01 Jan 1986

Interobserver Agreement in the Diagnosis of Stroke Type

Cynthia R. Gross

David Shinar

Jay P. Mohr

Daniel B. Hier
Missouri University of Science and Technology, hierd@mst.edu

et. al. For a complete list of authors, see https://scholarsmine.mst.edu/chem_facwork/3331

Recommended Citation

C. R. Gross and D. Shinar and J. P. Mohr and D. B. Hier and L. R. Caplan and T. R. Price and P. A. Wolf and C. S. Kase and I. G. Fishman and S. Calingo and S. C. Kunitz, "Interobserver Agreement in the Diagnosis of Stroke Type," *Archives of Neurology*, vol. 43, no. 9, pp. 893 - 898, JAMA Neurology, Jan 1986.
The definitive version is available at https://doi.org/10.1001/archneur.1986.00520090031012

Interobserver Agreement in the Diagnosis of Stroke Type

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

• **Interobserver agreement is essential to the reliability of clinical data from cooperative studies and provides the foundation for applying research results to clinical practice. In the Stroke Data Bank, a large cooperative study of stroke, we sought to establish the reliability of a key aspect of stroke diagnosis: the mechanism of stroke. Seventeen patients were evaluated by six neurologists. Interobserver agreement was measured when diagnosis was based on patient history and neurologic examination only, as well as when it was based on results of a completed workup, including a computed tomographic scan. Initial clinical impressions, based solely on history and one neurologic examination, were fairly reliable in establishing the mechanism of stroke (ie, distinguishing among infarcts, subarachnoid hemorrhages, and parenchymatous hemorrhages). Classification into one of nine stroke subtypes was substantially reliable when diagnoses were based on a completed workup. Compared with previous findings for the same physicians and patients, the diagnosis of stroke type was generally more reliable than individual signs and symptoms. These results suggest that multicentered studies can rely on the independent diagnostic choices of several physicians when common definitions are employed and data from a completed workup are available. Furthermore, reliability may be less for individual measurements such as signs or symptoms than for more-complex judgments such as diagnoses.**

(*Arch Neurol* **1986;43:893-898)**

This study examines interobserver agreement for diagnosis of stroke type in the Stroke Data Bank (SDB). The centers participating in the SDB are the New York Neurological Institute, and the Departments of Neurology at the 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] Diagnostic reliability is the reproducibility or consistency of diagnoses made independently by different physicians (interobserver agreement) or by an individual physician over time (intraobserver agreement). It is distinct from accuracy, which requires a standard for comparisons.[2] For stroke, as for many neurologic diseases, there is no single, definitive procedure or standard on which to base a diagnosis. Instead, diagnosis of stroke type is a decision based on the evaluation of some or all of the following data: patient observation, history and physical examination, laboratory findings such as electrocardiogram and echocardiogram, and roentgenographic images such as computed tomographic (CT) scans and angiograms. Combining all of this information into a diagnosis is a subjective clinical judgment that requires both interpretation of the component pieces of information and weighing their relative importance in terms of inferential value. Disagreements in diagnosis can arise from inconsistency or disagreement in the gathering of patient observations through history taking and examination, in the interpretation of diagnostic test results, or in synthesizing this information to arrive at a diagnosis.[3-5] Assessing and improving observer agreement is particularly important in cooperative studies when data are collected and clinical evaluations are made by different observers. It is then incumbent on the researchers to provide some estimates of the interobserver reliability of their data base so that the validity of conclusions about and relationships among the variables measured can be established.

When the reliability of some neurologic signs and symptoms has been investigated,[6-10] low levels of agreement have been found. However, despite the lack of agreement for some signs and symptoms, one cannot assume that agreement for stroke diagnosis is no higher than that of any component piece of information. This is because diagnosis is generally based on the evaluation of multiple pieces of information, many of which may be redundant. The expected redundancy of clinical information may negate errors in individual observations and

Accepted for publication Oct 22, 1985.

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

Reprint requests to University of Minnesota, College of Pharmacy and School of Nursing, 308 Harvard St SE, Minneapolis, MN 55455 (Dr Gross).

make the diagnosis more reliable than the component pieces. For this reason, reliability of diagnosis must be assessed directly. Two studies that have examined interobserver reliability for signs and symptoms of transient ischemic attacks[8,10] obtained low levels of agreements for many signs and symptoms, whereas two studies that evaluated the interobserver reliability of whether or not a transient ischemic attack occurred,[11,12] obtained a substantial level of agreement among neurologists.

