Agreement between physicians on assessment of outcome following severe head injury

ANDREW I. R. MAAS, M.D., REINDER BRAAKMAN, M.D., HUBERT J. A. SCHOUTEN, PH.D., JAN M. MINDERHOUD, M.D., AND ADRIAAN H. VAN ZOMEREN, M.D.

Departments of Neurosurgery and Biostatistics, Erasmus University, Rotterdam, and Department of Neurology, University Hospital, Groningen, The Netherlands

This study describes inter-observer agreement between physicians and consistency of one observer on repeated scoring in the assessment of outcome 6 to 12 months after severe head injury. Observer agreement is expressed by Kappa, taking chance agreement into account. The study was conducted in “live” and “non-live” situations, using five- and eight-category outcome scales. Kappa values were considerably higher when the five-category scale was used. However, even with the five-category scale, disagreement was considerable, and the results indicate that accurate predictions of the quality of survival will be difficult to attain. Observer agreement should be taken into account in prognosis studies, and when different therapeutic regimens are evaluated on the basis of outcome.

KEY WORDS □9 head injury □9 observer variability □9 outcome □9 prognosis □9 treatment evaluation

In recent years, there has been debate about the relative efficacy of different therapeutic regimens in altering the prognosis of severely head-injured patients. The clinical value of prognostic factors and feasibility of predicting outcome in severe head injury are influenced both by the rate of agreement between physicians in assessing early clinical signs, and by the rate of agreement about outcome. Studies on observer agreement on the best motor response of the arms, the Glasgow Coma Scale, and eye movements and pupil reactivity have already appeared in the literature. Little attention has yet been paid to outcome.

We report studies of the inter-observer agreement between physicians and consistency of one observer on repeated scoring in assessing the outcome of severely head-injured patients.

Clinical Material and Methods

Patients

The 37 patients in this study were admitted to the Neurological Department of the University Hospital Groningen (Cases 1–17) and to the Neurosurgical Department of the University Hospital Rotterdam-Dijkzigt (Cases 18–37). All patients were in coma for at least 6 hours following severe head injury. “Coma” is defined as inability to open the eyes, to obey commands, and to utter recognizable words.

Assessment of Outcome

Outcome was assessed between 6 months and 1 year after injury. Details of neurophysical examination (motor pattern, dysphasia, apraxia, ataxia, cranial nerve palsy, epilepsy), mental function (learning, short- and long-term memory, distractibility, drive, psychiatric disturbances), and social function (work, family life, leisure activities, relationship problems) were obtained and recorded. The ability to perform normal activities of daily life, such as walking, eating, washing, dressing, use of toilet facilities, capability of independent transportation in traffic, and use of money and telephone, was also assessed.

Posttraumatic complaints like headache, dizziness, lack of concentration, irritability, sluggishness, sleep disturbances, and emotional lability were also recorded. All these data together determine overall social outcome; they were recorded on precoded data forms. The overall social outcome was scored on the five-point Glasgow Outcome Scale and on an eight-point scale (Table 1). This latter scale is an extension...
TABLE 1
Summary of the eight-category outcome scale

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Death</td>
<td>Death</td>
</tr>
<tr>
<td>2: Vegetative State</td>
<td>non-sentient, not obeying commands, no verbal contact, no meaningful response; may have sleep-wake rhythm, may have spontaneous eye opening and ability to follow moving objects, may swallow food</td>
</tr>
<tr>
<td>3 &amp; 4: Severe Disability</td>
<td>conscious, but dependent</td>
</tr>
<tr>
<td>3:</td>
<td>communication is possible, minimally by emotional response, total or almost total dependency with regard to activities of daily life</td>
</tr>
<tr>
<td>4:</td>
<td>partial independence in activities of daily life, may require assistance for only one activity, such as dressing; many evident posttraumatic complaints and/or signs; resumption of former life and work not possible</td>
</tr>
<tr>
<td>5 &amp; 6: Moderate Disability</td>
<td>independent, but disabled</td>
</tr>
<tr>
<td>5:</td>
<td>independence in activities of daily life; for instance, can travel by public transport; not able to resume previous activities, either at work or socially; despite evident posttraumatic signs, resumption of activities at a lower level is often possible</td>
</tr>
<tr>
<td>6:</td>
<td>posttraumatic signs are present, which, however, allow resumption of most former activities, either full-time or part-time</td>
</tr>
<tr>
<td>7 &amp; 8: Good Recovery</td>
<td>capable of resuming normal occupational and social activities; there are minor physical or mental deficits or complaints</td>
</tr>
<tr>
<td>8:</td>
<td>full recovery without signs or symptoms</td>
</tr>
</tbody>
</table>

The five-point Glasgow Outcome Scale, and does not differ greatly from the expanded version as presented by Jennett, et al. of physicians. The average of these proportions taken over all patients is called the observed proportion of agreement ($p_o$). It is an estimate of the probability that two physicians agree on the outcome category when they judge the same patient. Agreement between two physicians may, however, occur by chance. The proportion of chance-expected agreements ($p_c$) under statistical independence is an estimate of the probability that two physicians agree on the outcome category when they judge different patients. In our experience, $p_c$ may reach fairly high values and has to be taken into account.

