ORIGINAL RESEARCH

Observer Agreement in the Assessment of Endovascular Aneurysm Therapy and Aneurysm Recurrence

H.J. Cloft
T. Kaufmann
D.F. Kallmes

BACKGROUND AND PURPOSE: Assessments of completeness of endovascular cerebral aneurysm therapy are commonly reported in the literature. We studied several aneurysm assessment scales with regard to observer variability, which directly affects validity of these scales.

MATERIALS AND METHODS: Initial aneurysm occlusion and occlusion at a follow-up angiogram at 3–6 months were assessed independently by 2 experienced observers. Assessments of each aneurysm were made using 3 different scales: 4-response (complete, dog ear, neck remnant, incomplete), 3-response (complete, near-complete, incomplete), and 2-response (complete or near-complete, incomplete). Assessments were also made of comparisons of initial treatment angiogram with follow-up angiogram using 2 different scales: 3-response (better, same, worse) and 2-point response (not worse, worse).

RESULTS: With assessments of both initial and follow-up angiograms, interobserver and intraobserver agreement was progressively worse with increasing response choices in the scales. Observer agreement on assessments of initial angiograms (κ values 0.48–0.67) was worse than that for follow-up angiograms (κ values 0.66–0.97). For the comparisons of the initial angiogram with the follow-up angiogram, there was worse observer agreement with the 3-response scale (κ values 0.64–0.71) than with the 2-response scale (κ values 0.78–0.89).

CONCLUSION: Interobserver and intraobserver variability are inherent to assessment scales of completeness of cerebral aneurysm therapy. Observer variability is substantially better in scales that offer fewer observer responses. However, scales with fewer observer responses may not identify aneurysm subgroups that have differing risks of recurrence and/or rehemorrhage.

The goal of endovascular therapy of cerebral aneurysms is to eliminate blood flow to as much of the aneurysm as can safely be achieved. Completeness of occlusion of an aneurysm is commonly used as a measure of the success of therapy. Angiographic assessments of completeness of aneurysm therapy have been widely used but never formally validated. Raymond et al[1,2] described a scale with 4 categories: "complete," "dog ear," "residual neck," and "residual aneurysm." Perhaps because of subjectivity in classifying aneurysms as "residual neck" and "residual aneurysm," and an uncertainty about the clinical relevance of such a discrimination, Raymond et al[3] later simplified their scheme to 3 classifications; "complete," "residual neck," and "residual aneurysm." Murayama et al[4] classified completeness of aneurysm therapy as "complete," "neck remnant," and "incomplete." Other authors have described a classification based on percentage of aneurysm filling (eg, ">90% occlusion"[5-8] or ">95% occlusion"[9-11]). Such percentages are based not on any real quantitative measurement but on subjective assessments not inherently different from those described by Raymond et al[1-3] and Murayama et al.[4]

As new technology emerges that offers potential improvements in completeness and durability of cerebral aneurysm therapy, it becomes more important to validate scales used for assessment of aneurysm therapy. Valid assessments of completeness and durability of aneurysm therapy will be needed to judge the efficacy of such new technology in clinical trials. Thus, we undertook a study to assess the observer variability in several scales that assess the completeness of aneurysm therapy and change in completeness of aneurysm occlusion on follow-up angiography.

Received February 7, 2006; accepted after revision May 4.

From the Department of Radiology, Mayo Clinic, Rochester, Minn.

Address correspondence to Harry J. Cloft, MD, PhD, Department of Radiology, Mayo Clinic, 200 First St SW, Rochester, MN 55905; e-mail: cloft.harry@mayo.edu

Materials and Methods

Aneurysms Studied and Readers

Angiograms from a total of 125 aneurysms obtained immediately before and after endovascular therapy and 83 angiograms obtained 3 to 6 months after endovascular therapy were assessed using the scales described below. These angiograms were obtained in the course of the HydroCoil for Endovascular Aneurysm Occlusion (HEAL) registry. Each angiogram was assessed by 2 readers who were interventional neuroradiologists (H.J.C. and D.F.K.), each with more than 7 years of experience with endovascular aneurysm therapy. Both readers were employed at the same institution and worked closely together. Each reader made assessments on 2 occasions separated by at least 30 days in an attempt to diminish recall bias. The readers had not participated in the treatment of any of the aneurysms included in the study, and they were blinded to all patient information other than the images to be read. Each reader was blind to the other observer's assessments and to his own previous assessments.