The purpose of the present study was to establish a paradigm for assessing the reliability of diagnosis data for a multicentered study of the clinical course of stroke. Agreement among neurologists for the diagnosis of stroke type was studied both as a function of the particular diagnosis and as a function of the amount of data available to the neurologist (eg, neurologic history and examination, with and without additional data such as laboratory findings, CT scans, and angiograms).

SUBJECTS AND METHODS

Participants

The subjects were 17 hospitalized stroke patients at the New York Neurological Institute who agreed to participate in the study. They ranged in age from 36 to 89 years, with a median age of 59 years. The observers in this study were six senior neurologists from the four SDB centers who are directly responsible for data collection in their centers. All are experienced in clinical neurology, with a special interest in stroke. These neurologists jointly developed the forms, agreed on the levels of definition of each data item, and were familiar with these definitions as they are described in the SDB coding manual.

Procedure and Design

The data in this study were collected in two phases. In the first phase all of the patients were interviewed and examined four times.[9] Patients were evaluated once in the morning and once in the afternoon on two consecutive days, each time by a different neurologist. Neurologists were assigned patients according to a blocked design that paired each neurologist with every other neurologist for at least five patients. Evaluations consisted of a neurologic history and examination, an initial clinical impression of the suspected mechanism or cause involved, and a rating on a five-point scale of the confidence level in the diagnostic judgment. The only information that was given to the neurologists before their examination of each patient was the patient's name, age, hospital admission date, and associated medical illnesses. In the case of aphasic patients the setting of the initial illness and initial in-hospital course of illness were also provided.

In the second phase, conducted three months later, a summary of the completed workup compiled for each patient was distributed to all of the neurologists for their in-depth diagnostic assessment. The data provided to the neurologists included initial evaluation findings, medical and neurologic history findings, neurologic examination findings, electrocardiogram, laboratory data, and 35-mm slides of CT scans (available for all patients) and angiograms (available for four patients). The six neurologists evaluated each of the 17 patients based on these data and then filled out an SDB diagnosis form (Figure) according to common definitions.[13]

The data analyses were designed to answer four questions:

1. When judgments are based solely on history taking and one neurologic examination, to what extent do neurologists agree in their clinical impressions of stroke diagnosis?
2. How and to what extent is the initial clinical impression affected by additional data?
3. What is the level of agreement among neurologists when the results of a completed workup are available?
4. What is the impact of a personally obtained neurologic examination and patient history on diagnosis?

Statistical Methods

The κ statistic was used to measure the level of agreement among neurologists. It is based on a formula developed by Fleiss[14] that 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. The κ statistic approaches -1.00 for complete disagreement and $+1.00$ for perfect agreement. When the agreement is equivalent to that expected by chance, κ equals 0. The significance of κ is tested by dividing it by its SE.[15] This ratio is distributed as a standard normal variate. In general, whenever κ is greater than .80 the agreement can be considered excellent; κ greater than .40 but less than or equal to .80 indicates moderate to substantial agreement; κ greater than .20 but less than or equal to .40 indicates fair agreement; and κ less than or equal to .20 indicates slight or poor agreement.[16] The sensitivity and specificity of the initial clinical impressions relative to the final diagnoses were also estimated.[17] In this context, sensitivity is the probability that an individual with a particular final diagnosis was assigned that diagnosis as an initial impression. Similarly, specificity is the probability that an individual not given a particular final diagnosis was not given that diagnosis as an initial impression.

RESULTS

Observer Agreement in Initial Impression of Stroke Mechanism

In the first phase of this study, each of the 17 patients was examined by four of the six neurologists participating in the study in a blocked design. Based entirely on their findings from the patient history and examination, without the benefit of any ancillary data, the neurologists were asked to provide their initial clinical impressions in terms of their "best guess" of the "mechanism of stroke," using the response options listed in item 5J (Figure) and the SDB classification rules.[13] The assignments given each patient by four neurologists are shown in Table 1. When all the classifications of infarcts were collapsed into a single category, there was complete agreement among four neurologists on seven of the 17 patients (patients 4, 5, 7, 9, 11, 12, and 17), and the chance-corrected level of agreement was fair ($\kappa = .38$; $P < .001$). However, when all of the individual subtypes of infarct classifications were analyzed—as the neurologists coded them on the forms—there were no complete agreements at all, and the chance-corrected level of agreement was only slight ($\kappa = .15$; $P < .001$). The high significance level that accompanied this relatively low level of agreement indicated that while the agreement level was low, it substantially exceeded chance. These results indicated that clinical impressions that distinguish among infarcts, subarachnoid hemorrhages, and parenchymatous hemorrhages were fairly reliable, but distinctions among subtypes of infarcts based on a history and examination alone were quite unreliable.

