Discrepancies in Recorded Results from Duplicate Neurological History and Examination in Patients Studied for Prognosis in Cerebrovascular Disease

BY CHARLES SISK, M.D.,* DEWEY K. ZIEGLER, M.D.,† AND TURGUT ZILELI, M.D.*

Abstract:
Discrepancies in Recorded Results from Duplicate Neurological History and Examination in Patients Studied for Prognosis in Cerebrovascular Disease

Histories were taken and neurological examinations performed according to a carefully standardized protocol by two neurologists on 28 patients with histories of probable transient cerebrovascular insufficiency. Duplicate examinations were performed within a short interval of the initial examinations. Comparison of the findings on the two examinations showed extreme discrepancies in recording of presence or absence of specific symptoms and of signs on the neurological examination, including signs considered “objective”—for example, reflex asymmetry. Discussion between the examiners followed by another set of duplicate examinations reduced the discrepancies on the findings of neurological examination, but did very little to the discrepancies in the history. The necessity of realizing the variability of the history and examination in evaluating the course of, and therapy for, transient cerebrovascular ischemic attacks is emphasized.

ADDITIONAL KEY WORDS transient ischemic attacks standard protocol statistical analysis natural history

Introduction

To obtain valid data on any clinical research, the extent of inter-observer variability of response to the items chosen for study must be known. In recent years, we have been engaged in a detailed study of the prognosis of patients with transient ischemic attacks of cerebrovascular insufficiency. It appeared to us that inter-observer discrepancy of observation, as recorded both in historical data and in the findings on neurological examination, was considerably larger than is generally appreciated, and warranted study.

Our intent was to list and define specific items in the history and neurological examination to be checked at each visit of the patient, in the hope that the fine points of the natural history of the disease and, subsequently, responses to therapy would be more accurately elucidated. The purpose of this report is to document the discovered inter-observer variability of reported data after considerable care had been taken to avoid ambiguity, and to comment on its relevance to the problem of prognosis in these patients.

Methods

In a prospective study of patients at the Kansas University Medical Center with a diagnosis of probable transient cerebral ischemic attacks or cerebral infarction with minimal residual, two staff neurologists (D.Z. and T.Z.) performed...
duplicate examinations on 30 patients. Both examiners had had several years' experience in clinical neurology.

Patients were selected for analysis if (1) both examiners independently concluded that they had had definite symptoms and/or signs of a previous cerebral vascular insufficiency and (2) if they were considered mentally competent to provide a satisfactory neurological history. The printed protocol listed 20 neurological symptoms to be covered in the history (present or absent), and 32 items in the neurological examination, the results of which were recorded by checking one of the listed findings. To avoid discrepancy in methods of questioning and interpretation of findings, the examiners formulated a manual in which the definition of terms in all questions and responses, both on history and neurological examination, were agreed upon in detail.

For each study case, both neurologists independently completed this standardized protocol on the same day. Although time intervals between each examination were short, care was taken not to tire or disturb any patient by prolonged questioning. Each entire examination rarely lasted more than one hour. The mean age of the patients was 59 years with a range from 51 to 69 years. The shortest interval between any patient's most recent cerebral insult and subsequent neurological examination was one week, and the longest interval was 12 weeks.

Of the first 16 cases examined, 14 had adequate data for complete analysis. A careful study of the findings on these initial 14 examinations was followed by a discussion between examiners in an attempt to ascertain the cause for discrepancy of response and to minimize any recognizable semantic and interpretive differences. Subsequently, an additional 14 patients were examined in a similar fashion to determine if these differences had been corrected.

**Results**

**NEUROLOGICAL SYMPTOMS**

Table 1 lists neurological symptoms collected on all 28 patients. The first and last columns of the table contain the numbers of patients in which the examiners were in agreement for a specific symptom, either as to its presence (col. 1) or its absence (col. 4). The second and third columns contain the numbers of patients in which the neurologists disagreed in their interpretation of the neurological history.