The higher rate of agreement obtained by the physicians compared to chance agreement is $p_o - p_c$. The greatest possible improvement beyond chance is $1 - p_c$, occurring when all pairs of physicians agree. The ratio of obtained improvement to maximum improvement is termed Kappa: $Kappa = (p_o - p_c)/(1 - p_c)$, a measure of agreement corrected for chance. Kappa = 0 when only chance agreement exists, and Kappa = 1 when complete agreement exists. In this study, Kappa is denoted by $K_5$ and $K_8$ when five or eight outcome categories are used, respectively.

Relevant statistical papers have been written by Cohen, Fleiss, et al., and Schouten.

Results

**Inter-Observer Agreement**

Outcome was assessed in each of the 17 patients studied live (Cases 1–17) by four of a group of 10 different physicians; the results are summarized in Table 2. On the eight-category scale a
Outcome assessment of severe head injury

### TABLE 3

<table>
<thead>
<tr>
<th>Clinical Material</th>
<th>Five-Category Scale</th>
<th>Eight-Category Scale</th>
<th><em>po</em></th>
<th><em>pc</em></th>
<th><em>K5</em> ± SE</th>
<th><em>po</em></th>
<th><em>pc</em></th>
<th><em>K8</em> ± SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;live&quot; observations</td>
<td>0.86 0.40 0.77 ± 0.11</td>
<td>0.60 0.23 0.48 ± 0.10</td>
<td>0.90 0.48 0.83 ± 0.09</td>
<td>0.66 0.24 0.55 ± 0.10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cases 1–17†</td>
<td>0.72 0.34 0.58 ± 0.08</td>
<td>0.58 0.18 0.49 ± 0.08</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*po = observed proportion of agreement; *pc = chance expected proportion of agreement; *K5* = Kappa value for five-category scale; *K8* = Kappa value for eight-category scale.

† Four observers for each patient.
‡ Eleven observers for each patient.

---

**Intra-Observer Agreement**

Outcome was assessed in the non-live situation twice in 20 patients from Rotterdam by six observers, four from Rotterdam and two from Groningen. Likewise, 15 patients from Groningen were scored twice by four physicians from Rotterdam. Thus in all, 180 pairs of observations were obtained for intra-observer agreement analysis. On the eight-category scale a difference of one or more categories were seen in 60 cases. One observer scored a patient in the category "good recovery" the first time, and "severely disabled" in the second round of scoring. On six occasions a difference of two categories on the eight-category scale was noted.

**Intra-Observer Agreement**

Outcome was assessed in the non-live situation twice in 20 patients from Rotterdam by six observers, four from Rotterdam and two from Groningen. Likewise, 15 patients from Groningen were scored twice by four physicians from Rotterdam. Thus in all, 180 pairs of observations were obtained for intra-observer agreement analysis. On the eight-category scale a difference of one or more categories were seen in 60 cases. One observer scored a patient in the category "good recovery" the first time, and "severely disabled" in the second round of scoring. On six occasions a difference of two categories on the eight-category scale was noted.

---

**Intra-Observer Agreement**

Outcome was assessed in the non-live situation twice in 20 patients from Rotterdam by six observers, four from Rotterdam and two from Groningen. Likewise, 15 patients from Groningen were scored twice by four physicians from Rotterdam. Thus in all, 180 pairs of observations were obtained for intra-observer agreement analysis. On the eight-category scale a difference of one or more categories were seen in 60 cases. One observer scored a patient in the category "good recovery" the first time, and "severely disabled" in the second round of scoring. On six occasions a difference of two categories on the eight-category scale was noted.

---

**Intra-Observer Agreement**

Outcome was assessed in the non-live situation twice in 20 patients from Rotterdam by six observers, four from Rotterdam and two from Groningen. Likewise, 15 patients from Groningen were scored twice by four physicians from Rotterdam. Thus in all, 180 pairs of observations were obtained for intra-observer agreement analysis. On the eight-category scale a difference of one or more categories were seen in 60 cases. One observer scored a patient in the category "good recovery" the first time, and "severely disabled" in the second round of scoring. On six occasions a difference of two categories on the eight-category scale was noted.