The Effect of Additional Information on Diagnosis

To evaluate the benefit of a completed workup, the final diagnosis based on all of the available data was compared with the initial clinical impression made on the basis of a patient history and examination only (Table 2). The emphasis here is on intraobserver consistency in a before-vs-after situation, with each neurologist serving as his own standard for final diagnosis. Each of the 68 entries in Table 2 represents a pair of judgments made by a single neurologist about one patient. Along the main diagonal are the clinical impressions that matched the final diagnoses (eg, "infarct cause unknown" was both the initial impression and the final diagnosis seven times). By each cell frequency is the column percentage. These are estimates of the probability a particular final diagnosis will be made, given that the patient has been assigned a specific initial clinical impression.

Overall, the initial clinical impres-

Stroke Data Bank
Diagnosis of Stroke

ID # _____
(PI) FORM J

5J. Primary diagnosis — Etiology

☐
1 Infarction, cause unknown
2 Infarction with normal angiogram
3 Infarction with tandem arterial pathology
4 Embolism from cardiac source
5 Infarction due to atherosclerosis
6 Lacuna
7 Parenchymatous hemorrhage
8 Subarachnoid hemorrhage
9 Other (Stroke)

If other (5J = 9), answer 6J.

6J. Specify _____

Excerpt from the Stroke Data Bank diagnosis form (Form J, item 5) describing categories for diagnosis of stroke.

sion and the final diagnosis were the same about one half of the time (49%), and the chance-corrected level of agreement was fair ($\kappa = .38$; $P < .001$). Collapsing all of the subtypes of infarcts into a single category raised the observed proportion of before-after agreements to three fourths (76%) and raised the chance-corrected level of agreement to a substantial level ($\kappa = .60$; $P < .001$).

To evaluate the utility of the initial impression as a screening test of the final decision about the mechanism of stroke, the sensitivity and specificity to infarction and hemorrhage were estimated. For infarction, the sensitivity of the clinical impression for a final diagnosis of infarction was .92 and its specificity was .76. Similarly, for hemorrhage the sensitivity was .60 and the specificity was .90.

The column percentages (in parentheses) along the diagonal in Table 2 indicate that the initial clinical impressions of lacuna and subarachnoid hemorrhage were durable. However, the initial impression of parenchymatous hemorrhage changed in the final diagnosis about as often as it remained the same. Finally, the initial impression of the specific type of infarction—from an unknown cause, due to atherosclerosis, or due to embolism from a cardiac source—changed in the final diagnosis more often than it remained the same.

An interesting effect of the additional information provided for the final diagnosis is that it did not necessarily reduce ambiguity about the mechanism of stroke. This can be observed from the change in assignments of signs and symptoms to different kinds of infarctions. When asked to specify the appropriate mechanism of stroke without the benefit of the final workup, the participating neurologists were often willing to interpret the clinical signs and symptoms as indicative of an embolism or thrombosis. However, failure to confirm this with additional data often reduced this level of specificity to the less-specific category "infarction with a cause unknown" (see Table 2 entries in the first row). Thus, in the present study the additional information served to rule out specific mechanisms as much as it served to identify them.

It was also of interest to test the impact of the physicians' confidence in their diagnosis. Presumably, initial clinical impressions made with a high level of confidence would correspond better with the final diagnosis than initial impressions made with a low level of confidence. Accordingly, the initial impressions of the stroke mechanism associated with low confidence levels (3 or less on a scale of 1 to 5) were analyzed separately from those with high confidence levels (more than 3). The κ coefficients for the before-after agreements for the two sets of data were .34 ($P < .001$) and .39 ($P < .001$), respectively ($z = 0.47$, not significant). The lack of a difference between the two κ values indicates that the subjective confidence a physician has in an initial impression following the neurologic examination is not a good discriminating clue to the durability or persistence of that diagnosis.

