01 Jan 1987

Interobserver Reliability in the Interpretation of Computed Tomographic Scans of Stroke Patients

David Shinar

Cynthia R. Gross

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

Louis R. Caplan

et. al. For a complete list of authors, see https://scholarsmine.mst.edu/chem_facwork/3325

Recommended Citation

D. Shinar and C. R. Gross and D. B. Hier and L. R. Caplan and J. P. Mohr and T. R. Price and P. A. Wolf and C. S. Kase and I. G. Fishman and J. A. Barwick and S. C. Kunitz, "Interobserver Reliability in the Interpretation of Computed Tomographic Scans of Stroke Patients," *Archives of Neurology*, vol. 44, no. 2, pp. 149 - 155, JAMA Neurology, Jan 1987.
The definitive version is available at https://doi.org/10.1001/archneur.1987.00520140021012

Interobserver Reliability in the Interpretation of Computed Tomographic Scans of Stroke Patients

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

• Interobserver reliability in interpretation of computed tomographic images was studied by six senior neurologists who independently evaluated on a standardized Stroke Data Bank form the brain lesions of 17 patients. The results analyzed with κ statistics yielded moderate to substantial agreement on most items of interest including the stroke pathology and anatomy. In general, the levels of agreement were as high as previously reported for the diagnosis of the mechanism of the stroke, and much higher than on many stroke history items and items of neurologic examination. Excellent agreement was obtained for the detection of infarcts and intracerebral hemorrhage, and substantial agreement was obtained on whether the computed tomographic images were normal or indicative of small deep infarcts, superficial and deep infarcts, and aneurysms. The level of agreement on anatomy of the lesions was best for the frontal, parietal, and temporal lobes, putamen, cerebellum, and subarachnoid space. Implications for clinical research and diagnosis are discussed.

(*Arch Neurol* 1987;44:149-155)

This study focused on interobserver agreement in the interpretation of computed tomographic (CT) images and involved six senior neurologists who examined slides of CT scans obtained from 17 hospitalized stroke patients. It is the third in a series of studies focusing on quality assurance in the Stroke Data Bank (SDB). The centers participating in the study 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]

Observer errors and interobserver disagreements in medical data have received much attention in the last decade, and a review of their origins has been offered by Sackett.[2] In his taxonomy, a distinction is made among the following three major sources of disagreement among observers: variations among the examiners, variations in the nature of the examination, and variations (over time) in the examined, ie, the patient. Obviously, only the first source is relevant in studying interobserver reliability in the interpretation of radiographic images such as CT scans, since the other two are held constant.

Even in the absence of the last two sources of interobserver disagreements, it is now readily acknowledged that the interpretation of pictographic data such as roentgenogram and CT scans is subjective[3,4] and heavily influenced by individual differences among observers due to factors such as expectancy and past experience.[5]

Even among senior physicians, some degree of interobserver variation may be expected. Indeed, significant variations among experienced radiologists were obtained in a study of CT scans of a sample of patients in whom brain tumor had been suspected.[6] In stroke, where the CT scan is a major diagnostic tool, these differences may be of clinical significance, eg, in deciding on modes of therapy.

In the present study, the interobserver level of agreement in the interpretation of CT scans was relevant to the quality of the data in the SDB and of general interest concerning the stability of CT readings across physicians. The interpretation of CT scans

Accepted for publication Aug 25, 1986.

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

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

of the same 35-mm projection slides were made in each center by a senior neurologist specializing in stroke. A brief patient abstract without diagnosis was provided. Thus, it can be assumed that the level of expertise applied to the task of interpreting the CT scan was as good as any that could be expected in the course of a routine clinical evaluation. The reliability of CT interpretations obtained in this study is probably a good approximation of the upper bounds of reliability of CT interpretation for stroke lesions in general.

PATIENTS AND METHODS

Participants

The 17 subjects, ranging in age from 36 to 89 years, were all of the in-hospital patients with acute stroke who were hospitalized during a given week at the New York Neurological Institute, who agreed to participate in the study, and who were considered to be in stable condition. Specific characteristics of each patient have been documented elsewhere.[7]