Gross inspection of the table does not suggest that either neurologist was more prone to interpret a history as positive or negative. There are no statistically significant differences between the examiners as determined by marginal chi square analysis, but it is apparent that there was disagreement in obtaining a positive interpretation of vertigo, memory disturbance, dysarthria, paresis, paresthesia, syncope, and ataxia.

Efforts to standardize interpretation of the neurological history in the second study group of patients did not improve the distribution of agreements and disagreements.

**NEUROLOGICAL SIGNS**

The distribution of neurological signs is tabulated in table 2. Inspection of these distributions again reveals patterns essentially similar to those for the neurological history. In addition, it is apparent that Neurologist 2 (T.Z.) is significantly more inclined to record abnormal findings than Neurologist 1 (D.Z.) for “weakness” and fifth motor nerve abnor-

---

**Table 1**

<table>
<thead>
<tr>
<th>Symptom</th>
<th>Interpretation of symptoms</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diplopia</td>
<td>+</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>23</td>
</tr>
<tr>
<td>Vertigo</td>
<td>+</td>
<td>9</td>
<td>0</td>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>Memory disturbance</td>
<td>+</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>19</td>
</tr>
<tr>
<td>Convulsions</td>
<td>+</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>26</td>
</tr>
<tr>
<td>Amnesia</td>
<td>-</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>22</td>
</tr>
<tr>
<td>Dysarthria</td>
<td>+</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>Coma</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>25</td>
</tr>
<tr>
<td>Paresis</td>
<td>-</td>
<td>8</td>
<td>4</td>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td>Paresthesia</td>
<td>-</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>16</td>
</tr>
<tr>
<td>Monocular blindness</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>26</td>
</tr>
<tr>
<td>Binocular blindness</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>26</td>
</tr>
<tr>
<td>Aphasia</td>
<td>+</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>23</td>
</tr>
<tr>
<td>Headaches</td>
<td>+</td>
<td>7</td>
<td>0</td>
<td>4</td>
<td>17</td>
</tr>
<tr>
<td>Syncope</td>
<td>+</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>21</td>
</tr>
<tr>
<td>Ataxia</td>
<td>+</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>16</td>
</tr>
</tbody>
</table>

* D.Z.
† T.Z.
— indicates the examiner obtained positive history of symptoms.
+ indicates he did not obtain such history.

---

* Marginal chi square (one degree of freedom) = (a — d — 1)^2 where a = positive, negative disagreement and d = negative, positive disagreements.
TABLE 2

Neurological Examination

<table>
<thead>
<tr>
<th>Interpretation of signs</th>
<th>Neurologist 1*</th>
<th>Neurologist 2†</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
</tbody>
</table>

Visual field restriction‡ | 2   | 1   | 2   | 51  |
Nystagmus                 | 0   | 0   | 5   | 23  |
Light reaction abnormality| 2   | 0   | 2   | 52  |
Extraocular movement       |     |     |     |     |
  abnormality‡             | 2   | 0   | 0   | 54  |
Vibratory loss‡            | 0   | 1   | 3   | 53  |
Paresis                   | 2   | 0   | 1   | 25  |
Fifth motor dysfunction‡  | 1   | 0   | 6   | 49  |
Facial weakness‡           | 0   | 0   | 2   | 54  |
Ninth and tenth motor     | 0   | 0   | 2   | 26  |
Hypoglossal               | 0   | 0   | 2   | 26  |
Romberg                   | 0   | 3   | 0   | 25  |
Conjugate gaze            | 1   | 1   | 3   | 23  |
Station and gait          | 8   | 0   | 5   | 15  |
Weakness (paresis)        | 3   | 0   | 10  | 43  |
Cerebellar†               | 2   | 6   | 0   | 48  |
Extensor plantar response†| 2   | 2   | 7   | 45  |

* D.Z.
† T.Z.
‡ Total observations = 56 for each examination performed on both sides of body in each of 28 patients.
+ indicates the examiner found this neurological sign (abnormality) to be present.
- indicates examiner did not find this abnormality.