Interobserver Agreement With Final Workup Data

For those neurologists who examined a particular patient (four neurologists per patient drawn from the pool of six participating neurologists), the additional information provided in the final workup markedly increased the interobserver level of agreement. Agreement increased from a low level ($\kappa = .15$) for the initial specific clinical impression to a substantial level ($\kappa = .61$; $P < .001$) for the final diagnosis of the stroke mechanism. This high level of agreement was due to seven patients on whom all four neurologists agreed, and another eight patients on whom three of the four neurologists agreed. When all of the infarcts were collapsed into a single category, agreement improved ($\kappa = .69$; $P < .001$). When analyzed for all six neurologists, ie, the two who did not see the patient together with the four who did, a similar level of agreement for the mechanism of stroke was obtained ($\kappa = .64$; $P < .001$).

In general, as can be seen from Table 3, all of the neurologists agreed on the stroke mechanism for patients 1, 4, 6, 9, 10, and 11; five of six agreed on patients 2, 7, 13, 15, and 16; and four of six agreed on patients 3, 5, 12, 14, and 17. Thus, for all but one of the patients (patient 8), at least four of the six neurologists agreed on the mechanism of stroke. Furthermore, many of the "other" responses are essentially compatible variations of different categories. In fact, patient 8 may be a near-perfect match with all responses for either subarachnoid hemorrhage or its consequences (see footnote for Table 3).

The bottom row of Table 3 contains the partial κ values for each of the categories of stroke mechanism based on the findings of all six neurologists. These numbers can be interpreted as estimates of the probability that one neurologist will choose a particular mechanism given that another neurologist has already selected that mechanism, corrected for chance agreement. Thus, if one neurologist diagnoses parenchymatous hemorrhage, there is a very high probability ($P = .93$) that another neurologist will do the same (Table 3). On the other hand, if a neurologist diagnoses a subarachnoid hemorrhage, the probability that another neurologist will choose the same category is .52. In no instance did the majority of the neurologists concur that a patient had an

Table 1.—Distribution of Responses for Stroke Mechanism: Initial Clinical Impressions of Results From Phase I*

Patient No.	All Infarcts Combined	Initial Clinical Impressions: Infarctions: Cause Unknown	Infarctions: Tandem Pathologic Findings	Infarctions: Embolism, Cardiac Source	Infarctions: Atherosclerosis	Infarctions: Lacuna	Hemorrhages: Parenchymatous	Hemorrhages: Subarachnoid	Other
1	3	2	...	1	...	...	1	...	...
2	3	2	...	1	...	...	1	...	...
3	2	...	2	...	...	...	...	...	2
4	4	1	...	...	...	3	...	...	...
5	4	1	...	...	3	...	...	...	...
6	3	1	...	...	1	1	1	...	...
7	4	1	...	1	1	1	...	...	...
8	2	1	...	...	1	...	...	2	...
9	4	1	...	2	1	...	...	...	...
10	2	2	...	...	...	...	2	...	...
11	4	2	...	...	...	2	...	...	...
12	4	1	...	...	1	2	...	...	...
13	...	...	...	...	...	...	3	...	1
14	...	...	...	...	...	...	...	3	1
15	...	...	...	...	...	...	1	2	1
16	...	...	...	...	...	...	...	1	3
17	4	2	...	1	1	...	...	...	...
Total	**43**	**17**	**2**	**6**	**9**	**9**	**9**	**8**	**8**

*Cell entries indicate the number of neurologists who assigned a particular category to a given patient.

Table 2.—Comparison of Clinical Impressions With Final Diagnosis of Stroke Mechanism

Final Diagnosis of Stroke Mechanism	Initial Clinical Impressions of Stroke Mechanism: Infarctions 1	2	3	4	5	Hemorrhages 6	7	8 Other Strokes	Total
Infarctions									
1. Infarction, cause unknown	**7 (41)**	...	4 (66)	7 (78)	1 (11)	2 (22)	...	...	**21**
2. Infarction with tandem arterial pathologic findings	...	**2 (100)**	...	...	...	...	...	1 (13)	**3**
3. Embolism from cardiac source	1 (6)	...	**1 (17)**	...	...	...	...	...	**2**
4. Infarction due to atherosclerosis	1 (6)	...	...	**0 (0)**	...	...	...	...	**1**
5. Lacuna	4 (24)	...	1 (17)	...	**7 (78)**	...	...	...	**12**
Hemorrhages									
6. Parenchymatous hemorrhage	3 (18)	...	...	1 (11)	1 (11)	**5 (56)**	...	1 (13)	**11**
7. Subarachnoid hemorrhage	...	...	...	...	...	...	**7 (88)**	2 (25)	**9**
8. Other strokes	1 (6)	...	...	1 (11)	...	2 (22)	1 (12)	**4 (50)**	**9**
Total	**17**	**2**	**6**	**9**	**9**	**9**	**8**	**8**	**68**