---

**Intra-Observer Agreement**

Outcome was assessed in the non-live situation twice in 20 patients from Rotterdam by six observers, four from Rotterdam and two from Groningen. Likewise, 15 patients from Groningen were scored twice by four physicians from Rotterdam. Thus in all, 180 pairs of observations were obtained for intra-observer agreement analysis. On the eight-category scale a difference of one or more categories were seen in 60 cases. One observer scored a patient in the category "good recovery" the first time, and "severely disabled" in the second round of scoring. On six occasions a difference of two categories on the eight-category scale was noted.

---

**Intra-Observer Agreement**

Outcome was assessed in the non-live situation twice in 20 patients from Rotterdam by six observers, four from Rotterdam and two from Groningen. Likewise, 15 patients from Groningen were scored twice by four physicians from Rotterdam. Thus in all, 180 pairs of observations were obtained for intra-observer agreement analysis. On the eight-category scale a difference of one or more categories were seen in 60 cases. One observer scored a patient in the category "good recovery" the first time, and "severely disabled" in the second round of scoring. On six occasions a difference of two categories on the eight-category scale was noted.

---

**Intra-Observer Agreement**

Outcome was assessed in the non-live situation twice in 20 patients from Rotterdam by six observers, four from Rotterdam and two from Groningen. Likewise, 15 patients from Groningen were scored twice by four physicians from Rotterdam. Thus in all, 180 pairs of observations were obtained for intra-observer agreement analysis. On the eight-category scale a difference of one or more categories were seen in 60 cases. One observer scored a patient in the category "good recovery" the first time, and "severely disabled" in the second round of scoring. On six occasions a difference of two categories on the eight-category scale was noted.

---

**Intra-Observer Agreement**

Outcome was assessed in the non-live situation twice in 20 patients from Rotterdam by six observers, four from Rotterdam and two from Groningen. Likewise, 15 patients from Groningen were scored twice by four physicians from Rotterdam. Thus in all, 180 pairs of observations were obtained for intra-observer agreement analysis. On the eight-category scale a difference of one or more categories were seen in 60 cases. One observer scored a patient in the category "good recovery" the first time, and "severely disabled" in the second round of scoring. On six occasions a difference of two categories on the eight-category scale was noted.

---

**Intra-Observer Agreement**

Outcome was assessed in the non-live situation twice in 20 patients from Rotterdam by six observers, four from Rotterdam and two from Groningen. Likewise, 15 patients from Groningen were scored twice by four physicians from Rotterdam. Thus in all, 180 pairs of observations were obtained for intra-observer agreement analysis. On the eight-category scale a difference of one or more categories were seen in 60 cases. One observer scored a patient in the category "good recovery" the first time, and "severely disabled" in the second round of scoring. On six occasions a difference of two categories on the eight-category scale was noted.

---

**Intra-Observer Agreement**

Outcome was assessed in the non-live situation twice in 20 patients from Rotterdam by six observers, four from Rotterdam and two from Groningen. Likewise, 15 patients from Groningen were scored twice by four physicians from Rotterdam. Thus in all, 180 pairs of observations were obtained for intra-observer agreement analysis. On the eight-category scale a difference of one or more categories were seen in 60 cases. One observer scored a patient in the category "good recovery" the first time, and "severely disabled" in the second round of scoring. On six occasions a difference of two categories on the eight-category scale was noted.

---

**Intra-Observer Agreement**

Outcome was assessed in the non-live situation twice in 20 patients from Rotterdam by six observers, four from Rotterdam and two from Groningen. Likewise, 15 patients from Groningen were scored twice by four physicians from Rotterdam. Thus in all, 180 pairs of observations were obtained for intra-observer agreement analysis. On the eight-category scale a difference of one or more categories were seen in 60 cases. One observer scored a patient in the category "good recovery" the first time, and "severely disabled" in the second round of scoring. On six occasions a difference of two categories on the eight-category scale was noted.

---

**Intra-Observer Agreement**

Outcome was assessed in the non-live situation twice in 20 patients from Rotterdam by six observers, four from Rotterdam and two from Groningen. Likewise, 15 patients from Groningen were scored twice by four physicians from Rotterdam. Thus in all, 180 pairs of observations were obtained for intra-observer agreement analysis. On the eight-category scale a difference of one or more categories were seen in 60 cases. One observer scored a patient in the category "good recovery" the first time, and "severely disabled" in the second round of scoring. On six occasions a difference of two categories on the eight-category scale was noted.

