Interobserver Variability of Grading Scales for Aneurysmal Subarachnoid Hemorrhage

Lieveke A.R. Degen, MD; Sanne M. Dorhout Mees, MD; Ale Algra, MD; Gabriël J.E. Rinkel, MD

Background and Purpose—Worldwide, different scales are used to assess the clinical condition on admission after aneurysmal subarachnoid hemorrhage. In addition to the prognostic value, the inter-rater variability should be taken into account when deciding which scale preferably should be used. We assessed the interobserver agreement of the commonly used World Federation of Neurological Surgeons, the Hunt and Hess, and the Prognosis on Admission of Aneurysmal Subarachnoid Hemorrhage scales.

Methods—In a cohort of 50 subarachnoid hemorrhage patients, 103 paired assessments were performed on the 3 admission scales by 2 independent observers per assessment with a total of 57 different raters. Patients were assessed during the first week after the hemorrhage. The interobserver agreement was calculated using quadratic (weighted) kappa statistics.

Results—The weighted kappa value of the Prognosis on Admission of Aneurysmal Subarachnoid Hemorrhage scale was 0.64 (95% CI, 0.49–0.79), of the World Federation of Neurological Surgeons scale was 0.60 (95% CI, 0.48–0.73), and of the Hunt and Hess scale was 0.48 (95% CI, 0.36–0.59).

Conclusions—The Hunt and Hess scale showed the lowest interobserver agreement, whereas agreement of the World Federation of Neurological Surgeons and Prognosis on Admission of Aneurysmal Subarachnoid Hemorrhage scales was similar with overlapping CI. (***Stroke***. 2011;42:1546-1549.)

Key Words: aneurysm ■ clinical condition ■ observer variability ■ subarachnoid hemorrhage

A neurysmal subarachnoid hemorrhage (SAH) still has a poor prognosis. Approximately one-third of the patients die and one-third of the surviving patients remain dependent on care for their activities of daily life.[1] The clinical condition at admission is an important factor in predicting functional outcome after SAH. Several scales are used to measure the clinical condition at admission. The Hunt and Hess (H&H) scale is a 5-category grading scale that classifies patients with SAH according to their surgical risk.[2] Another widely used scale is that of the World Federation of Neurological Surgeons (WFNS).[3] This scale was published in 1988 and is based on the Glasgow Coma Scale (GCS), but with focal deficits (motor deficit or aphasia) comprising 1 additional level for patients with a GCS score of 13 or 14. The cut-off points are based on consensus, not on a formal analysis. The Prognosis on Admission of Aneurysmal Subarachnoid Hemorrhage scale (PAASH) has recently been developed by Japanese investigators.[4] In developing this scale, the cut-off points between the categories were selected by calculating at which point 2 consecutive categories corresponded to a statistically different outcome at 6 months. The WFNS and PAASH scales both have a good discriminatory ability with regard to patient prognosis, but the PAASH scale has a more gradual increase in risk for poor outcome in successive categories than the WFNS.[5] A summary of the 3 scales on clinical condition at admission is shown in Table 1. The WFNS and H&H scales are both commonly used in clinical practice and in research; the PAASH scale is not yet widely used because the scale has been introduced only recently.

For the sake of uniformity, preferably a single scale is used among different centers to assess the clinical condition on admission in SAH patients. In addition to the prognostic value, the inter-rater variability should be taken into account when deciding which scale should be applied. Knowledge of interobserver variability is important both in a clinical setting and in a research setting. For clinicians, it is important to know whether different grading at 2 assessments at different points in the clinical course really means a change in the clinical situation or whether it can be the result of interobserver variation. For researchers, it is important to know if series of patients with dissimilar grading at baseline are really dissimilar. Also, in clinical trials with multiple assessors, a good interobserver agreement of the scale used to assess baseline clinical condition is pivotal to assure proper comparison between the treatment groups.

So far, interobserver variability of the different scales has only been investigated in a few observations.[6,7] In this study, we have compared the inter-rater variability of the WFNS, H&H, and PAASH scales in patients with aneurysmal SAH.