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

Procedures and Design

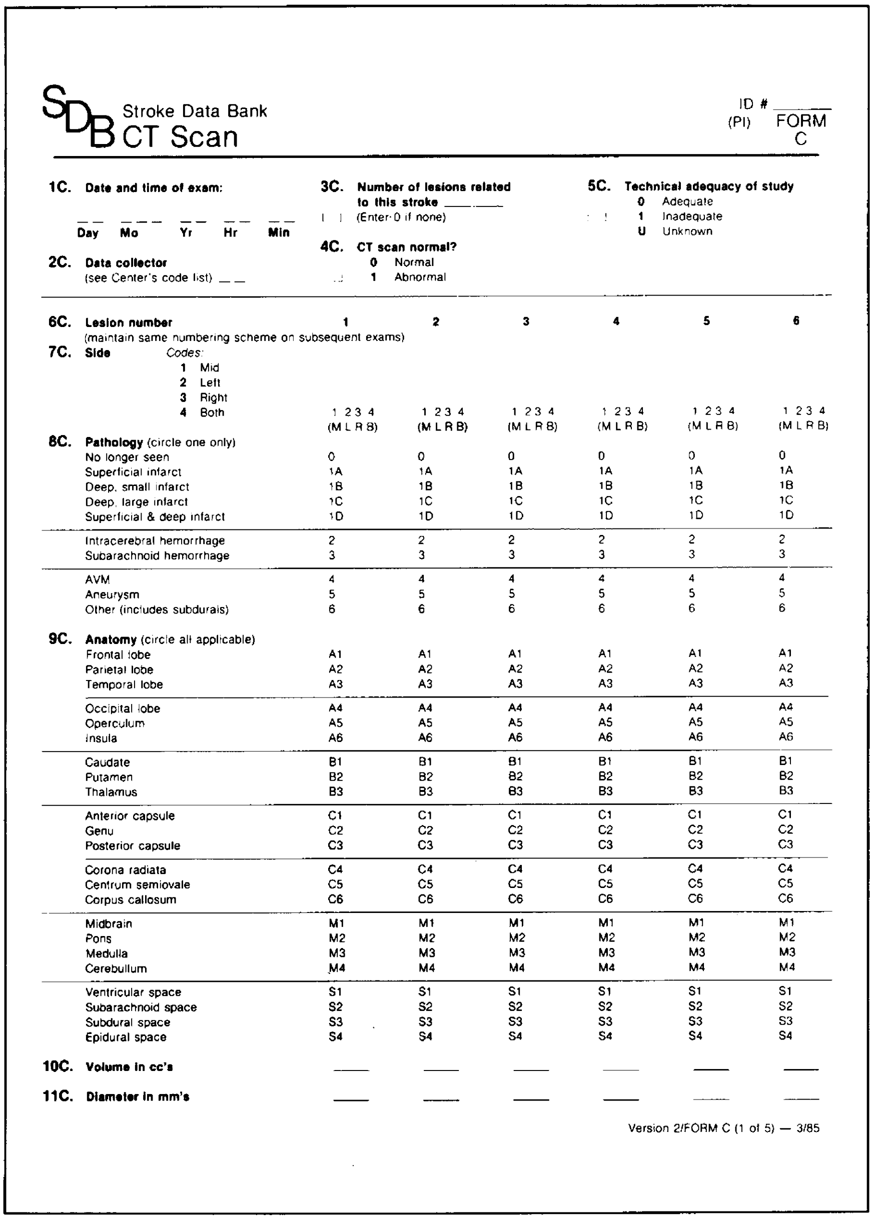
During a visit to the New York Neurological Institute, each of the six neurologists personally and independently interviewed and examined ten to 12 of the 17 patients according to a blocked design as part of a study on interobserver agreement in neurologic assessment.[7]

Three months later, back at their own centers, each of the six neurologists received and independently reviewed 35-mm projection slide copies of the CT scans of all 17 patients. The procedure for interpretation of the CT scans was modeled to approximate the usual mode of interpretation in the SDB. The same forms were used to record data, and the timing of the CT scans was delayed by up to five days from stroke onset so as to provide data that were as clear-cut as possible. The only difference was in the use of 35-mm slides as compared with the actual complete CT folder.

In addition to the CT scans, the neurologists also received for each patient a brief abstract containing the patient's age, hospital admission date, associated medical illness, neurologic history, and findings of the neurologic examination. The package did not contain any references to the patient's diagnosis or the interpretation of the CT scans by the physicians caring for the patient.

Form and Materials

Neurologic history and the results of the neurologic examination on each patient were recorded on the standard SDB Neurologic History and Neurologic Examination forms.[7] The interpretation of the CT images was recorded by all the neurologists on the SDB CT form (Figs 1 and 2). The form contains 21 items describing the nature of the lesion(s) and can accommodate the description of up to six lesions.



SDB Stroke Data Bank
CT Scan

ID # ______
(PI) FORM C

1C. **Date and time of exam:**
__ ___ __ __ __
Day Mo Yr Hr Min

2C. **Data collector**
(see Center's code list) __ __

3C. **Number of lesions related to this stroke** ______
| | (Enter 0 if none)

4C. **CT scan normal?**
0 Normal
1 Abnormal

5C. **Technical adequacy of study**
0 Adequate
1 Inadequate
U Unknown

6C. **Lesion number** (maintain same numbering scheme on subsequent exams)	1	2	3	4	5	6
7C. **Side** *Codes:* 1 Mid 2 Left 3 Right 4 Both	1 2 3 4 (M L R B)	1 2 3 4 (M L R B)	1 2 3 4 (M L R B)	1 2 3 4 (M L R B)	1 2 3 4 (M L R B)	1 2 3 4 (M L R B)
8C. **Pathology** (circle one only)						
No longer seen	0	0	0	0	0	0
Superficial infarct	1A	1A	1A	1A	1A	1A
Deep, small infarct	1B	1B	1B	1B	1B	1B
Deep, large infarct	1C	1C	1C	1C	1C	1C
Superficial & deep infarct	1D	1D	1D	1D	1D	1D
Intracerebral hemorrhage	2	2	2	2	2	2
Subarachnoid hemorrhage	3	3	3	3	3	3
AVM	4	4	4	4	4	4
Aneurysm	5	5	5	5	5	5
Other (includes subdurals)	6	6	6	6	6	6
9C. **Anatomy** (circle all applicable)						
Frontal lobe	A1	A1	A1	A1	A1	A1
Parietal lobe	A2	A2	A2	A2	A2	A2
Temporal lobe	A3	A3	A3	A3	A3	A3
Occipital lobe	A4	A4	A4	A4	A4	A4
Operculum	A5	A5	A5	A5	A5	A5
Insula	A6	A6	A6	A6	A6	A6
Caudate	B1	B1	B1	B1	B1	B1
Putamen	B2	B2	B2	B2	B2	B2
Thalamus	B3	B3	B3	B3	B3	B3
Anterior capsule	C1	C1	C1	C1	C1	C1
Genu	C2	C2	C2	C2	C2	C2
Posterior capsule	C3	C3	C3	C3	C3	C3
Corona radiata	C4	C4	C4	C4	C4	C4
Centrum semiovale	C5	C5	C5	C5	C5	C5
Corpus callosum	C6	C6	C6	C6	C6	C6
Midbrain	M1	M1	M1	M1	M1	M1
Pons	M2	M2	M2	M2	M2	M2
Medulla	M3	M3	M3	M3	M3	M3
Cerebullum	M4	M4	M4	M4	M4	M4
Ventricular space	S1	S1	S1	S1	S1	S1
Subarachnoid space	S2	S2	S2	S2	S2	S2
Subdural space	S3	S3	S3	S3	S3	S3
Epidural space	S4	S4	S4	S4	S4	S4
10C. **Volume in cc's**	____	____	____	____	____	____
11C. **Diameter in mm's**	____	____	____	____	____	____