---

**Intra-Observer Agreement**

Outcome was assessed in the non-live situation twice in 20 patients from Rotterdam by six observers, four from Rotterdam and two from Groningen. Likewise, 15 patients from Groningen were scored twice by four physicians from Rotterdam. Thus in all, 180 pairs of observations were obtained for intra-observer agreement analysis. On the eight-category scale a difference of one or more categories were seen in 60 cases. One observer scored a patient in the category "good recovery" the first time, and "severely disabled" in the second round of scoring. On six occasions a difference of two categories on the eight-category scale was noted.

---

**Intra-Observer Agreement**

Outcome was assessed in the non-live situation twice in 20 patients from Rotterdam by six observers, four from Rotterdam and two from Groningen. Likewise, 15 patients from Groningen were scored twice by four physicians from Rotterdam. Thus in all, 180 pairs of observations were obtained for intra-observer agreement analysis. On the eight-category scale a difference of one or more categories were seen in 60 cases. One observer scored a patient in the category "good recovery" the first time, and "severely disabled" in the second round of scoring. On six occasions a difference of two categories on the eight-category scale was noted.
When observers used the five-category Glasgow Outcome Scale, a change in scores between moderate and severe disability was seen eight times, between moderate disability and good recovery 25 times, and between severe disability and good recovery once. Thus, in all 180 pairs of observations, disagreement on the five-category scale was seen in 34 cases (19%). Results of statistical analysis of the intra-observer agreement are summarized in Table 6.

**TABLE 6**

<table>
<thead>
<tr>
<th>Physician</th>
<th>$p_o$</th>
<th>$K_o \pm SE$</th>
<th>$p_a$</th>
<th>$K_a \pm SE$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Groningen patients</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R1</td>
<td>0.93</td>
<td>0.89 ± 0.11</td>
<td>0.80</td>
<td>0.73 ± 0.14</td>
</tr>
<tr>
<td>R2</td>
<td>0.93</td>
<td>0.87 ± 0.12</td>
<td>0.87</td>
<td>0.80 ± 0.12</td>
</tr>
<tr>
<td>R3</td>
<td>0.67</td>
<td>0.43 ± 0.21</td>
<td>0.40</td>
<td>0.22 ± 0.16</td>
</tr>
<tr>
<td>R4</td>
<td>0.93</td>
<td>0.89 ± 0.11</td>
<td>0.87</td>
<td>0.82 ± 0.12</td>
</tr>
<tr>
<td><strong>Rotterdam patients</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R1</td>
<td>0.65</td>
<td>0.49 ± 0.14</td>
<td>0.65</td>
<td>0.56 ± 0.13</td>
</tr>
<tr>
<td>R2</td>
<td>0.75</td>
<td>0.61 ± 0.14</td>
<td>0.65</td>
<td>0.57 ± 0.12</td>
</tr>
<tr>
<td>R3</td>
<td>0.90</td>
<td>0.85 ± 0.10</td>
<td>0.75</td>
<td>0.69 ± 0.12</td>
</tr>
<tr>
<td>R4</td>
<td>0.60</td>
<td>0.40 ± 0.15</td>
<td>0.45</td>
<td>0.34 ± 0.13</td>
</tr>
<tr>
<td>G1</td>
<td>0.95</td>
<td>0.92 ± 0.08</td>
<td>0.75</td>
<td>0.69 ± 0.13</td>
</tr>
<tr>
<td>G2</td>
<td>0.65</td>
<td>0.51 ± 0.12</td>
<td>0.55</td>
<td>0.48 ± 0.11</td>
</tr>
</tbody>
</table>

* $K_o =$ Kappa value for the five-category scale; $K_a =$ Kappa value for the eight-category scale; $p_o =$ observed proportion of agreement.

