Diagnosis of Transient Ischemic Attacks:
Improvement of Interobserver Agreement by a
Check-List in Ordinary Language

P.J. Koudstaal, M.D.,* J. van Gun, M.D.,† A. Staal, M.D.,*
H.J. Duivenvoorden, Ph.D.,* J.G.M. Gerritsma, Ph.D.,† and C.L. Kraaijveld, M.D.‡

SUMMARY To try and improve the interobserver agreement for the diagnosis of TIA, we used a check-list in which the symptoms were recorded in plain language, instead of in abstract diagnostic terms such as amaurosis fugax. Criteria for a diagnosis of TIA were similarly phrased and recommended to all observers. Eight senior neurologists and ten neurology residents interviewed 72 patients in random pairs. In 64 cases the observers agreed on the diagnosis (kappa value = 0.77). After a short discussion between the two observers the agreement increased to a maximum (kappa = 1.0). However, in 29 of the 144 interviews the diagnosis would have been different, had the recommended criteria been fully applied. In 28 of these the observer had diagnosed TIA on insufficient evidence. Six “misinterpretations” led to disagreement for the diagnosis and only these were corrected by the observers during their discussion. The present design has led to a maximal agreement for the diagnosis of TIA between two observers, but the agreement between such a pair and the common diagnostic criteria was not yet ideal. The precision of the diagnosis could be improved if details of the required symptoms are discussed in general as well as for each patient.

Stoke Vol 17, No 4, 1986

THE DIAGNOSIS OF TIA is subject to considerable interobserver variation.1,4 In a previous study we found an agreement rate between seasoned neurologists of 0.65 (kappa value, adjusted for chance agreement).3

One of the reasons of interobserver disagreement may be that symptoms are usually recorded in diagnostic terms such as amaurosis fugax or dysarthria.2 This tendency to record inference rather than evidence is encouraged because international criteria6 for the diagnosis of TIA are phrased in the same manner. Therefore, the actual cause for disagreement for the diagnosis of TIA might be the lack of an exact definition of each of these diagnostic terms. Another reason might be that the content of acquired information differs between observers. For instance, repetition of the history tends to uncover new data.7

Measures for improving the interobserver agreement follow from the considerations mentioned above: (1) The symptoms should be recorded in ordinary words. If the diagnostic criteria are phrased in exactly the same way, diagnostic terms are redundant and interpretation is simplified. (2) Agreement could be improved if the observers employ a check-list.8,9 This is likely to improve uniformity because it discourages the interviewers from omitting parts of the history. (3) After the observers have independently interviewed the patient and recorded their conclusion, they should be offered the opportunity of mutual consultation.3,7

The main purpose of this study was to explore whether the interobserver agreement for the diagnosis of TIA could be improved by the three measures mentioned above. As a base-line we used the results of our previous study. Other studies could not be used for comparison since most investigated only the agreement for the individual symptoms and since chance agreement (kappa statistics) was not taken into account. In those cases in which our measures had failed, we wanted to investigate the reasons for the remaining disagreement in order to find new possibilities for improving the uniformity of the diagnosis. We therefore tried to answer the following questions: (1) Are disagreements on the diagnosis of TIA mainly explained by a difference in the content of the history or by a difference in interpretation? (2) Is the agreement rate among observers on the diagnosis of TIA dependent on the symptom category involved (e.g. vision, muscle strength, etc.)? (3) To what extent are recommended criteria for the diagnosis of TIA actually used by the observers and would the diagnosis be different if these were strictly applied?

Patients and Methods

During the study period, we selected all patients in whom a diagnosis of TIA was considered by the referring physician or by one of the residents of the department of neurology. Some patients had a minor physical deficit, but not so obvious that it could influence the history. Every patient was independently interviewed, within two days, by two physicians. These were either senior neurologists or residents with at least one year of clinical training in neurology. Eight senior neurologists and ten residents participated in the study. Six of the senior neurologists had also participated in our previous study. All belonged to the same University department of neurology. If a diagnosis of TIA was first suggested by one of the participating physicians, two others interviewed the patient. They were paired according to tables for random numbers10 into 32 pairs.
of two senior neurologists and 40 pairs of one senior neurologist and one resident.