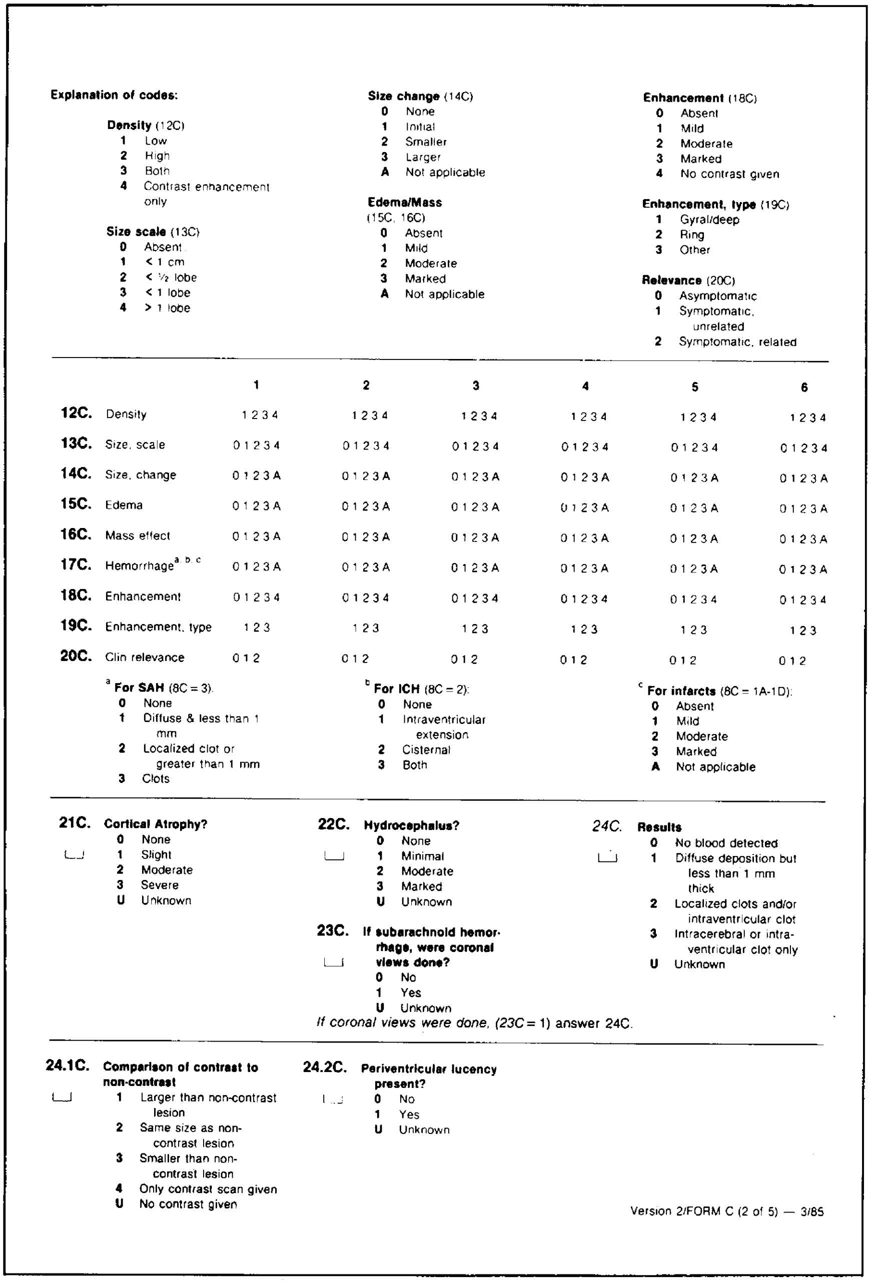
Version 2/FORM C (1 of 5) — 3/85

Fig 1.—Stroke Data Bank computed tomographic (CT) scan form, page 1.

For the CT evaluation, the neurologists were provided with high-resolution 35-mm slides of the original CT scan images. All slides were considered to be of adequate technical quality by the participating neurologists. Figure 3 contains the CT images from patient 2.