Table 1.   Summary of the Hunt and Hess Scale, the World Federation of Neurological Surgeons Scale, and the Prognosis on Admission of Aneurysmal Subarachnoid Hemorrhage Scale

Grade	Hunt and Hess Scale	WFNS Scale	PAASH Scale
I	Asymptomatic, or minimal headache and slight nuchal rigidity	GCS score 15	GCS score 15
II	Moderate to severe headache, nuchal rigidity, no neurological deficit other than cranial nerve palsy	GCS score 13–14 without focal deficit	GCS score 11–14
III	Drowsiness, confusion, or mild focal deficit	GCS score 13–14 with focal deficit	GCS score 8–10
IV	Stupor, moderate to severe hemiparesis, possibly early decerebrate rigidity, and vegetative disturbances	GCS score 7–12	GCS score 4–7
V	Deep coma, decerebrate rigidity, moribund appearance	GCS score 3–5	GCS score 3

GCS indicates Glasgow Coma Scale; PAASH, Prognosis on Admission of Aneurysmal Subarachnoid Hemorrhage; WFNS, World Federation of Neurological Surgeons.

Subjects and Methods

Subjects

A cohort of consecutive patients with aneurysmal SAH admitted to our hospital between April 2009 and March 2010 was enrolled in our study. Patients were included only when they had been admitted within 4 days after the hemorrhage. The diagnosis of aneurysmal SAH was confirmed with the presence of blood on CT (or, in absence thereof, on presence of xanthochromia in the cerebrospinal fluid) in combination with an aneurysm shown on CT or angiography.

To obtain sufficient observations in a short time window, patients could be seen on the day of admittance to our hospital and on 2 other days during the first week of hospitalization. Each time, the patient was assessed by 2 different raters. Assessments were performed blinded to the assessments of the other rater. One of the researchers was always present at time of assessments and took care that the raters independently assessed the patients and independently completed the assessment forms to ensure blinding. A rater could not rate the same patient more than once. Because SAH patients can sometimes deteriorate rapidly, particularly on the first day, no more than 30 minutes were allowed between paired assessments.

The raters consisted of 4 neurologists (with a total of 12 assessments; 5.8%), 17 residents in neurology or neurosurgery (47 assessments; 22.8%), 5 neurology interns (28 assessments; 13.6%), 3 research nurses (51 assessments; 24.8%), and 28 nurses working on the neurology medium care or intensive care unit (68 assessments; 33.0%). Because the PAASH scale and the WFNS scale are derived directly from the GCS, and because all medical and paramedical personnel in our department are familiar with this scale, it was not necessary to provide any training to the raters in the use of these scales. If raters were not familiar with the H&H scale, then they were trained in using it. The training was provided by 1 researcher. If patients were intubated or had aphasia, then a clinically possible GCS score was assigned based on interpretation of the rater and derived from their eye and motor scores. Only 5 patients were assessed this way.

The study was approved by the ethics committee of our hospital. The committee decided that informed consent could be waived because evaluation of the clinical condition of the patient is part of standard daily care.

Data Analysis

For each scale, the inter-rater agreement was analyzed with STATA version 10.1 (StataCorp LP, College Station, TX). Quadratic κ statistics were used, which correct for agreement by chance. If there is perfect agreement between 2 observers, then κ is 1. When κ is 0, the inter-rater agreement is no greater than would be expected by chance. Because all scales consisted of >2 categories, the weighted κ, which takes partial agreements into account, was also calculated. We used interpretation of κ for the agreement between clinical measures as follows: poor ($\kappa<0.20$), fair ($\kappa=0.21-0.40$), moderate ($\kappa=0.41-0.60$), good ($\kappa=0.61-0.80$), and very good agreement ($\kappa=0.81-1.00$).[8]

Results

In a cohort of 50 patients, 103 pairs of assessments were performed. The patients' characteristics are given in Table 2. The assessments between 103 pairs of observers and unweighted and weighted κ values are presented in Table 3. The weighted κ of the PAASH scale was 0.64 (95% CI, 0.49–0.79), of the WFNS scale was 0.60 (95% CI, 0.48–0.73), and of the H&H scale was 0.48 (95% CI, 0.36–0.59). Perfect agreement for the 2 observers was obtained for 81 (79%) with the PAASH scale, for 75 (73%) assessments with the WFNS scale, and for 64 (62%) with the H&H scale. For the PAASH scale, there were no assessments in which scores differed >1 grade. For the WFNS scale there were 5 assessments and for the H&H scale there were 6 assessments in which scores differed by 2 grades.