During the study period (November 1983 to May 1984), 79 patients were found eligible. Seventy-two patients entered the study, 48 men and 24 women, ranging in age between 30 and 84 years, with a mean age of 60.3 years. Seven patients could not be included for the following reasons: more than two days between the interviews (2), absence of one or both observers (3), patient not consenting (1) or too low intelligence of the patient (1).

For the history a check-list was used, in which the observers had to record the symptoms as closely as possible in the words of the patient. For every item such as vision, muscle strength or speech, the check-list contained a number of possible symptoms in ordinary language, which could be ticked by the observers (Table 1). If none of these terms were adequate, the observer could write down the symptom himself, again in plain words. If the patient had experienced different kinds of attacks, these could be separately recorded, up to a maximum of three. Next on the list were multiple-choice questions about the mode of onset of the symptoms (in a split second, within seconds, minutes, hours or days), including the synchronicity of different symptoms, the duration of the attack, the mode of disappearance and the number of attacks. The last items of the check-list concerned the diagnosis: TIA, specified according to the vascular territory (carotid, either carotid or vertebrobasilar, vertebrobasilar, or unknown) or no TIA, including an alternative diagnosis. Finally, the observers could list the symptoms they considered strong arguments pro or contra TIA.

The recommended criteria for a diagnosis of TIA were based on arbitrary, but internationally accepted criteria, identical to those employed in a former study. For the purpose of the study these criteria were "translated" from abstract diagnostic terms into ordinary language. For instance, amaurosis fugax was defined as a complete loss of vision or black vision of one eye or of the upper or lower half of the visual field, with the exclusion of blurred, distorted or grey vision. The mode of onset should be within a few seconds and the duration of the attack at least one minute. The translation was made by the most senior neurologists of the department, who did not participate in the interviews. All observers were advised to use these criteria, which were included in the check-list as a supplement.

After both observers had independently taken the history and recorded their conclusions, a short discussion followed between them. They were instructed to exchange the arguments for their diagnosis, including the presumed vascular territory in case of a TIA. After this consultation they again independently recorded the final diagnosis, with an explanation if there was a change of opinion.

The degree of agreement between the two observers was measured by kappa statistics. Kappa = (Po - Pe)/(1 - Po), where Po is the observed percentage of agreement, and Pe is the percentage of agreement that is to be expected by chance when judgments are statistically independent. Kappa = 0.0 when there is just chance agreement and kappa = 1.0 when there is perfect agreement.

The interobserver agreement for the items "mode of onset," "mode of disappearance," "number of attacks" and "duration of the attack" was assessed by comparing all check-lists two by two. Next, the mean kappa value for all pairs of observers was determined.

All data from the check-lists were analyzed by computer. For each recorded symptom we checked whether it had been interpreted according to the recommended criteria. We also reconstructed the diagnosis that would have been reached if these criteria had been strictly applied. This computer diagnosis served as the "gold standard."

To explore the reasons for possible idiosyncrasies in interpretation we performed an enquiry among the observers in which they could express their own view on the importance of various symptoms.

**Results**

Before the Discussion between the Two Observers

a. Interobserver Agreement for the Diagnosis

In 39 cases, the observers agreed that the diagnosis was TIA, in 25 cases they both concluded "no TIA."

### Table 1 Part of the Check-list

<table>
<thead>
<tr>
<th>Vision</th>
<th>1st kind of attack*</th>
<th>2nd kind of attack*</th>
<th>3rd kind of attack*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>Nature of the abnormality</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>blurred, foggy</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>black, blind</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>grey</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>distorted</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>flickering</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>image moving with change of posture</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>double</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>also with one eye covered</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>horizontal</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>vertical</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>oblique</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>Distribution of the abnormality</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>one eye</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>both eyes</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>part of the visual field</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>left or right half</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>upper or lower half</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>central</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>peripheral</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>otherwise (nature and distribution)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*When the patient had experienced different kinds of attacks, these could be recorded separately.
Taken together, there was agreement in 64 of the 72 patients, which results in a kappa value of 0.77 ($P_o = 0.89$; $P_e = 0.52$). The agreement rate between two senior neurologists was not significantly different from that between a senior neurologist and a resident (kappa value 0.76 and 0.78, respectively). Agreement on the vascular territory, if a division was made between carotid, either carotid or vertebrobasilar attacks on the one hand and vertebrobasilar or unknown on the other, reached a kappa value of 0.65 ($P_o = 0.85$; $P_e = 0.57$).

b. The Agreement Rate for the Diagnosis of TIA According to the Symptom Category Involved

Table 2 shows the interobserver agreement on the diagnosis of TIA for each of the symptom categories. It appears that the agreement rate is not better for the identification of TLAs of one kind versus another.