Deep tendon reflexes were graded as normal and bilaterally symmetrical, asymmetrical, or absent (table 3). Findings on biceps, patellar and Achilles reflexes were selected for analysis from the 28 patients. Each digit in the table, therefore, refers to finding of one reflex pair (e.g., left and right biceps reflex) by the two examiners. Reflexes were recorded as present and symmetrical, asymmetrical, or absent. "Asymmetrical" includes all degrees of asymmetry, and does not identify if one side is hyperreflexia or hyporeflexia.

Table 3 shows the striking finding that Neurologist 1 found reflex asymmetry in 19 reflex pairs which Neurologist 2 classified as normal, whereas in the reverse situation Neurologist 2 classified only two reflex pairs as asymmetrical where Neurologist 1 found them normal. As was to be expected in a large number of reflex pairs, there was agreement as to asymmetry (21). In none of the 21 instances in which both examiners found asymmetrical reflexes was there disagreement as to the laterality of the reflex hyperactivity.

Neck bruits were graded on a 1 to 4 scale. There was good agreement for carotid bruits with 22 equally graded (17 graded as absent; three graded as 2; and two graded as 3 with no disagreement as to laterality). Of the six disagreements, one was graded as 2 and 3; one as 3 and 4; three as 0 and 2; and one as 1 and 4.

An attempt to minimize differences in interpretation of the neurological examination met with more success than that effected in reconciling differences in the history. For example, a definite semantic difference between neurologists in the interpretation of "paresis" (weakness) was recognized and appropriate adjustment made. (See Discussion.)

TABLE 3

Grading of Paired Patellar Achilles, and Biceps Deep Tendon reflexes by Independent Examinations*

<table>
<thead>
<tr>
<th>Reflex grading</th>
<th>Neurologist 1 Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Present and symmetrical</td>
</tr>
<tr>
<td>Present and symmetrical</td>
<td>40</td>
</tr>
<tr>
<td>Abnormal and asymmetrical</td>
<td>2</td>
</tr>
<tr>
<td>Absent</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>43</td>
</tr>
</tbody>
</table>

* Number of patients examined = 28. Total observations 84, with three of paired reflex observations (biceps, patellar, Achilles) being recorded by each examiner on each case.
DISCREPANCIES IN NEUROLOGICAL HISTORY AND EXAMINATION

This adjustment resulted in only three disagreements in this item in the second study group as opposed to seven in the first.

Discussion

The data suggest that for every case with a symptom interpreted as positive by both neurologists, there are one or more cases in which they disagree. Such a finding is of practical importance, since it implies—given different neurologists on follow-up examination—that there is at least an equal likelihood that a second observer will interpret a positive history differently for a majority of neurological symptoms. Even more disconcerting is the suggestion that the distribution of agreements and disagreements for the neurological history follows a binomial pattern.* In other words, if one were to toss a coin repeatedly and assign heads as indicating a positive neurological history and tails a negative history and tabulate the numbers of paired tosses which were two heads, a head and a tail, and two tails, the observed patterns in table 1 and the expected patterns for coin tossing would be quite similar.† This suggests that the probability of two neurologists' eliciting a positive history of the same symptom is a random phenomenon and is no more predictive of a "true" positive history than the arbitrary classification of cases as normal or abnormal depending on the toss of a coin.

It is of interest that whereas elicitation of history may be considered to have an irreducible "subjective" element, the findings on neurological examination should be highly "objective"—one meaning of which should be minimal reliance on observer interpretation. The somewhat surprising findings were that some discrepancies appeared in the recording of pupillary light reaction, function of the fifth, ninth, tenth and twelfth cranial nerves, and conjugate gaze, and several disagreements were found in "normality-abnormality" of station and gait, weakness of an extremity, nystagmus, and extensor plantar response. Similarly a striking disparity was found in recording deep tendon reflexes (table 3), with 19 observations in which one observer found asymmetry and another did not.

Marked inter-observer variability in stethoscopical findings has been studied (1). It is of interest that in our study agreement between the two examiners was greatest on auscultation of neck bruits despite the possibility for variation in interpretation of sounds from this area (2).

One can only speculate on possible reasons for the discrepancies in the elicited histories and neurological signs. Feinstein (3) has discussed at length this general subject of sources of error in the gathering of clinical data, and we will mention only a few. Concerning the history, it is possible that the interview atmosphere between patient and examiner may facilitate memory in one case more than another. One examiner may be more interested in certain symptoms, and thus ask questions in a slower and more searching fashion.