Assessment Scales

Each initial post-treatment angiogram and each 3–6-month follow-up angiogram was assessed using 3 different scales of completeness of aneurysm therapy. The 4-response scale included the categories "complete," "dog ear," "residual neck," and "residual aneurysm," as described previously.[1,2] The 3-response scale included the categories "complete," "near-complete," and "incomplete." The 2-response scale included the categories "complete or near-complete," and "incomplete." Change between initial postprocedure angiograms and

Table 1: Comparison of interobserver agreement for assessment of aneurysm treatment using 4-, 3-, and 2-response scales and "better/same/worse" scale

Assessment	Concordant Rate p	κ (95% CI)
Initial angiogram		
4-Response	.58	0.50 (0.38–0.61)*
3-Response	.71	0.54 (0.42–0.67)*
2-Response	.94	0.63 (0.40–0.87)†
Follow-up angiogram		
4-Response	.63	0.66 (0.55–0.77)*
3-Response	.72	0.67 (0.55–0.79)*
2-Response	.96	0.87 (0.74–0.99)†
Comparing initial with follow-up		
Better/same/worse	.80	0.68 (0.55–0.81)*
Not worse/worse	.95	0.89 (0.78–0.99)†

Note:—CI indicates confidence interval.
* Weighted κ.
† Simple κ.

follow-up angiograms was assessed using 2 scales. The 3-response scale included the categories "better," "same," and "worse." The 2-response scale included the categories "not worse" and "worse."

Statistical Analysis. Proportion of concordant readings (p) and κ statistics were calculated to determine the relative interobserver and intraobserver agreement of the various aneurysm assessment scales. The proportion of concordant readings was calculated as the number of concordant readings divided by the total number of readings. Simple κ values were determined for assessments that had 2 responses, and weighted κ values were determined for assessments that had more than 2 possible responses. The statistical analysis was performed by using SAS 8.02 software (SAS Institute, Cary, NC).

The κ values obtained were interpreted relative to the criteria of Fleiss.[12] According to Fleiss, values $\leq$0.40 represent poor agreement, values between 0.40 and 0.75 represent fair to good agreement, and values $>$0.75 represent excellent agreement.

Results

The results are summarized in Tables 1 and 2.

Angiographic Completeness of Aneurysm Therapy

In general, scales with more allowable responses were associated with less agreement than scales with fewer allowable responses. The 4-response scale yields the least interobserver and intraobserver agreement for both the initial angiogram and the follow-up angiogram. The 3-response scale yields more agreement than the 4-response scale. The 2-response scale yields more agreement than the 3-response scale.

Intraobserver agreement, stated as κ, for follow-up angiograms assessed with the 4-response and 3-response scales ranged from 0.75 to 0.78, whereas for interobserver agreement, they ranged from 0.66 to 0.67. This indicates less interobserver than intraobserver agreement in the categorization of aneurysms when more choices are available in the scale.

For initial postprocedure angiogram assessments, the κ values for interobserver and intraobserver comparisons were 0.48 to 0.67 for all scales used. For follow-up angiogram assessments, the κ values for interobserver and intraobserver comparisons were 0.66 to 0.97 for all scales used. Thus, both interobserver and intraobserver agreement was better on follow-up angiograms than initial angiograms for all scales used. Examples of cases with discordant readings are shown in Fig 1.

Assessment of Change at Follow-up Angiogram

For assessments of angiographic change on the follow-up angiogram relative to the initial angiogram, the κ values for interobserver and intraobserver comparisons were 0.64 to 0.71 for the "better, same, worse" scale, whereas they were 0.78 to 0.89 for the "not worse, worse" scale.