Statistical Method

The κ statistic used to measure the level of agreement was developed by Fleiss[8] to provide a numerical measure of agreement among multiple raters on variables that are scored on a nominal scale (qualitative classification, without ordering). 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 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 SE.[9] This ratio is distributed as a standard normal variate. It has been suggested that 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.[10]



Explanation of codes:

Density (12C)
1 Low
2 High
3 Both
4 Contrast enhancement only

Size scale (13C)
0 Absent
1 < 1 cm
2 < ½ lobe
3 < 1 lobe
4 > 1 lobe

Size change (14C)
0 None
1 Initial
2 Smaller
3 Larger
A Not applicable

Edema/Mass (15C, 16C)
0 Absent
1 Mild
2 Moderate
3 Marked
A Not applicable

Enhancement (18C)
0 Absent
1 Mild
2 Moderate
3 Marked
4 No contrast given

Enhancement, type (19C)
1 Gyral/deep
2 Ring
3 Other

Relevance (20C)
0 Asymptomatic
1 Symptomatic, unrelated
2 Symptomatic, related

		1	2	3	4	5	6
12C.	Density	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4
13C.	Size, scale	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4
14C.	Size, change	0 1 2 3 A	0 1 2 3 A	0 1 2 3 A	0 1 2 3 A	0 1 2 3 A	0 1 2 3 A
15C.	Edema	0 1 2 3 A	0 1 2 3 A	0 1 2 3 A	0 1 2 3 A	0 1 2 3 A	0 1 2 3 A
16C.	Mass effect	0 1 2 3 A	0 1 2 3 A	0 1 2 3 A	0 1 2 3 A	0 1 2 3 A	0 1 2 3 A
17C.	Hemorrhage[a, b, c]	0 1 2 3 A	0 1 2 3 A	0 1 2 3 A	0 1 2 3 A	0 1 2 3 A	0 1 2 3 A
18C.	Enhancement	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4
19C.	Enhancement, type	1 2 3	1 2 3	1 2 3	1 2 3	1 2 3	1 2 3
20C.	Clin relevance	0 1 2	0 1 2	0 1 2	0 1 2	0 1 2	0 1 2

[a] **For SAH** (8C = 3).
0 None
1 Diffuse & less than 1 mm
2 Localized clot or greater than 1 mm
3 Clots

[b] **For ICH** (8C = 2):
0 None
1 Intraventricular extension
2 Cisternal
3 Both

[c] **For infarcts** (8C = 1A-1D):
0 Absent
1 Mild
2 Moderate
3 Marked
A Not applicable

21C. **Cortical Atrophy?**
□
0 None
1 Slight
2 Moderate
3 Severe
U Unknown

22C. **Hydrocephalus?**
□
0 None
1 Minimal
2 Moderate
3 Marked
U Unknown

23C. **If subarachnoid hemorrhage, were coronal views done?**
□
0 No
1 Yes
U Unknown

If coronal views were done, (23C = 1) answer 24C.

24C. **Results**
□
0 No blood detected
1 Diffuse deposition but less than 1 mm thick
2 Localized clots and/or intraventricular clot
3 Intracerebral or intraventricular clot only
U Unknown

24.1C. **Comparison of contrast to non-contrast**
□
1 Larger than non-contrast lesion
2 Same size as non-contrast lesion
3 Smaller than non-contrast lesion
4 Only contrast scan given
U No contrast given

24.2C. **Periventricular lucency present?**
□
0 No
1 Yes
U Unknown

Version 2/FORM C (2 of 5) — 3/85

Fig 2.—Stroke Data Bank computed tomographic scan form, page 2.

RESULTS

Of the different items in the SDB CT form, the most important are probably those describing the pathology and anatomy of the observed lesions (items 8C and 9C). Initial review of these items indicated that there was a need to collapse across lesions rather than analyze each lesion separately. This was done because in several cases a lesion labeled lesion 1 by one neurologist was labeled lesion 2 by another and, thus, although the neurologists may have been in perfect agreement on both lesions, a separate analysis of each lesion (by lesion number) would have yielded an erroneous impression of a low level of agreement.

Pathology

Table 1 summarizes the six neurologists' responses to the observed pathology (item 8C). In this table, each patient is represented by a row and each of the alternative sources of pathology is represented by a column. For 13 of the 17 patients, only one type of pathology was noted. For the additional four patients, more than one type of lesion was noted by some of the neurologists, and these are noted in Table 1. At the gross level of distinction among "Normal CT," "Infarct," "Hemorrhage," and "Other," the agreement among the neurologists was excellent, $\kappa = .90$. (Wherever a neurologist coded more than one category, the one consistent with the most frequent choice of the other neurologists was selected. This approach will yield the closest agreement. A more conservative approach, of selecting the pathology of lesion 1 only, still yielded very good agreement, $\kappa = .78$.) This is reflected by the cell entries indicating that at this level of discrimination all six neurologists agreed on 13 patients and five out of the six neurologists agreed on three of the remaining four patients. Even at high levels of specificity (involving all ten categories of pathology) agreement was substantial, $\kappa = .61$. (The conservative approach yielded $\kappa = .56$.)

The level of agreement among neurologists for each of the specific categories is indicated by the partial κ values (last line in Table 1). The partial κ values denote the chance-corrected probability that if one randomly selected neurologist chooses a given category, another randomly selected neurologist will choose the same category. The partial κ values vary widely from excellent ($\kappa_6 = 1.00$) for intracerebral hemorrhage to poor ($\kappa_4 < .2$) for large deep infarcts. The low frequency with which this category was cited may be partially responsible for its low agreement level.[11] Subarachnoid hemorrhage and arteriovenous malformation were never cited; so we have no data to determine their reliability.

Interobserver agreement was substantial ($\kappa_p > .6$) for identifying small deep infarcts and aneurysms as well as for indicating no abnormalities at all.