It is also possible that inappropriate grading of responses may have, in part, been responsible for the high proportion of disagreements between examiners in both history and examination. Rather than classifying responses as either positive or negative, an improved percentage of agreements might have been attained by recording interpretations as positive, negative or equivocal. Perhaps an even more semiquantitative grading of responses would have been more appropriate. McCance, Watt and Hall, (4) in a study to evaluate the reliability of the plantar response, attempted to classify this neurological sign in a more semiquantitative manner. In an investigation of 41 psychogeriatric patients the authors recorded plantar responses as down, equivocally down, equivocal, equivocally up, and up. By this classification system, 59% agreement was found between two independent examinations by different observers. This poor agreement was only increased to 68% by collapsing the categories into three groups and interpreting a down and equivocally down as an agreement and an equivocally up and up as an agreement. Therefore, different grading systems did not appreciably decrease the percentage of disagreements for the plantar response. It is of further interest to note from these authors' data that an individual neurologist's agreement with himself on repeated attempts to elicit an

* Binomial distribution = \( P^2 + 2PQ + Q^2 \) where \( P^2 \) = probability of positive agreements; \( 2PQ \) = probability of disagreements; \( Q^2 \) = probability of negative agreements.

†This model assumes a probability of positives \( \leq 0.20 \) for any individual symptom or sign.
extensor plantar response in a group of individuals was essentially the same as the percent agreement between two neurologists on the same testing occasion. Whether this surprisingly low agreement of an examiner with himself is due to inherent variability or is the result of physiological variability of the patient was not evaluated. It appears in their data that intra-observer variability may be another major source of error which will require careful standardization in well-designed clinical studies.

Although based on a limited number of cases, the present study clearly documents that attention to the observer-variability problem is essential in clinical investigations concerning the course of patients with cerebrovascular disease. Even under the most optimal circumstances—examinations performed by trained neurologists, standardized protocols and careful definition of a positive or negative finding—only the most gross symptoms and, perhaps, signs can be expected to be reproducible. Until more efforts are directed to the study of this problem, clinical investigations in neurology will continue to report conflicting results.

We do not nihilistically imply that it is impossible to obtain any “true” histories and consistent neurological examinations. To do so would constitute an assault on common sense similar to that felt by Dr. Samuel Johnson when told of Bishop Berkeley’s questioning of the “real” world existing behind phenomena. We believe, however, that the present study shows the magnitude of variability of observed findings, particularly in the history. Undoubtedly this variability is at its maximum, for various reasons, in histories involving the “transient-attacks” syndrome.

There have been disagreements in the literature concerning the natural course of ischemic cerebrovascular disease, and concerning its course when modified by treatment. Some studies of use of anticoagulants in particular have shown striking decrease in transient ischemic attacks, and others have shown no difference between other control and treated populations (5, 6). In light of these disagreements, it would seem essential to know the limits of accuracy of the language in which we define results. Yet clinical results on some therapeutic studies have been given in terms of patients being “better” or “unimproved” or only slightly more specific terms, with no documentation of reproducibility of the clinical history and physical examination. Our results suggest that an unknown factor in these discrepant reported results is the “validity” of the history, and, to a lesser extent, the examination. It is of interest that similar problems in definition of data have occasionally been reported in non-neurological diseases (7). Can this inter-observer variability be controlled or overcome in order to derive “true” histories and examinations? The persistence of variability after examiners’ discussions shows the stubbornness of the problem; yet it seems apparent that awareness of this kind of variability of patient response may lead to greater freedom from artifacts in the future.

References

Discrepancies in Recorded Results from Duplicate Neurological History and Examination in Patients Studied for Prognosis in Cerebrovascular Disease

CHARLES SISK, DEWEY K. ZIEGLER and TURGUT ZILELI

Stroke. 1970;1:14-18
doi: 10.1161/01.STR.1.1.14

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://stroke.ahajournals.org/content/1/1/14