With regard to the time scale of the symptoms, which is an important factor in the diagnostic decision, the mean kappa value after two by two analysis was 0.46 for the mode of onset, 0.54 for the mode of disappearance, 0.80 for the duration of the attack and 0.82 for the number of attacks.

c. Application of the Diagnostic Criteria

After reconstruction of the diagnosis on the basis of a computer analysis, which applied the recommended criteria as a “gold standard” to the results of each observer, it appeared that in 29 of the 144 interviews the diagnosis should have been different. These deviations involved 16 of the 18 physicians. In only one of these cases the physician diagnosed “no TIA” — according to the criteria — the recorded symptoms were sufficient for a diagnosis of TIA. In the other 28 cases “soft” symptoms were interpreted as a TIA, against the recommended criteria. This is illustrated in figure 1. In the same figure the results of the enquiry among the observers are given. The enquiry (histogram on the right) shows that the observers held widely different views on the interpretation of various symptoms, but most did not take extreme positions on the interpretation of equivocal symptoms. Comparison with the left side of figure 1 shows that such symptoms were more often interpreted as a TIA, when they came from the patient’s mouth than when they had to be judged on a questionnaire. The most striking example is the interpretation of blurred or foggy vision. Eight observers rated this symptom as absolutely incompatible with the diagnosis of TIA, six found it not sufficient for the diagnosis and four were neutral. Yet in practice, six interviews in which the patient actually mentioned blurred or foggy vision led to a diagnosis of TIA. In all six cases the observer found the symptom strong evidence for the diagnosis. No difference was found between the seasoned neurologists and the residents in the kind or number of “misinterpretations.”

During and after the Discussion between the Two Observers

a. Interobserver Agreement for the Diagnosis

After their discussion the two observers agreed in 43 cases on a diagnosis of TIA and in 29 cases on no TIA, which means maximal agreement (kappa value of 1.0). The consensus for the vascular territory increased to a kappa value of 0.77 ($P_o = 0.90$; $P_e = 0.57$).
b. Causes of Interobserver Disagreement

Of the eight cases in which the observers reached opposite conclusions before their discussion, a difference in information accounted for the disagreement in only two. This was easily corrected during the discussion. For instance, one patient had told the first observer only irrelevant and vague symptoms in spite of tenacious questioning, whereas the second observer obtained a clear history of a right-sided weakness, with sudden onset, lasting for fifteen minutes. After hearing this, the first observer also diagnosed TIA. The remaining six cases of disagreement before the discussion could be explained by a difference in interpretation, in all cases because one of the two observers had diagnosed TIA against the recommended criteria. This concerned the nature of the main symptom in two cases, the duration of the attack in two cases, the mode of onset in one case and a combination of the main symptom and the mode of onset in one case. In all six cases the “unorthodox” observer changed his interpretation according to the rules.

c. Application of the Diagnostic Criteria

In table 3, the common diagnoses of the observer pairs — after their discussion — are compared with the diagnoses that should have resulted from full application of the recommended criteria (computer diagnosis). In 57 of the 72 cases the observers’ diagnoses equalled those after computer analysis. A striking finding was that eight pairs of observers agreed on a diagnosis of TIA against the recommended criteria. This combined dissent concerned the main symptom in three cases (grey vision in all three), the mode of onset in three cases (within minutes 1, not all symptoms at the same time 1, in minutes and with a march 1), duration of the attack (seconds only) in one case and a combination of the main symptom (blurred vision) and mode of onset (within minutes) in one case. In the remaining seven cases the computer analysis resulted in a diagnosis of TIA from the information of one observer and a conclusion “no TIA” from that of the other, which means that one of the pair had not applied the recommended criteria. In six of these seven cases the rules were broken to diagnose TIA, in only one case to diagnose no TIA.