Discussion

Our study elucidates the observer variability inherent in the angiographic assessment of completeness of cerebral aneurysm therapy. With assessments of both initial and follow-up angiograms for completeness of aneurysm therapy, interobserver and intraobserver agreement was progressively worse with increasing response choices in the aneurysm assessment scale. Likewise, for the comparisons of the initial angiogram to the follow-up angiogram, there was worse observer agreement with the 3-response ("better, same, worse") scale than with the 2-response ("not worse, worse") scale. These findings are supported by previous work demonstrating that the value of κ tends to increase as the number of categories is decreased, thus indicating better agreement when fine distinctions are eliminated.[13]

There was less agreement on initial angiograms than on 3–6-month follow-up angiograms. This finding can be explained by a high degree of subjectivity in the assessment of areas within an aneurysm that still fill with contrast material immediately after endovascular aneurysm therapy. The interpreter is left to guess whether these areas will thrombose over the next few hours.

Our study used 2 readers of very similar training and practice background. Interobserver variability between readers with more varied backgrounds might be even higher.

The degree of aneurysm occlusion after endovascular therapy is really a continuous variable. Therefore, commonly used scales for grading the degree of aneurysm occlusion apply an ordinal scale on what is really a continuous variable. This system forces a reader to categorize an outcome into a discrete category, which is rather subjective and can contribute to observer variability (Fig 1). This practice evolved because it is quite difficult to accurately quantify the continuous variable of "completeness of aneurysm occlusion" using 2-dimensional images of irregular 3D objects. The creation of ordinal categories forces the image interpreter to subjectively categorize the completeness of aneurysm treatment into one of the categories based on 2-dimensional images. 3D angiographic imaging now available might provide a more quantitative assessment of aneurysm occlusion, but standardized performance of such 3D measurement techniques is not currently developed enough to be practical for use on all, or even most, of the angiographic equipment used at each of the medical centers that might participate in a multicenter study of aneurysm therapy. Thus, the use of ordinal scales for the assessment of completeness of aneurysm occlusion remains as an important tool that potentially allows for the comparison of treatment data from multiple medical centers and multiple studies.

Because cerebral aneurysm therapy is intended to prevent subarachnoid hemorrhage, the future risk of hemorrhage from a cerebral aneurysm is ideally what would be measured to judge the success of therapy. Because endovascular therapy is rather successful at preventing subarachnoid hemorrhage,

Table 2: Comparison of intraobserver agreement for assessment of aneurysm treatment using 4-, 3-, and 2-response scales, and "better/same/worse" scale

	Observer 1		Observer 2	
Assessment	Concordant Rate *p*	κ (95% CI)	Concordant Rate *p*	κ (95% CI)
Initial angiogram				
4-Response	.70	0.67 (0.57–0.77)*	.63	0.48 (0.34–0.62)*
3-Response	.74	0.65 (0.54–0.76)*	.69	0.50 (0.35–0.65)*
2-Response	.91	0.66 (0.47–0.84)†	.94	0.56 (0.28–0.85)†
Follow-up angiogram				
4-Response	.75	0.78 (0.69–0.87)*	.77	0.78 (0.68–0.88)*
3-Response	.78	0.75 (0.65–0.86)*	.83	0.77 (0.65–0.88)*
2-Response	.99	0.97 (0.91–1.00)†	.96	0.89 (0.77–1.00)†
Comparing initial with follow-up				
Better/same/worse	.82	0.71 (0.58–0.84)*	.78	0.64 (0.49–0.78)*
Not worse/worse	.93	0.80 (0.66–0.94)†	.92	0.78 (0.63–0.92)†

Note:—CI indicates confidence interval.
* Weighted κ.
† Simple κ.

Fig 1. Examples of observer variability in the assessment of completeness of cerebral aneurysm therapy. A carotid aneurysm before (*A*) and immediately after (*B*) endovascular therapy, which was assessed variably as "complete," "dog ear," and "incomplete."

hemorrhagic events in these patients are quite uncommon. With such a low event rate for subarachnoid hemorrhage after endovascular therapy, studies of very large numbers of patients over a number of years would be necessary for rate of subarachnoid hemorrhage to be a practical outcome measure. The International Subarachnoid Aneurysm Trial (ISAT)[14] is a trial that has enrolled enough patients (1073 treated with coils) and follows them over a long enough period of time (at least 5 years) that it can make a meaningful assessment of the rehemorrhage rate. However, it is not practical for all studies of outcomes of endovascular cerebral aneurysm therapy to be so large. Thus, investigators use completeness of coil therapy and angiographic recurrence as surrogate markers for risk of future hemorrhage, making the assumption that risk of future hemorrhage correlates with degree of persistent filling of the aneurysm with contrast during angiography.