For combined superficial and deep infarcts, the agreement was moderate ($\kappa_5 = .47$), and for superficial infarcts it was fair ($\kappa_2 = .29$). In these categories, the major source of disagreement was on whether a lesion was "superficial" only or "superficial and deep" (Table 1, patients 2, 3, 8, 14, 16, and 17). Note, further, that the κ values for superficial infarcts and for deep infarcts are extremely conservative since they are based on agreements that the infarcts were *only* superficial *or only* deep but not both. For example, for patient 2, although superficial infarct was noted by only two neurologists (Table 1), all six neurologists agreed that there was a superficial infarct since four of the six neurolo-

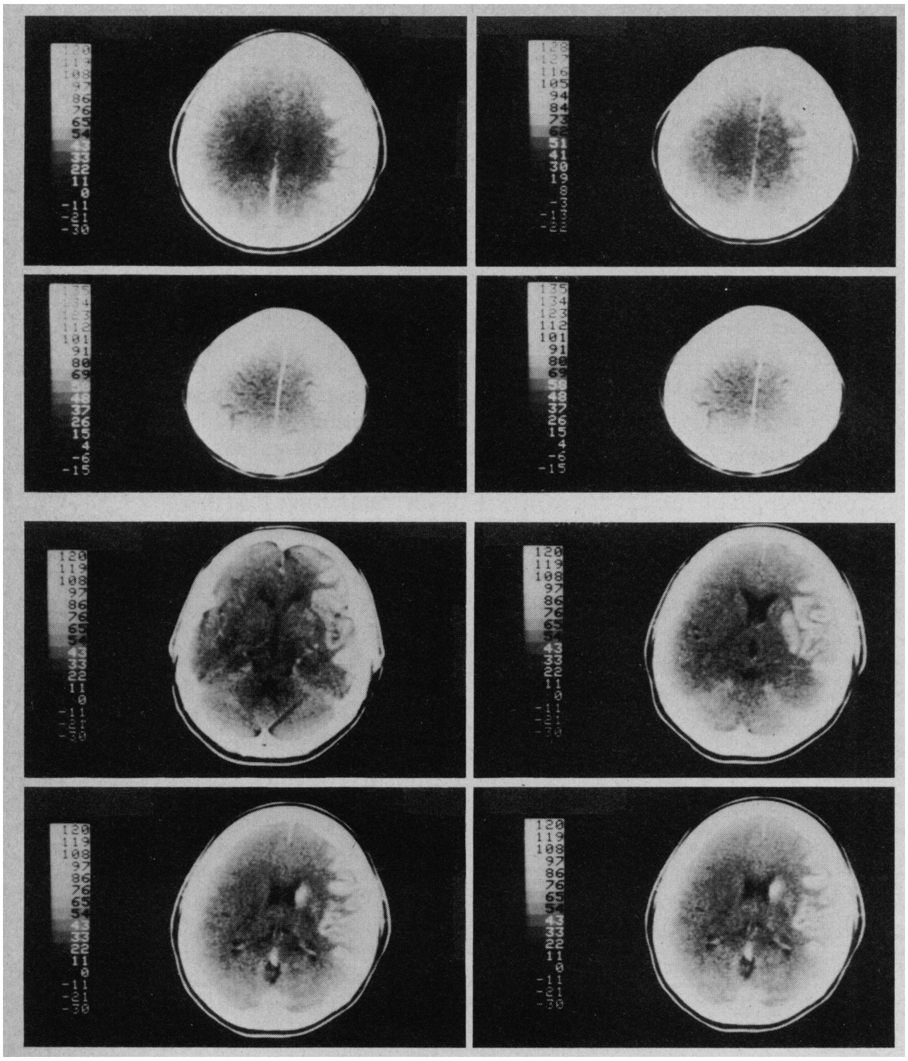


Fig 3.—High-resolution 35-mm slide of original computed tomographic image for patient 2.

gists recorded "superficial and deep."

Anatomy

Agreement on the specific anatomic region(s) of the lesion(s) (item 9C) could not be summarized by a single numeric value since, for most patients, lesions were noted in more than one location, and the indicated number of locations varied among neurologists. The distribution of the neurologists' responses for each of the 23 anatomic locations specified in item 9C is presented in Table 2. As can be seen from the "Total" row and column entries, the six neurologists listed 212 locations for the 17 patients, or an average of two locations per patient by each neurologist. Qualitatively, it can be observed that the crude level of agreement (not chance-corrected) was quite good, as follows: for nine of 17 patients all six neurologists agreed, and for five additional patients, five of six neurologists were in agreement on the location of at least one lesion or the absence of any lesions (patient 5).

The data obtained for patient 2 (Fig 3) illustrates the problem of providing a single statistic for agreement on this item. All six neurologists noted lesions in the frontal and parietal lobes. Four of six neurologists also noted lesions in the temporal lobe while, of the other two neurologists, one checked the operculum while the other checked the operculum as well as the insula. Four of the six neurologists also noted lesions in the caudate. Of these four neurologists, two neurologists also noted a lesion in the putamen, one neurologist noted a lesion in the anterior capsule, and one neurologist noted lesions in the putamen and the anterior capsule.