**TABLE 7**

<table>
<thead>
<tr>
<th>Authors &amp; Year</th>
<th>Feature</th>
<th>Feature</th>
<th>Mean Kappa Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Braakman, et al., 1977</td>
<td>motor response (Glasgow Coma Scale)</td>
<td></td>
<td>0.66</td>
</tr>
<tr>
<td>van den Berge, et al., 1979</td>
<td>inequality of pupils</td>
<td></td>
<td>0.61</td>
</tr>
<tr>
<td>Hunt &amp; Hess classification, 1982</td>
<td>spontaneous eye movements</td>
<td></td>
<td>0.46</td>
</tr>
<tr>
<td>Nishioka classification, 1983</td>
<td>oculocephalic response</td>
<td></td>
<td>0.49</td>
</tr>
<tr>
<td>pupil reactivity to light</td>
<td></td>
<td></td>
<td>0.64</td>
</tr>
<tr>
<td>Lindsay, et al., 1982</td>
<td>inter-observer agreement $K_s$</td>
<td></td>
<td>0.73</td>
</tr>
<tr>
<td>Maas, et al., 1983</td>
<td>inter-observer agreement $K_s$</td>
<td></td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td>intra-observer agreement $K_s$</td>
<td></td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>intra-observer agreement $K_s$</td>
<td></td>
<td>0.59</td>
</tr>
</tbody>
</table>

* $K_s =$ mean Kappa value for five-category scale; $K_8 =$ mean Kappa value for eight-category scale.

**Discussion**

Observer disagreement may arise for various reasons: 1) change in response of the patient, for instance when the examination takes place at different times; 2) observer bias (that is, prejudice causing perception bias); 3) difference between physicians in eliciting the response; 4) difference in interpretation of the same response by various physicians; and 5) errors in recording observations.

In our study, observer disagreement will mainly result from observer bias, different interpretation of the patient’s condition by various observers, and errors in recording data. Jennett, et al., emphasized that a source of bias in assessing overall outcome arises from judging this outcome in relation to the known initial severity of injury. This bias is particularly likely to occur when outcome is judged by those who were responsible for early management. Outcome should preferably be assessed by physicians not involved in previous management and without knowledge of the severity of injury or of preceding scores. It is difficult to meet this requirement in outpatient departments; it could only be fulfilled in the non-live part of the present study as details of the clinical course were withheld. It was interesting that some of the physicians in our study expressed difficulty in scoring outcome without knowledge of the immediate post-injury condition.

A most important source of disagreement is whether information is obtained from the patient or from the relatives. The patients in this study often presented a far more optimistic view of their state and of their abilities than their relatives. A minimal requirement for the assessment of outcome is that both the patient and his relatives are interviewed personally by a physician, preferably not involved in the management, either at home or in the outpatient department. Assessment of outcome based on information obtained by telephone or from letters, either from the patient or his family, may be highly unreliable.

In our opinion, a written definition of the various outcome categories should be consulted before the actual scoring takes place. Often an observer reported that he was contemplating a score of 5, for instance, but then changed his mind to a 6 after he consulted the definition of the various categories.

The question arises whether a five- or an eight-category outcome scale is to be preferred. For the description of outcome in individual patients, many clinicians believe that the five-category scale is too inaccurate, because patients who differ markedly in their condition may be classified in one group. The low Kappa values obtained in the use of the eight-category scale, however, cast doubt on the reliability of practical application of the eight-category scale. For research purposes, such as prognosis studies and in the comparison of results of different therapeutic regimens, the five-category Glasgow Outcome Scale is definitely preferable.

Our study demonstrates considerable disagreement, even in the use of the five-category scale. In the live study, unanimous agreement on the five-category scale was seen in 13 of 17 patients studied (76%), but in only six patients on the eight-category scale (35%). In the non-live situation, disagreement was similar. Even the same observer’s scores varied remarkably when assessment was repeated. The results show slightly better agreement than results obtained in agreement studies of early prognostic signs in severe
Outcome assessment of severe head injury

head injury,2,10,20 or with systems for grading patients with subarachnoid hemorrhage11 (Table 7).

Our results have important clinical implications. It is hard to achieve a sharp prediction of the degree of disability following severe head injury, since the disagreement between the physicians who were actually scoring outcome was considerable. The problems in comparing different series of patients with severe head injury will be even greater when the definitions of the various outcome categories used in different centers are not identical, and when assessment of outcome depends on information obtained by letter or telephone, thus inducing a lower observer agreement rate. Observer variation in assessing outcome must be taken into account when alternative therapeutic regimens are compared.

Acknowledgments

The authors thank their colleagues who took part in the assessment of outcome.

References

16. Schouten HJA: Measuring pairwise interobserver agreement when all subjects are judged by the same observers. Stat Neerland 36:45–61, 1982

Manuscript received July 28, 1982.
Address for Drs. Minderhoud and van Zomeren: Department of Neurology, University Hospital, Groningen, The Netherlands.
Address reprint requests to: Andrew I. R. Maas, M.D., Department of Neurosurgery, Academisch Ziekenhuis Rotterdam-Dijkzigt, Erasmus University, Dr. Molewaterplein 40, 3015 GD Rotterdam, The Netherlands.