Another factor that makes assessment of degree of aneurysm occlusion important is that the completeness of aneurysm therapy at the time of treatment is somewhat predictive of subsequent risk of recurrence. For aneurysms that were angiographically completely treated, Murayama et al[4] reported a recurrence rate of 1.1% for small aneurysms with small necks, 12.5% for small aneurysms with wide necks, and 20% in large aneurysms. For aneurysms with a neck remnant, the recurrence rates were 7.7% for small aneurysms with small necks, 29.4% for small aneurysms with wide necks, and 44.4% for large aneurysms. Raymond et al[3] reported major recurrence rates of 9% for aneurysms initially completely occluded, 23% for aneurysms with residual neck, and 47% for aneurysms with residual aneurysm. Thornton et al[11] reported a recurrence rate of 1.8% for aneurysms that were "100%" occluded, and 26% for aneurysms with "residual neck." As an aneurysm progressively recanalizes after endovascular therapy, the degree of protection of the patient from rehemorrhage is expected to decrease. Because the presence of a residual neck is predictive of future recanalization, it is a worthwhile distinction to make.

Scales with fewer observer responses may not identify aneurysm subgroups that have differing risks of recurrence and/or rehemorrhage. Without identification of lesser degrees of aneurysm remnant and recurrence, it would not be possible to ascertain such issues as the extent to which a minor recurrence predicts a higher risk of future hemorrhage and the risk that a minor recurrence will turn into a major recurrence. In choosing a scale for use in a trial of endovascular aneurysm therapy, a balance must be struck between reducing observer variability by offering fewer responses and improving identification of important subgroups by offering more responses.

Change in degree of aneurysm occlusion on follow-up angiography is a critical variable to record because it specifically assesses the major weakness of endovascular therapy relative to surgery (ie, aneurysm recurrence). Scales of degree of aneurysm occlusion can be deceptive with regard to recurrence, because an aneurysm may have a worsening degree of occlu-

sion at follow-up angiography, yet it may not change categories on a grading scale. For example, an aneurysm treated and initially classified as "residual aneurysm" may show interval increase in the size of the residual aneurysm cavity at follow-up angiography and yet remain classified simply as "residual aneurysm"; ie, the degree of occlusion of the aneurysm changes, but the classification does not. Raymond et al[3] defined a recurrence as "any increase in size of the remnant" and defined a recurrence as "major" if "it was saccular and its size would theoretically permit retreatment with coils." Murayama et al[4] defined "recanalization" as a "more than 10% increase in contrast filling of the aneurysm." We chose to make the assessments as objective as possible by making the scales qualitative rather than quantitative.

One might suspect that an important marker of failure of aneurysm therapy is the need for retreatment of the aneurysm. The necessity of aneurysm retreatment, however, is not something that can be objectively measured as an end point for scientific research. The decision to retreat is quite subjective in many cases. The risk of treating relative to not treating a recurrent aneurysm can be quite difficult to ascertain in many cases, because the risk of future rupture of a given partially treated aneurysm is largely unknown. In our experience, recurrences are often more technically challenging to treat than the original, untreated aneurysm. Physicians vary greatly in terms of how aggressively they would treat an aneurysm recurrence with balloon remodeling, an adjunctive stent, or perhaps referral for surgery. Patients may refuse retreatment for a variety of reasons. The treating physician might have subjective reservations about further therapy based on experiences at the time of treatment ("you had to be there") or based on intimate knowledge of the patient ("you have to know this patient"). The operating physician may be biased against retreating an aneurysm that he or she originally treated because of an unwillingness to admit failure of treatment. Conversely, a physician may be biased toward retreating an aneurysm that he or she originally treated because of a perception of the recurrence as a personal failure that he or she would like to confront. Ideally, such personal biases should be left out of patient care decisions, but they undoubtedly have some effect on management decisions. With so much uncertainty, bias, and subjectivity, simply tallying whether an aneurysm is retreated has little value in scientifically assessing the success of aneurysm treatment.