Discussion

In the absence of objective criteria for the diagnosis of TIA, improvement of the accuracy of the diagnostic process is not possible. All our efforts should therefore be directed at improving the precision of the diagnosis. In the present study, the use of (1) plain language in recording and interpreting the history and (2) a checklist in multiple-choice format resulted in an improvement of the interobserver agreement for the diagnosis of TIA in comparison with our previous results (kappa value 0.77 against 0.65 in the earlier study). The improvement could not be attributed only to a better
agreement between the six observers who participated in both studies. In contrast with our previous study, neurology residents also participated, but if anything, this would lead to a lower agreement rate in view of the greater number of observers\textsuperscript{13, 14} and the inclusion of less experienced physicians.\textsuperscript{15, 16} This makes the improvement the more convincing. The agreement for the vascular territory of presumed TIA\textsubscript{s} also improved (kappa value 0.65 against 0.36 in the previous study\textsuperscript{3}). After (3) a discussion between the two observers, the agreement on the diagnosis was maximal (kappa value 1.0).

In achieving maximal interobserver agreement for the diagnosis of TIA we answered the main purpose of this study. On the other hand, analysis of the underlying data showed that some sources of error had remained. First, the observers showed striking differences in the classification and chronological assessment of the individual symptoms, but apparently this hardly affected agreement on the ultimate diagnosis. This does not mean that the differences were always subtle: in 7 of the 72 patients no diagnosis could be made when the check-lists of both observers were analyzed by computer, because of essential differences in recorded data. This finding might be important if the aid of complex computer systems is used to diagnose TIA from the information of the physician.\textsuperscript{17} The validity of such a diagnostic procedure clearly depends on the ability of the physician to transmit the history as neutrally as possible to the computer. Studies on the improvement of agreement should concentrate on optimal classification of symptoms by nature and time course. This could be achieved by carefully adapting the check-list and by training the observers.

Second, analysis of the check-lists showed that a considerable number of symptoms were not interpreted according to the recommended criteria. This was at odds with our aim of improving the precision of the diagnosis. Apparently, the presence of such criteria in no way guarantees uniformity in interpretation. In 29 of the 144 interviews the initial diagnosis would have been different if the recommended criteria had been fully applied. It is striking that the rules were far more often broken to diagnose "TIA" (28 cases) than to diagnose "no TIA" (only one case). From the enquiry it appeared that the observers held widely different views on the interpretation of various symptoms, which partly explains why the recommended criteria were not consistently adhered to. These criteria, however, originated from internationally accepted diagnostic guidelines,\textsuperscript{6} which had been used for many years in our department and were implicitly endorsed by all participating observers. But these traditional criteria were phrased in abstract diagnostic terms, while the idiosyncrasies were apparently related to the "translation" of these criteria into plain language. For example, the observers have no difficulty in agreeing that amaurosis fugax is consistent with a TIA, but they often disagree on the actual definition of amaurosis. Even when observers did agree in theory that a particular symptom did not qualify for the diagnosis of TIA, they might still make this diagnosis when they had heard the same symptom from a real patient. It must of course be remembered that the recommended criteria are arbitrary guidelines, but the impressive variance of opinion among the observers about "atypical" symptoms makes it clear that more explicit definitions are necessary.

Although the number of interpretations against the recommended criteria is rather disappointing, there is good evidence that this might be further improved. In two previous studies\textsuperscript{18, 19} attention has been drawn to the necessity of prior discussion and agreement on the pieces of evidence required for a diagnosis. In both studies this subsequently improved the interobserver agreement on the ultimate diagnosis. Unfortunately, it is impossible to infer from the published data to what extent this diagnosis involved consultation of previously agreed criteria. Yet in all probability the number of "misinterpretations" in our study would have been smaller if the recommended criteria had not been merely added as an appendix to each form, but instead had been thoroughly discussed with all participating physicians before the start of the study.

Finally, it is surprising that the eight pair-wise deviations from the recommended rules were not discovered by the observers during their discussion. Apparently, in the absence of precise instructions for checking each other's history and subsequent interpretation, the observers quickly found out that they agreed on the diagnosis and then decided that they had nothing to talk about. This was confirmed by many observers after the study.