Agreement on each of the specific regions was calculated with the κ coefficient by considering the probability of selecting each region vs not selecting it. These κ coefficients are presented at the bottom row of Table 2. First, for four of the five major brain regions—hemisphere, deep, cerebellum, and extracerebral spaces—the interobserver agreement was substantial, $.65 \leq \kappa \leq .84$. There was also substantial interobserver agreement in deciding whether the CT was normal or abnormal, $\kappa = .68$. Interobserver agreements on identifying lesions in the brain stem were essentially undefined with the present sample of cases, since none of the neurologists identified a lesion in either the midbrain or the medulla, and only one neurologist selected the pons. A larger or a different sample might have provided sufficient data to analyze the level of agreements on lesions in the brain stem.

With respect to the specific locations within the hemispheres, the deep structures, the extracerebral spaces, and the miscellaneous categories, the levels of agreement varied widely. Substantial agreement was obtained for the frontal, temporal, and parietal lobes and the subarachnoid space. Agreement was moderate for deep lesions in the caudate, putamen, and the anterior and posterior capsule; it was only fair for lesions noted in the genu and ventricular space. Interobserver agreement was poor and statistically nonsignificant in only the following four regions: the operculum, the insula, the centrum semiovale, and the thalamus; and poor, but significantly above chance, for the corona radiata. As with pathology, here too the poor agreement levels may be due in part to the low frequency with which some of these sites were noted.[11] Because there were either no citings or only one, it was impossible to evaluate interobserver agreement for the occipital lobes, corpus callosum, midbrain, pons, medulla, and the subdural and epidural spaces.

Other CT Findings

Agreements among neurologists were also evaluated on the additional items listed in Table 3. The level of agreement was statistically significant for all items. Agreement was substantial ($\kappa \geq .60$) on whether the CT was normal (item 4C), and the general location of the lesion (item 7C—side: left, right, mid, or any combination of the three).

The agreements on lesion density (item 12C—high, low, or both), the number of lesions related to this

Table 1.—The Distribution of the Neurologists' Response With Respect to the Pathology Seen on CT Scan for Each Patient*

	Infarcts					Hemorrhage					
Patient	CT Normal	Superficial	Deep Small	Deep Large	Superficial and Deep	ICH	SAH	AVM	Aneurysm	Other	Total
1	4	. . .	2	. . .	. . .	. . .	. . .	. . .	. . .	. . .	6
2	. . .	2	. . .	. . .	4	. . .	. . .	. . .	. . .	. . .	6
3	. . .	4	. . .	. . .	2	. . .	. . .	. . .	. . .	. . .	6
4	. . .	. . .	6	. . .	. . .	. . .	. . .	. . .	. . .	. . .	6
5	5	. . .	. . .	1	. . .	. . .	. . .	. . .	. . .	. . .	6
6	. . .	. . .	. . .	. . .	. . .	6	. . .	. . .	. . .	. . .	6
7	. . .	. . .	5	1	. . .	. . .	. . .	. . .	. . .	. . .	6
8	. . .	1	. . .	. . .	5†	. . .	. . .	. . .	. . .	. . .	6
9	. . .	. . .	. . .	. . .	6	. . .	. . .	. . .	. . .	. . .	6
10	. . .	. . .	. . .	. . .	. . .	6	. . .	. . .	. . .	. . .	6
11	. . .	. . .	6	. . .	. . .	. . .	. . .	. . .	. . .	. . .	6
12	. . .	. . .	4	1	1	. . .	. . .	. . .	. . .	. . .	6
13	. . .	. . .	. . .	. . .	. . .	6‡	. . .	. . .	. . .	. . .	6
14	. . .	3	. . .	. . .	3	. . .	. . .	. . .	. . .	. . .	6
15	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	5§	1	6
16	. . .	3‖	. . .	. . .	3	. . .	. . .	. . .	. . .	. . .	6
17	. . .	3	. . .	. . .	3	. . .	. . .	. . .	. . .	. . .	6
Total	**9**	**16**	**23**	**3**	**27**	**18**	**0**	**0**	**5**	**1**	**102**
Partial κ values	.68¶	.29¶	.76¶	.03	.47¶	1.00¶	. . .	. . .	.79¶	−.01	. . .

*CT indicates computed tomography; ICH, intracerebral hemorrhage; SAH, subarachnoid hemorrhage; and AVM, arteriovenous malformation.

†One of the neurologists who entered superficial and deep lesion also entered ICH, another neurologist also entered aneurysm, and a third neurologist entered "other" category.

‡One of the neurologists who entered ICH also entered AVM.

§One of the neurologists who entered aneurysm also entered superficial infarct, and two other neurologists also entered SAH.

‖One of the neurologists who entered superficial infarct also entered AVM.

¶Significant at $P < .001$.

stroke (item 3C—0, 1, 2, or 3) and the mass effect of the largest region (item 16C—absent, mild, moderate, or marked) were moderate ($\kappa > .40$).

The Relationship of CT Interpretation to Diagnosis

At the same time that the neurologists completed the CT forms for this study, they also diagnosed the primary mechanism of stroke.[12] To examine the relationships between the two, the neurologists' responses of the "Primary Diagnosis" of stroke (item 5J on the Diagnosis form) were cross-tabulated with their responses to the pathology (item 8C on the CT form). Since the diagnosis categories partially overlap the pathology categories of item 8C in the CT form,[13] the correlation between the two provides a measure of intraobserver consistency. The cross-tabulation of these two variables is provided in Table 4. The contingency coefficient, an χ^2-related measure of association, was .75, indicating a substantial level of correspondence between these variables. Further, some specific relationships are of interest. First, the most commonly diagnosed condition, infarction, was also recognized as such on the CT. Also, the diagnosis of intracerebral hemorrhage (ICH) was always associated with an ICH code on the CT. In contrast, the diagnosis of subarachnoid hemorrhage (SAH) was not once associated with a CT code of SAH. Instead, other CT codes such as infarction (eight of 13) and aneurysm (four of 13) were selected as the relevant CT manifestation of patients with a primary diagnosis of SAH (these are not unexpected associations).

