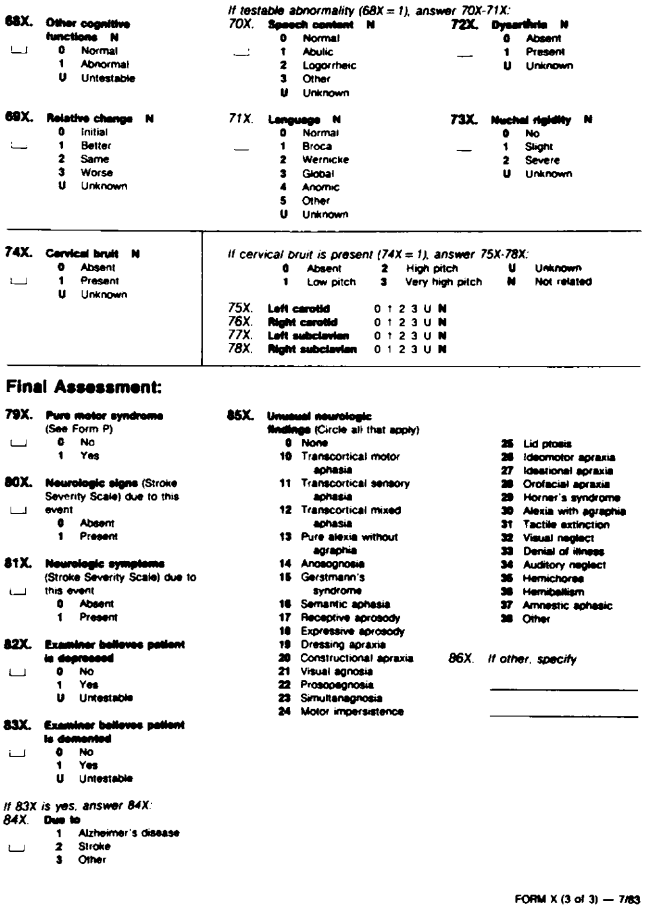
63X. Relative change N: 0 Initial; 1 Better; 2 Same; 3 Worse; U Unknown

If testable abnormality (62X = 1), answer 64X-67X:

64X. Monocular	N	0 Absent	1 Left	2 Right	3 Both	U Unknown
65X. Quadrantanopia	N	0 Absent	1 Left	2 Right	3 Both	U Unknown
66X. Hemianopia	N	0 Absent	1 Left	2 Right	3 Both	U Unknown
67X. Hemineglect	N	0 Absent	1 Left	2 Right	3 Both	U Unknown

FORM X (2 of 3) — 7/83

Fig 3.—Stroke Data Bank Neurologic History and Neurologic Examination form, page 2.



If testable abnormality (68X = 1), answer 70X-71X:

68X. Other cognitive functions N: 0 Normal; 1 Abnormal; U Untestable

69X. Relative change N: 0 Initial; 1 Better; 2 Same; 3 Worse; U Unknown

70X. Speech content N: 0 Normal; 1 Abulic; 2 Logorrheic; 3 Other; U Unknown

71X. Language N: 0 Normal; 1 Broca; 2 Wernicke; 3 Global; 4 Anomic; 5 Other; U Unknown

72X. Dysarthria N: 0 Absent; 1 Present; U Unknown

73X. Nuchal rigidity N: 0 No; 1 Slight; 2 Severe; U Unknown

74X. Cervical bruit N: 0 Absent; 1 Present; U Unknown

If cervical bruit is present (74X = 1), answer 75X-78X:

0 Absent; 1 Low pitch; 2 High pitch; 3 Very high pitch; U Unknown; N Not related

75X. Left carotid	0 1 2 3 U N
76X. Right carotid	0 1 2 3 U N
77X. Left subclavian	0 1 2 3 U N
78X. Right subclavian	0 1 2 3 U N

Final Assessment:

79X. Pure motor syndrome (See Form P): 0 No; 1 Yes

80X. Neurologic signs (Stroke Severity Scale) due to this event: 0 Absent; 1 Present

81X. Neurologic symptoms (Stroke Severity Scale) due to this event: 0 Absent; 1 Present

82X. Examiner believes patient is depressed: 0 No; 1 Yes; U Untestable

83X. Examiner believes patient is demented: 0 No; 1 Yes; U Untestable

If 83X is yes, answer 84X:

84X. Due to: 1 Alzheimer's disease; 2 Stroke; 3 Other

85X. Unusual neurologic findings (Circle all that apply):

0 None	10 Transcortical motor aphasia	11 Transcortical sensory aphasia
12 Transcortical mixed aphasia	13 Pure alexia without agraphia	14 Anosognosia
15 Gerstmann's syndrome	16 Semantic aphasia	17 Receptive aprosody
18 Expressive aprosody	19 Dressing apraxia	20 Constructional apraxia
21 Visual agnosia	22 Prosopagnosia	23 Simultanagnosia
24 Motor impersistence	25 Lid ptosis	26 Ideomotor apraxia
27 Ideational apraxia	28 Orofacial apraxia	29 Horner's syndrome
30 Alexia with agraphia	31 Tactile extinction	32 Visual neglect
33 Denial of illness	34 Auditory neglect	35 Hemichorea
36 Hemiballism	37 Amnestic aphasic	38 Other

86X. If other, specify: _____

FORM X (3 of 3) — 7/83

Fig 4.—Stroke Data Bank Neurologic History and Neurologic Examination form, page 3.

ues) because the κ coefficient is a more conservative measure of agreement for two reasons. First, it has an additional stringent requirement that the observations between neurologists should not only be consistently related but actually be identical. Thus, if one neurologist consistently evaluates a patient as slightly better (on some arbitrary quantitative scale) than another neurologist the correlation between the two could be very high but the κ coefficient of agreement would be relatively low. Second, all disagreements are equally weighted, unlike the case of a correlation coefficient where disagreements are weighted by the magnitude of the discrepancy.

Finally, low interobserver reliability does not necessarily indicate poor agreement among observers but rather poor agreement between observations. Since for many items the neurologists must depend on either subjective information (such as a verbal report for many of the neurologic history items), or objective responses that may vary over time, it is very likely that some of the interobserver variability is due to actual changes in the patient's course of performance. This was, in fact, underscored in the present study when one of the patients confided in one of the examining neurologists who appeared particularly sympathetic, that she wanted to tell him "the real story," something she "never told anyone else." Thus, low interobserver reliability may be just as indicative of an item (or patient) that is not stable over a time as it is indicative of the variability among the observers. In most studies (including the present study) the ultimate source of the variability is impossible to assess.

Implications for Clinical Research

Some of the interobserver variability is due not so much to difficulties in discriminating between levels within a given item but in distinctions between items. Once the items involved are identified, both the data collection forms and the data analysis can be improved. An illustration of this was obtained in the assessment of previous strokes and previous TIAs. In Table 3 it can be observed that patient No. 3 was considered to have never had a TIA by two neurologists and to have had a TIA by two other neurologists. An examination of the comparable table for previous stroke (item 8N) indicated a reciprocal result for the same patient, and subsequent discussion with the examiners showed that the same event was described by two neurologists as a TIA and by two other neurologists as a stroke. Given the fact that the patient had some communication problems and the fact that the patient now has substantial disability, and allowing for some variability on how the event was described (note that it was one to six months ago) it is very easy to understand how such disagreement may occur. By considering strokes and TIAs together as "a previous ischemic episode," and regarding the item as a two-category item with either a "no/never" response or a "yes" response the κ was recalculated and was found to be 0.60 ($P < .001$). Such a finding can provide insight not only to the reliability of the items and potential reliability of new items, but also to directions in reformulating questions and data

items both for collection and for analysis. Thus, in this particular case, it may be appropriate to replace the present items 4N and 8N with the more robust measure of "previous ischemic episode." Even if this is not done on the data forms, for some analytic purposes it may be worth pooling these two variables together with the knowledge that they yield a more consistently coded item.

Neurologic Assessment: State of the Art

The results of the present study and the two previous studies by Sisk et al[9] and Tomasello et al[10] can provide an appreciation of the state of the art of interobserver consistency in neurologic assessment. Note that all three studies involved various amounts of efforts specifically directed at improving the interobserver agreements beyond that which would be obtained by randomly pairing clinical neurologists without prior preparation.