In conclusion, the present design has resulted in the maximal interobserver agreement for the diagnosis of TIA. However, as has been noted by others,\textsuperscript{2} we found that focusing only on the agreement between two observers does not exclude the possibility that both are "wrong" in the sense of not applying common diagnostic guidelines. This implies that the homogeneity of the group of patients classified as TIA can be further improved by a thorough discussion on the exact grounds for the diagnostic decision, and by a mutual check on the adherence to previously agreed rules.

Acknowledgments
The authors wish to thank all their colleagues who participated with admirable enthusiasm throughout this study, Dr. M. Vermeulen for valuable comment and Mrs. J. Doornbosch-Konijn and Mrs. R.M. van der Hoven for secretarial assistance.

References

The Effect of Intravenous Lidoflazine on
Serotonin-Induced Cerebral Vascular Contraction —
An In Vivo Study

ROBERT H. ROSENWASSER, M.D., RONALD F. TUMA, PH.D., AND WILLIAM A. BUCHHEIT, M.D.

SUMMARY Lidoflazine, a piperazine derivative with known selectivity for vascular smooth muscle, was evaluated as a possible agent for prophylaxis of cerebral vascular contraction induced by subarachnoid perfusion with serotonin. The animals treated with serotonin (5 × 10^-6 M), had a 60% reduction in the diameter of basilar artery but when pretreated with Lidoflazine (1 mg/kg) intravenously, only had a 20% reduction in diameter (p < 0.01). Lidoflazine, when administered intravenously at a slow rate will not adversely lower systemic blood pressure and can prevent the contraction of cerebral vessels when the stimulus for contraction is in the subarachnoid space.

CEREBRAL VASOSPASM after subarachnoid hemorrhage is a major cause of morbidity and mortality in this patient population. 1,4 Much interest has evolved in the use of calcium antagonists in the prophylaxis and reversal of cerebral vasospasm. 5 Many drugs evaluated experimentally in the prophylaxis of vasospasm have either been applied topically to a vessel or have been shown to have no effect when given intravenously because of poor lipid solubility or instability in solution. 6-8 Lidoflazine, a piperazine derivative and calcium antagonist, is a potent coronary artery dilator which has been used in Europe for the treatment of exertional and vasospastic angina. The purpose of this study was to determine the efficacy of this drug when administered intravenously to prophylax against basilar artery contraction produced by subarachnoid perfusion with serotonin.

Methods
Experiments were carried out in Sprague-Dawley rats weighing approximately 350 grams. The animals were initially anesthetized with 1.5% Halothane, nitrous oxide, and oxygen. After completion of tracheostomy, pancuronium bromide 0.05 mcg/kg was administered intravenously for skeletal paralysis and ventilation was controlled. Cannulae were placed in the femoral artery for determination of blood gases and measurement of blood pressure, the femoral vein for administration of fluids and drugs, and the internal jugular vein for measurement of central venous pressure. Anesthesia was then maintained with Fentanyl 20 mcg/kg and oxygen, a technique to minimize cerebral vascular reactivity and alterations in cerebral blood flow. 10, 11 The arterial blood pressure and central venous pressure were maintained in a physiologic range. The partial oxygen pressure was kept at 100–150 mm Hg and the partial carbon dioxide pressure was maintained at 38-42 mm Hg. Clivectomy was then performed and the dura was opened. The basilar artery and its branches, as well as pial circulation were readily identified. The arachnoid was opened and the subarachnoid space was perfused in the following manner:

From the Cerebrovascular Research Laboratory, Departments of Neurosurgery and Physiology, Temple University Hospital, Philadelphia, Pennsylvania. Address correspondence to: Robert H. Rosenwasser, M.D., Department of Neurosurgery, Temple University, 3401 North Broad Street, Philadelphia, Pennsylvania 19140.
Diagnosis of transient ischemic attacks: improvement of interobserver agreement by a check-list in ordinary language.

P J Koudstaal, J van Gijn, A Staal, H J Duivenvoorden, J G Gerritsma and C L Kraaijeveld

Stroke. 1986;17:723-728
doi: 10.1161/01.STR.17.4.723

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/17/4/723