To assess the impact of the personal contact and examination on the interpretation of the CT image, the neurologists were divided into two groups as a function of whether or not they saw the patients. This was possible since each of the neurologists personally examined ten to 12 of 17 patients in a blocked design, so that each patient was examined by a different subset of four of six neurologists.[7] An analysis of interobserver agreement and image interpretation, as a function of whether or not the neurologist actually saw and personally examined the patient, did not reveal any statistically significant effects. Thus, the direct contact did not influence the CT interpretation beyond the effects of the written documentation available to all physicians on each patient.

COMMENT

The purpose of this study was to evaluate interobserver reliability. Since in most stroke cases in the SDB, as well as in other studies, the CT scan serves as the "gold standard," the neurologists' performance relative to an external criterion or truth could not be assessed. However, in general, the reliability of any measure determines the upper bound of its validity, in particular, for a collaborative study such as the SDB it is a prerequisite to valid data and must be assessed.

This study was conducted in a relatively realistic environment, ie, the neurologists examined the CT images in their own offices, were under no time pressure to reach their conclusions, and had at their disposal the results of the neurologic examination, the neurologic and medical history and, for ten to 12 of 17 patients, also had the benefit of a personal examination of the patient. These features make this study different from most of the other studies on the validity and reliability of CT interpretations,[6,14] interpretations of other radiologic data,[15] and interpretations of other analog data[16] and more relevant to the actual level of interobserver reliability that exists in a clinical setting. In comparison with the cited studies, this study's features would also tend to increase the level of interobserver agreement.

The level of agreement among the

Table 2.—Distribution of Neurologists' Responses With Respect to Anatomy of Lesions Seen on CT Scans*

Patient	No. Lesions Seen	Hemisphere							Deep Structures									Brain Stem†	
		Fron	Par	Tem	Occ	Opr	Ins	Any	Cau	Put	Thal	AC	G	PC	CR	CS	Any	Pons	Any
1	4	...	...	...	...	...	...	...	...	...	2	...	...	1	...	...	2	...	...
2	...	6	6	4	...	2	1	6	4	3	...	2	...	...	...	...	4	...	...
3	...	3	6	3	1	1	1	6	1	1	...	...	...	...	1	1	3	...	...
4	...	...	...	...	...	...	...	...	...	5	...	...	...	2	3	1	6	...	...
5	5	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	1	1
6	...	...	...	...	...	...	...	...	...	6	...	...	...	3	2	...	6	...	...
7	...	...	...	...	...	...	...	...	2	3	...	1	1	1	3	1	6	...	...
8	...	2	2	...	...	...	...	4	...	...	...	...	...	...	...	...	...	...	...
9	...	6	6	5	...	1	1	6	4	4	...	5	3	3	3	3	5	...	...
10	...	2	3	6	...	1	1	6	...	...	...	...	...	...	...	...	...	...	...
11	...	...	...	...	...	...	...	...	...	1	1	1	...	5	...	...	6	...	...
12	...	1	1	...	...	1	...	2	...	...	1	...	...	5	...	3	6	...	...
13	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
14	...	6	1	...	...	...	...	6	...	...	...	...	...	...	...	...	...	...	...
15	...	1	...	...	...	...	...	1	...	...	...	...	...	...	...	...	...	...	...
16	...	6	...	...	...	...	1	6	...	...	...	...	...	...	1	...	1	...	...
17	...	6	3	...	...	2	1	6	...	...	...	...	...	...	1	...	1	...	...
Total	**9**	**39**	**28**	**18**	**1**	**8**	**6**	**49**	**11**	**23**	**4**	**9**	**4**	**20**	**14**	**9**	**46**	**1**	**1**
κ value	.68	.71	.65	.70	−.01	.02	−.06	.84	.41	.54	.06	.44	.27	.43	.17	.20	.68	−.01	−.01

*All κ values >.20 are significant at $P < .001$. $\kappa = .17$ is significant at $P = .01$. Fron indicates frontal lobe; par, parietal lobe; tem, temporal lobe; occ, occipital lobe; opr, operculum; ins, insula; cau, caudate; put, putamen; thal, thalamus; AC, anterior capsule; G, genu; PC, posterior capsule; CR, corona radiate; CS, centrum semiovale; ven, ventricular space; and sub, subarachnoid space.

†Following locations were not cited at all: corpus callosum, midbrain, medulla, subdural space, and epidural space.

‡Sum total does not include nine reports of "no lesions seen."

Table 3.—Level of Agreement Among Neurologists on CT-Related Variables*

Variable No.	Name	κ
3C	No. of lesions related to this stroke	.56
4C	CT scan normal	.60
7C	Side (of lesion)	.65
12C	Lesion density	.59
13C	Lesion size (of largest lesion)	.37
15C	Edema (largest value)	.34
16C	Mass effect (largest value)	.52

*CT indicates computed tomography.