Discussion

Our study shows that the H&H scale has the lowest κ and κ of the WFNS and PAASH scales are similar, with overlapping CI.

In a previous study, the PAASH scale has been validated in an independent SAH patient population and it was compared with the WFNS scale.[5] Both scales showed a good prognostic value for patient outcome, but the PAASH scale showed a more gradual increase in risk for poor outcome in ascending categories. Furthermore, the PAASH scale is easier to apply in clinical practice because it is based solely on the GCS. Therefore, we have a preference for the PAASH scale to assess clinical condition in SAH patients.

Table 2.   Baseline Characteristics

Characteristics	N (%)
Total	50
Women	36 (72)
Mean age (range)	54 (19–77)
Total no. of ratings	103
No. of ratings per patient	
1	14 (28)
2	20 (40)
3	15 (30)
4	1 (2)
Mean time between observers for ratings, min (range)	7 (1–30)

Table 3.   Assessments Between 103 Pairs of Observations and Kappa Values for the World Federation of Neurological Surgeons, Hunt and Hess, and the Prognosis on Admission of Aneurysmal Subarachnoid Hemorrhage Scales

Observer 2	1	2	3	4	5	Total
World Federation of Neurological Surgeons Observer 1						
1	37	9				46
2	9	27	1	2		39
3		2	3			5
4		3	1	4		8
5				1	4	5
Total	46	41	5	7	4	103
Unweighted kappa value (95% CI)						0.58 (0.44–0.71)
Weighted kappa value (95% CI)						0.60 (0.48–0.73)
Prognosis on Admission of Aneurysmal Subarachnoid Hemorrhage Observer 1						
1	37	8				45
2	10	40	1			51
3		1				1
4			2	2		4
5					2	2
Total	47	49	3	2	2	103
Unweighted kappa value (95% CI)						0.62 (0.46–0.78)
Weighted kappa value (95% CI)						0.64 (0.49–0.79)
Hunt and Hess Observer 1						
1	18	4	2			24
2	12	32	3			47
3	4	9	8	1		22
4			2	4	1	7
5				1	2	3
Total	34	45	15	6	3	103
Unweighted kappa value (95% CI)						0.45 (0.33–0.57)
Weighted kappa value (95% CI)						0.48 (0.36–0.59)

CI indicates confidence interval.

In another previous study, an unweighted κ value of 0.43 ($P<0.001$) was found for the H&H scale, which is similar to the κ value of 0.45 we found.[6] As in our study, few patients with H&H grade 4 or 5 were included.

In a study from Baltimore, the interobserver variability of the WFNS scale, the H&H scale and a GCS-based grading system was examined.[7] This GCS-based grading system was distinct from the PAASH scale we used, because the 15-point GCS was categorized intuitively in a 5-grade scale and not based on calculations between the GCS and outcome. The resulting categories from the Baltimore scale differ from these in the PAASH scale. The unweighted κ values were 0.27 (P of κ statistic$=0.027$) for the WFNS scale, 0.41 ($P=0.0005$) for the H&H scale, and 0.46 ($P=0.0002$) for the GCS-based grading system. The κ value of the WFNS scale is considerably lower than in our study. This difference might be attributable to the fact that only 15 paired assessments were performed in the study. Also, the time between the assessments was not described and a weighted κ value was not calculated in that study.

Because the WFNS scale and the PAASH scale are both based on the GCS, our high κ values could be explained by a good interobserver agreement of the GCS. The interobserver variability of the GCS has been extensively studied. Most studies found a good reproducibility between different observers;[9,10] however, another study found a moderate agreement when comparing inexperienced raters with expert raters.[11] Our personnel is well-trained in using the GCS, which can explain the good agreement of the WFNS and PAASH scales, and might be worse when used by inexperienced raters.