*Cell entries represent the pair of diagnostic assignments for mechanism of stroke made by each examining neurologist about a patient after phase I (initial clinical impression) and after phase II (final diagnosis). (Numbers in parentheses are the column percentages.)

infarction due to embolism or atherosclerosis. However, the low reliability levels for these diagnoses may reflect their low frequency in our study sample.[18]

Finally, the relationship between the interobserver level of agreement and the level of detail of the information sought was also examined (Table 4). Looking at the data for each combination of examining and reviewing neurologists, it can be readily observed that the level of agreement is a function of the level of detail required. The greater the detail in diagnosis, the lower the interobserver agreement.

The Impact of the Personally Performed History and Examination

As comprehensive as any data collection form may be, it may still not be able to capture the whole image of the patient as the clinical neurologist does. It was important to analyze how interobserver agreement was affected by whether or not a particular neurologist saw that patient or only had access to the written workup. This analysis was possible since patients were seen by different subsets of four of the six neurologists. Each neurologist saw approximately the same number of patients, each pair of neurologists was teamed for at least five patients, and the two neurologists who did not see a given patient were not the same for all patients but rotated among the six. Thus, interobserver agreement on the final diagnosis was examined separately for three groups: (1) the four neurologists who

Table 3.—Distribution of Responses for Stroke Mechanism: Final Diagnosis Results From Phase II

Patient No.	All Infarcts Combined	No. of Final Diagnoses: Infarctions: Cause Unknown	Infarctions: With Normal Angiogram	Infarctions: Tandem Pathologic Findings	Infarctions: Embolism, Cardiac Source	Infarctions: Atherosclerosis	Infarctions: Lacuna	Hemorrhages: Parenchymatous	Hemorrhages: Subarachnoid	Other Stroke*
1	6	6	...	...	...	...	...	...	...	...
2	6	5	1	...	...	...	...	...	...	...
3	5	1	...	4	...	...	...	...	...	1
4	6	...	...	...	...	...	6	...	...	...
5	6	4	...	...	...	2	...	...	...	...
6	...	...	...	...	...	...	...	6	...	...
7	6	1	...	...	...	...	5	...	...	...
8	...	...	...	...	...	...	...	...	3	3
9	6	6	...	...	...	...	...	...	...	...
10	...	...	...	...	...	...	...	6	...	...
11	6	...	...	...	...	...	6	...	...	...
12	6	4	...	...	...	...	2	...	...	...
13	...	...	...	...	...	...	...	5	...	1
14	...	...	...	...	...	...	...	...	4	2
15	...	...	...	...	...	...	...	...	5	1
16	...	...	...	...	...	...	...	...	1	5
17	6	4	...	...	2	...	...	...	...	...
Total	**59**	**31**	**1**	**4**	**2**	**2**	**19**	**17**	**13**	**13**
Partial κ†	.96	.64	−.01	.58	.18	.18	.83	.93	.52	.35

*Specified as "other": patient 3, operative embolism; patient 8, infarction due to vasospasm, infarction secondary to vasospasm, spasm with aneurysm; patient 13, arteriovenous malformation; patient 14, operative infarction, infarction secondary to vasospasm; patient 15, infarction due to vasospasm; and patient 16, planned iatrogenic embolus, embolism (not cardiac source), iatrogenic embolism, iatrogenic embolism, and infarction from embolism due to procedure.

†Partial κ is a measure of chance-corrected agreement between raters on a particular response category. All values above .12 are significant at $P < .05$.

Table 4.—κ Coefficients for Final Diagnosis

Final Diagnosis	κ Coefficient*: 4 Examining Neurologists	2 Reviewing Neurologists†	All Neurologists
Primary diagnosis—cause	.61	.54	.64
Primary diagnosis—cause, recoded to 4 categories‡	.69	.80	.76

*All coefficients are significant at $P < .001$.