Unfortunately, a direct comparison of the three studies is not possible because neither of the previous two studies used chance-corrected measures of agreement. Sisk et al[9] only presented raw frequencies of agreements and disagreements and Tomasello et al[10] used what they termed the "index of crude agreement," the ratio of subjects with agreement in responses to the total number of patients, limited to patients with one or more positive or abnormal ratings. Since their patients were not in-hospital stroke patients but rather patients identified as having had reversible cerebral ischemia,[10] it is likely that in their study the frequency of patients who would be identified as "normal" with respect to neurologic signs and symptoms would be greater than in the present study. The fact that all their neurologic examination items consisted of dichotomies of either "present" or "absent" would act to yield greater raw agreement among observers. Despite these factors they only obtained 25% agreement on sensory signs while the present study yielded a 50% agreement on sensory deficits (item 45X). The percent agreement on weakness in both studies was essentially identical (79% here *v* 76% in their study). Comparison of the two studies and their extent of agreement on neurologic signs shows that, in general, the present study yielded higher agreement levels. However, because the items were not identically defined, and the measure of agreement based on percent agreement is not a particularly useful statistic for research purposes, these comparisons are of limited usefulness.

In summary, because the present assessments are at least as reliable as those made in the previous studies, it is recommended that the statistics presented in Tables 5 and 9 be used as a current benchmark for interobserver consistency in neurologic assessments and, by implication, provide some initial robust criteria for evaluating relationships among these variables.

The results of this study have significant implications both to clinical stroke research in general and the SDB in particular. They demonstrate that research based on human observations and verbal communications is prone to interobserver variations which can account for significant differences in results among studies investigating the same phenomena. The unique contribution of the present study is that it provided sensitive quantitative measures of these variations, thus providing an insight into the extent of the problem in general, and for each of the variables studied in particular.

In the SDB, effort is now under way to identify the sources of interobserver disagreements with the goal of either changing data items, changing categories within items, or changing the definitions of individual items so that they can be more operational and less ambiguous. Once these suggestions are implemented in the data bank a reevaluation of interobserver agreement will be made.

Finally, in light of the present results it is recommended that other types of data be evaluated in a similar manner. In the SDB, studies of interobserver reliability in stroke diagnosis and computed tomography interpretation are also under way.

Financial support was provided by National Institutes of Health contracts N01-NS-2-2397, 2398, 2399, and 2302.

The authors express their gratitude to Fenwick Nichols, MD, Thomas Tatemichi, MD, and Sharon Fast, RN, for their assistance in preparing background material on the patients; to Karlin Richardson for the use of her computer program; to John Melski, MD, and Robert Stern, MD, who helped develop the Stroke Data Bank forms; and to Deborah Trout for secretarial assistance. William Weiss, Chief, Office of Biometry and Field Studies, NINCDS provided support and encouragement at every stage of the study, and is gratefully acknowledged.

References

1. Kunitz SC, Gross CR, Heyman A, et al: The Pilot Stroke Data Bank: Definition, design, and data. *Stroke* 1984;15:740-746.

2. Theodossi A, Knill-Jones RP, Skene A, et al: Inter-observer variation of symptoms and signs in jaundice. *Liver* 1981;1:21-32.

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. 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.

5. Kolmannskog F, Larsen S, Swensen T, et al: Reproducibility in observer variation at computed tomography and ultrasound of the normal pancreas. *Acta Radiol Diagn* 1983;24:21-25.

6. Kundel HL, Nodine CF: Interpreting chest radiographs without visual search. *Radiology* 1975;116:527-532.

7. Theodossi A, Skene AM, Portmann B, et al: Observer variation in assessment of liver biopsies including analysis by kappa statistics. *Gastroenterology* 1980;79:232-241.

8. Reuben A, Johnson AL, Cotton PB: Is pancreatogram reliable? A study of observer variation and error. *Br J Radiol* 1978;51:956-962.

9. Sisk C, Ziegler DK, Zileli T: Discrepancies in recorded results from duplicate neurological history and examination in patients studied for prognosis in cerebral vascular disease. *Stroke* 1970;1:14-18.

10. Tomasello F, Mariani F, Fieschi C, et al: Assessment of inter-observer differences in the Italian Multicenter study on reversible cerebral ischemia. *Stroke* 1982;13:32-35.

11. Garland LH: The problem of observer error. *Bull NY Acad Med* 1960;36:570-584.

12. Fleiss JL: Measuring nominal scale agreement among many raters. *Psychol Bull* 1971; 76:378-382.

13. 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.

14. Landis JR, Koch GG: The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159-174.