Scales to assess angiographic completeness of aneurysm therapy and recurrence will continue to be necessary for the development of new endovascular therapies. Such scales are of critical importance in assessing the efficacy of new aneurysm therapies in clinical trials. To minimize bias, assessment of angiogram studies of aneurysm treatment should be made by a central reader rather than the treating physician. Because of the potential for interobserver variability, especially with assessment scales with more than 2 responses, it may be useful to have more than 1 reader make the assessments and to have readers try to review difficult cases together to reach consensus.

Conclusions

Both interobserver and intraobserver variability are inherent to assessment scales of completeness of cerebral aneurysm therapy. Observer variability is substantially lower in scales that offer fewer observer responses. However, scales with fewer observer responses may not identify aneurysm subgroups that have differing risks of recurrence and/or rehemorrhage. These limitations must be considered when using such scales in clinical trials that assess the efficacy of endovascular cerebral aneurysm treatment.

References

1. Raymond J, Roy D, Bojanowski M, et al. **Endovascular treatment of acutely ruptured and unruptured aneurysms of the basilar bifurcation.** *J Neurosurg* 1997;86:211–19
2. Roy D, Raymond J, Bouthillier A, et al. **Endovascular treatment of ophthalmic segment aneurysms with Guglielmi detachable coils.** *AJNR Am J Neuroradiol* 1997;18:1207–15
3. Raymond J, Guibert F, Weill A, et al. **Long-term angiographic recurrences after selective endovascular treatment of aneurysms with detachable coils.** *Stroke* 2003;34:1398–403
4. Murayama Y, Nien YL, Duckwiler G, et al. **Guglielmi detachable coil embolization of cerebral aneurysms: 11 years' experience.** *J Neurosurg* 2003;98:959–66
5. Murayama Y, Vinuela F, Duckwiler GR, et al. **Embolization of incidental cerebral aneurysms by using the Guglielmi detachable coil system.** *J Neurosurg* 1999;90:207–14
6. Eskridge JM, Song JK. **Endovascular embolization of 150 basilar tip aneurysms with Guglielmi detachable coils: results of the Food and Drug Administration multicenter clinical trial.** *J Neurosurg* 1998;89:81–86
7. Kuether TA, Nesbit GM, Barnwell SL. **Clinical and angiographic outcomes, with treatment data, for patients with cerebral aneurysms treated with Guglielmi detachable coils: a single-center experience.** *Neurosurgery* 1998;43:1016–25
8. Richling B, Gruber A, Bavinski G, et al. **GDC-system embolization for brain aneurysms—location and follow-up.** *Acta Neurochir (Wien)* 1995;134:177–83
9. Cognard C, Weill A, Spelle L, et al. **Long-term angiographic follow-up of 169 intracranial berry aneurysms occluded with detachable coils.** *Radiology* 1999;212:348–56
10. Byrne JV, Sohn MJ, Molyneux AJ, et al. **Five-year experience in using coil embolization for ruptured intracranial aneurysms: outcomes and incidence of late rebleeding.** *J Neurosurg* 1999;90:656–63
11. Thornton J, Debrun GM, Aletich VA, et al. **Follow-up angiography of intracranial aneurysms treated with endovascular placement of Guglielmi detachable coils.** *Neurosurgery* 2002;50:239–49
12. Fleiss JL. *Statistical Methods for Rates and Proportions.* New York: Wiley; 1981
13. Kundel HL, Polansky M. **Measurement of observer agreement.** *Radiology* 2003;228:303–08
14. Molyneux A, Kerr R, Stratton I, et al. **International Subarachnoid Aneurysm Trial (ISAT) of neurosurgical clipping versus endovascular coiling in 2143 patients with ruptured intracranial aneurysms: a randomised trial.** *Lancet* 2002;360:1267–74