†Based on 16 patients only.

‡Categorization was as follows: primary diagnosis included infarction (cause unknown), infarction with normal angiogram, infarction with tandem arterial pathologic findings, embolism from cardiac source, infarction due to atherosclerosis, and lacuna, all recoded as infarctions; parenchymatous hemorrhage, recoded as parenchymatous hemorrhage; subarachnoid hemorrhage, recoded as subarachnoid hemorrhage; and other, recoded as other.

examined that patient in the first phase of the study; (2) the two neurologists who did not examine the patient in the first phase of the study; and (3) all six neurologists (Table 4).

The agreements among the four neurologists who personally examined the patient and between the two neurologists who only reviewed the written workup were $\kappa = .61$ and $\kappa = .54$, respectively ($z = .57$, not significant). Agreement among all six was slightly higher ($\kappa = .64$). Apparently, the personal contact between the patient and physician (which took place three months before their review of the written workup and CT scans) did not influence the choice of diagnosis in a manner that could be detected by a difference in the level of agreement.

COMMENT

This study indicates that experienced neurologists, collaborating in a common research effort and using common definitions and data-collection forms, can achieve high levels of agreement with respect to stroke mechanism. Furthermore, final diagnoses made after review of a completed workup are more reliable (although often less specific) than initial impressions.

With regard to the reliability of clinical data and judgments, Koran[4] has suggested that increasing either the number of observers or diagnostic categories lowers the level of interobserver agreement. We found that a nine-category classification was substantially reliable and that the level of agreement among six neurologists was as high as that for four or two. Furthermore, although collapsing diagnoses into a small number of categories led to greater reliability, even a nine-category classification was substantially reliable.

A diagnosis is generally more reliable than many of the component observations that contribute to it. Using the same patients and physicians as in the current investigation, Shinar et al[9] found κ coefficients for individual stroke signs and symptoms that were generally lower than the κ values for stroke type observed in the present study. For example, interobserver agreements are low ($\kappa \leq .25$) for some neurologic history items, such as course of illness within the first hour of onset, presence of focal deficits at onset, and reported use of antiplatelet agents or anticoagulants. Of over 40 neurologic signs, only five were more reliable than the diagnosis of stroke type: extraocular movements, swallowing, articulation, lateralized weakness, and facial weakness.

The observation that elementary findings, many of which are unreliable, are synthesized to form a reliable diagnosis is consistent with previous studies of medical decision making.[19] To deal effectively with all of the

individual data items of a case, the physician reduces the complexity of the problem by aggregating sets of items into a coherent concept.[20] It has been shown that people do not generally use more information to make more accurate decisions but rather use it to select a few key items or filter out irrelevant items.[21] This selection process is guided by what has been labeled as the confirmation bias: the tendency to attend to information that confirms a chosen hypothesis and ignore information that refutes it.[22]

According to the model proposed by Eddy and Clanton,[20] the physician chooses either a small aggregation of findings (eg, a syndrome) or a salient, perhaps striking, individual finding and uses it to generate a list of possible diagnoses. Additional findings are then used to select a final diagnosis from this list. With this model, disagreements in diagnosis might arise even if identical findings are observed and recorded by each clinician. Two clinicians may choose different findings as their keys for assembling lists of possible diagnoses ("differential diagnosis") or weigh additional findings differently in the process of selecting the final diagnosis. If this model is correct, then a fruitful approach to improving the reliability of diagnosis would be to identify the findings physicians use in making certain diagnoses and develop from these findings diagnostic criteria that can be standardized among observers. The classification rules for the mechanism of stroke detailed by Mohr et al[13] are illustrative of this approach.

The present SDB classification rules emphasize distinguishing among subtypes of infarctions. The findings of the present study indicate that an additional classification that would permit combining subarachnoid hemorrhages with vasospasm due to aneurysm and infarction due to vasospasm into a single category would improve the reliability of the scheme.

The increase in agreement that occurs with additional data (ie, a CT scan or angiogram) is substantial. It suggests that these tests are very informative. Furthermore, it suggests that the levels of agreement reported in this study are probably lower bound estimates for the agreement on estimates of the SDB diagnoses and those of other studies with similar methodologies. This is because, in practice, additional data are collected from family members, nurses, charts, and notes; examinations are repeated; additional data or opinions, such as those of the neuroradiologist, are obtained; and the final diagnosis is typically arrived at by a consensus of two or more neurologists. These methods have been identified by Feinstein[19] and the McMaster University Study.[23] as strategies for preventing or minimizing clinical disagreements.