A limitation of our study might be that we rated patients not only at admission but also on 2 other days during their first week of admission, whereas the scales are meant to be used as a prognostic indicator only on the day of admission. We used the scales in a different setting to limit our time of research. Yet, the aim of this study was not to assess the prognostic accuracy of the scales, but to compare the interobserver variability. The prognostic accuracy had been reported previously.[5] Because we kept the time interval in

between the assessments short, we consider that the difference in setting has barely affected our results.

Another limitation is that few patients with a poor clinical condition were included. Therefore, less precise statements can be made about the interobserver variability of patients who were assigned a grade 4 or 5.

Strengths of our study are the large number of observations in comparison to other studies and the large number of observers from different backgrounds and having levels of education and experience. Moreover, in most instances raters did not have an equivalent involvement in the care of the particular patient at the time of paired assessments. Despite this large number of and variation in observers and unequal involvement in care of the assessed patient, moderate and good κ values were found. Therefore, we expect that our results can be generalized to other settings in which there are many different professions involved in stroke care.

Another strength is the short time interval between the assessments. For that reason, it is unlikely that patients had fluctuations in the severity of their symptoms between the first and the second assessment.

In conclusion, the H&H scale showed the lowest interobserver agreement, whereas agreement of the WFNS and PAASH scales were similar, with overlapping CI. Given the similarity in interobserver reliability of the WFNS and PAASH scales, the easier applicability and the more gradual risk for poor outcome in successive categories of the PAASH scale than the WFNS scale, we prefer the PAASH scale to assess clinical condition in SAH patients.

Acknowledgments

Disclosures

References

1. Nieuwkamp DJ, Setz LE, Algra A, Linn FH, de Rooij NK, Rinkel GJ. Changes in case fatality of aneurysmal subarachnoid haemorrhage over time, according to age, sex, and region: a meta-analysis. *Lancet Neurol*. 2009;8:635–642.
2. Hunt WE, Hess RM. Surgical risk as related to time of intervention in the repair of intracranial aneurysms. *J Neurosurg*. 1968;28:14–20.
3. Report of World Federation of Neurological Surgeons Committee on a Universal Subarachnoid Hemorrhage Grading Scale. *J Neurosurg*. 1988;68:985–986.
4. Takagi K, Tamura A, Nakagomi T, Nakayama H, Gotoh O, Kawai K, Taneda M, Yasui N, Hadeishi H, Sano K. How should a subarachnoid hemorrhage grading scale be determined? A combinatorial approach based solely on the Glasgow Coma Scale. *J Neurosurg*. 1999;90:680–687.
5. van Heuven AW, Dorhout Mees SM, Algra A, Rinkel GJ. Validation of a prognostic subarachnoid hemorrhage grading scale derived directly from the Glasgow Coma Scale. *Stroke*. 2008;39:1347–1348.
6. Lindsay KW, Teasdale G, Knill-Jones RP, Murray L. Observer variability in grading patients with subarachnoid hemorrhage. *J Neurosurg*. 1982;56:628–633.
7. Oshiro EM, Walter KA, Piantadosi S, Witham TF, Tamargo RJ. A new subarachnoid hemorrhage grading system based on the Glasgow Coma Scale: a comparison with the Hunt and Hess and World Federation of Neurological Surgeons Scales in a clinical series. *Neurosurgery*. 1997;41:140–148.
8. Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. *BMJ*. 1992;304:1491–1494.
9. Teasdale G, Knill-Jones R, van der Sande J. Observer variability in assessing impaired consciousness and coma. *J Neurol Neurosurg Psychiatry*. 1978;41:603–610.
10. Lindsay KW, Teasdale GM, Knill-Jones RP. Observer variability in assessing the clinical features of subarachnoid hemorrhage. *J Neurosurg*. 1983;58:57–62.
11. Rowley G, Fielding K. Reliability and accuracy of the Glasgow Coma Scale with experienced and inexperienced users. *Lancet*. 1991;337:535–538.

Stroke

American Heart Association

American Stroke Association®

Interobserver Variability of Grading Scales for Aneurysmal Subarachnoid Hemorrhage

Lieveke A.R. Degen, Sanne M. Dorhout Mees, Ale Algra and Gabriël J.E. Rinkel

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/42/6/1546