Table 4.—The Relationship Between Primary Diagnosis and the CT-Based Pathology*

CT Pathology	Primary Diagnosis				
	Infarcts	ICH	SAH	Other	Total
Not seen	9	...	...	...	**9**
Infarct	50	...	8	11	**69**
ICH	...	17	...	1	**18**
Aneurysm	...	...	4	1	**5**
Other	...	...	1	...	**1**
Total	**59**	**17**	**13**	**13**	**102**

*Entries indicate number of combinations out of 17 patients times six neurologists. CT indicates computed tomography; ICH, intracerebral hemorrhage; and SAH, subarachnoid hemorrhage.

neurologists was generally quite high for most of the items studied despite the fact that agreement was measured with a chance-corrected statistic, ie, the κ coefficient. As could have been expected, the level of agreement rose as the category under evaluation became more general. Thus, the agreement on whether or not the CT was normal was higher than the agreement on the specific type of lesion or its location. Similarly, the level of agreement on whether the pathology was an infarct or a hemorrhage was higher than the level of agreement on the type of infarct; and the level of agreement on the general location of the lesion (eg, lobar vs the deep structures) was higher than the level of agreement on the specific location within these general categories.

Among the component entities making up the general categories, the variability in agreements was quite high. As expected, agreement on entities that have fuzzy boundaries (such as "large deep infarct" vs a "small deep infarct" and anatomic structures such as the insula and the operculum) is not as high as the agreement on entities that have sharper boundaries (such as intracerebral hemorrhage and anatomic structures such as the frontal and parietal lobes). One implication of such results is that emphasis should be placed on developing diagnostic categories that have relatively sharp demarcations that allow them to be distinguished easily from other categories.

The levels of agreement in the interpretation of the CT scan are much higher than the interobserver levels of agreement obtained for recording many neurologic signs and symptoms.[7] Since the data for both studies were obtained from the same patients and relied on the same neurologists for interpretation, the higher levels of agreement in the interpretation of the CT image can be directly attributed to the absence of situation-

	Extracerebral Spaces†			
Cerebellum	Ven	Sub	Any	Total No. of Lesions
. . .	. . .	. . .	. . .	3
. . .	. . .	. . .	. . .	28
. . .	. . .	. . .	. . .	19
. . .	. . .	. . .	. . .	11
. . .	. . .	. . .	. . .	1
. . .	. . .	. . .	. . .	11
. . .	. . .	. . .	. . .	12
. . .	1	1	2	6
. . .	. . .	. . .	. . .	44
. . .	. . .	. . .	. . .	13
. . .	. . .	. . .	. . .	8
. . .	. . .	. . .	. . .	12
5	3	1	3	9
. . .	. . .	. . .	. . .	7
. . .	. . .	6	6	7
. . .	. . .	. . .	. . .	8
. . .	. . .	. . .	. . .	13
5	**4**	**8**	**11**	**212‡**
.79	.27	.73	.65	

al variations in the nature of examinations and the variations within each patient (over time and between examinations), which are the other two major sources of observer error.[2] Thus, the higher reliability of the interpretation of the CT image provides strong support for its use in understanding stroke.

The interobserver range of agreements on the CT scan and the diagnosis[12] were quite similar. Thus, the partial κ coefficients for the diagnosis item of primary cause of stroke were .96 for infarcts and .93 for ICH, while the partial κ values for the corresponding CT item of pathology (item 8C) were .88 for infarcts and 1.00 for intracerebral hemorrhage. The reason for the high interobserver agreement on the diagnosis—even though it does include all three sources of observer error—is that diagnosis is probably based on redundant sources of information including CT scans. This also explains the high correlation between the CT item of pathology and the diagnosis item of primary mechanism of stroke.

The lack of a significant effect of previous examination of the patient is at a slight variance with the finding obtained by McNeil et al,[17] in which neurologists' ability to detect intracranial disease was improved by 3% when provided with a complete patient history. However, in that study only a simple yes or no determination was required, and the sample size was large (N = 84), and both factors acted to increase the level of significance without necessarily affecting the magnitude of the effect.

Financial support was provided by contracts NO1-NS-2-2397, 2398, 2399, and 2302 issued by the National Institutes of Health, Bethesda, Md.

The authors express their gratitude to Fenwick Nichols, MD, Thomas Tatemichi, MD, and Sharon Fast, RN, for their assistance in preparing the patient summaries, and to Karlin Richardson and Sandra Calingo for use of their computer programs. William Weiss, Chief, Office of Biometry and Field Studies, retired, provided support to initiate this project, and his support 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. 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.

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

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

5. Kundel HL, Follette PS: Visual search patterns and experience with radiological images. *Radiology* 1972;103:523-528.

6. Swets JA, Pickett RM, Whitehead SF, et al: Assessment of diagnostic technologies. *Science* 1979;205:753-759.

7. 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.

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

9. 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.

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

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

12. Gross CR, Shinar D, Mohr JP, et al: Interobserver agreement in the diagnosis of stroke type. *Arch Neurol* 1986;43:893-898.

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

14. Turner DA, Ramachandran PC, Ali AA, et al: Brain scanning with the Anger multiplane tomographic scanner as a primary examination. *Radiology* 1976;121:125-129.

15. Musch DC, Landis JR, Higgins ITT, et al: An application of κ-type analysis to interobserver variation in classifying chest radiographs for pneumoconiosis. *Stat Med* 1984;3:73-83.

16. Holman CDJ, James IR, Heenan PJ, et al: An improved method of analysis of observer variation between pathologists. *Histopathology* 1982;6:581-589.

17. McNeil BJ, Hanley JA, Funkenstein HH, et al: Paired receiver operating characteristic curves and the effect of history on radiographic interpretation. *Radiology* 1983;149:75-77.