It should be noted that subjects in this study were preselected on the basis of having had a stroke. The inclusion of subjects whose stroke status was questionable could have reduced the levels of agreement on the diagnosis of stroke type. The issue of agreement on primary diagnosis (ie, Is the disease present or not?) is of greater significance in certain other neurologic diseases, such as multiple sclerosis[24] and transient ischemic attacks,[11,12] where similar investigations with preselected populations reported disagreements among expert observers as to whether or not the disease was present.

Investigators in multicentered clinical studies frequently use diagnostic subtypes for determining study eligibility and for subgroup analyses. The present study indicates that diagnostic subtypes, at least for stroke, can be reliably ascertained, even with a detailed classification scheme. This reliability depends on common definitions and data collection forms as well as a large body of data from a completed clinical workup. Studies lacking common definitions or varying the amount of information available to each clinician risk a potentially serious reduction in the level of interobserver agreement for diagnosis.

Financial support was provided by National Institutes of Health, Bethesda, Md, 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 the patient summaries, to Karlin Richardson for use of her computer program, and to Anita Roth for her assistance in the preparation of this manuscript. William Weiss, Chief, Office of Biometry and Field Studies, retired, provided support to initiate this project 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. Koran LM: The reliability of clinical methods, data and judgments: I. *N Engl J Med* 1975; 293:642-646.

3. Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario: Clinical disagreement: I. How often it occurs and why. *Can Med Assoc J* 1980;123:499-504.

4. Koran LM: The reliability of clinical methods, data and judgments: II. *N Engl J Med* 1975; 293:695-701.

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

6. Aoki N, Horibe H, Ohno Y, et al: Epidemiological evaluation of funduscopic findings in cerebrovascular diseases: III. Observer variability and reproducibility for funduscopic findings. *Jpn Circ J* 1977;41:11-17.

7. McCance C, Watt JA, Hall DJ: An evaluation of the reliability and validity of the plantar response in a psychogeriatric population. *J Chronic Dis* 1968;21:369-374.

8. 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.

9. Shinar D, Gross CR, Mohr JP, et al: Interobserver variability in the assessment of neurologic history and examination in the Stroke Data Bank. *Arch Neurol* 1985;42:557-565.

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

11. Kraaijeveld CL, van Gijn J, Schouten HJA, et al: Interobserver agreement for the diagnosis of transient ischemic attacks. *Stroke* 1984;15:723-725.

12. Calanchini PR, Swanson PD, Gotshall RA, et al: Cooperative study of hospital frequency and character of transient ischemic attacks: IV. The reliability of diagnosis. *JAMA* 1977;238:2029-2033.

13. Mohr JP, Nichols FT, Tatemichi TK: Classification and diagnosis of stroke. *Int Angiol* 1984;3:431-439.

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

15. Fleiss JL, Nee JCM, Landis JR: Large sample variance of κ in the case of different sets of raters. *Psychol Bull* 1979;86:974-977.

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

17. Armitage P: *Statistical Methods in Medical Research.* Boston, Blackwell Scientific Publications Inc, 1971.

18. Walter SD: Measuring the reliability of clinical data: The case for using three observers. *Rev Epidemiol Sante Publique* 1984;32:206-211.

19. Feinstein AR: *Clinical Judgment.* Baltimore, Williams & Wilkins, 1967.

20. Eddy DM, Clanton CH: The art of diagnosis. *N Engl J Med* 1982;306:1263-1268.

21. Payne JW: Information processing theory: Some concepts and methods applied to decision research, in Wallsten TS (ed): *Cognitive Processes in Choice and Decision Behavior.* Hillsdale, NJ, Lawrence Erlbaum Assoc Inc, 1980.

22. Schustack MW, Sternberg RJ: Evaluation of evidence in causal inference. *J Exp Psychol Gen* 1981;110:101-120.

23. Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario: Clinical disagreement: II. How to avoid it and how to learn from one's mistakes. *Can Med Assoc J* 1980;123:613-617.

24. Westlund KB, Kurland LT: Studies on multiple sclerosis in Winnipeg, Manitoba and New Orleans, Louisiana. *Am J Hyg* 1953;57:380-396.