Noninvasive Angiographic Results of Clipped or Coiled Intracranial Aneurysms: An Inter- and Intraobserver Reliability Study


*AJNR Am J Neuroradiol* 2021, 42 (9) 1615-1620
doi: https://doi.org/10.3174/ajnr.A7236
http://www.ajnr.org/content/42/9/1615
Noninvasive Angiographic Results of Clipped or Coiled Intracranial Aneurysms: An Inter- and Intraobserver Reliability Study


ABSTRACT

BACKGROUND AND PURPOSE: Noninvasive angiography is commonly used to assess the outcome of surgical or endovascular treatment of intracranial aneurysms in clinical series or randomized trials. We sought to assess whether a standardized 3-grade classification system could be reliably used to compare the CTA and MRA results of both treatments.

MATERIALS AND METHODS: An electronic portfolio composed of CTAs of 30 clipped and MRAs of 30 coiled aneurysms was independently evaluated by 24 raters of diverse experience and training backgrounds. Twenty raters performed a second evaluation 1 month later. Raters were asked which angiographic grade and management decision (retreatment; close or long-term follow-up) would be most appropriate for each case. Agreement was analyzed using the Krippendorff α (αk) statistic, and the relationship between angiographic grade and clinical management choice, using the Fisher exact and Cramer V tests.

RESULTS: Interrater agreement was substantial (αk = 0.63; 95% CI, 0.55–0.70); results were slightly better for MRA results of coiling (αk = 0.69; 95% CI, 0.56–0.76) than for CTA results of clipping (αk = 0.58; 95% CI, 0.44–0.69). Intrarater agreement was substantial to almost perfect. Interrater agreement regarding clinical management was moderate for both clipped (αk = 0.49; 95% CI, 0.32–0.61) and coiled subgroups (αk = 0.47; 95% CI, 0.34–0.54). The choice of clinical management was strongly associated with the size of the residuum (mean Cramer V = 0.77 [SD, 0.14]); but complete occlusions (grade 1) were followed more closely after coiling than after clipping (P = .01).

CONCLUSIONS: A standardized 3-grade scale was found to be a reliable and clinically meaningful tool to compare the results of clipping and coiling of aneurysms using CTA or MRA.

ABBREVIATION: $\alpha_k$ = Krippendorff α

The main goal of intracranial aneurysm treatments is to prevent ruptures or rebleeding. However, because such events may be devastating, many clinicians verify angiographic results to determine the success of therapy in each patient.1-3 Ruptures or rebleeding are relatively infrequent. Thus, angiographic results are often selected as outcome measures in clinical trials comparing aneurysm treatments.4-8 However, the repeatability of angiographic outcome measures must be verified before widespread use.

Conventional angiography, the criterion standard to diagnose aneurysms and assess the results of treatment, has been increasingly replaced by noninvasive CTA and MRA in the past decades.9,10 One problem is that noninvasive angiographic modalities

Received February 16, 2021; accepted after revision April 28.

From the Department of Radiology (A.B., B.F., G.G., W.B., D.R., A.W., D.I., F.G., L.L.-G., J.R.), Department of Medicine, Division of Neurology (G.J.), and Division of Neurosurgery (Z.K.), and Department of Radiology (J.S.G.), Health Sciences Centre, Halifax, Nova Scotia, Canada; Departments of Neurology, Neurosurgery, and Radiology (T.N.N.), Boston Medical Center, Boston, Massachusetts; Departments of Interventional Neuroradiology (J.-C.G.) and Neurosurgery (E.M.), Hôpital de la Cavale Blanche, Centre Hospitalier Régional et Universitaire de Brest, Brest, France; Department of Neuroradiology (J.S.G.), Centre Hospitalier Régional et Universitaire de Tours, Tours, France; Department of Neuroradiology (G.F.), University Hospital of Limoges, Limoges, France; Department of Medical Imaging (J.-F.H.), University Hospital Timoine Assistance Publique - Hôpitaux de Marseille, Marseille, France; University Hospital of Neuchâtel, Switzerland; and Department of Surgery (T.E.D.), Division of Neurosurgery, Walter C. Mackenzie Health Sciences Centre, University of Alberta Hospital, Edmonton, Alberta, Canada.
Table 1: Characteristics of patients and aneurysms included in the portfolio

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Clipped (n = 30)</th>
<th>Coiled (n = 30)</th>
<th>Total (n = 60)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Women</td>
<td>24 (80%)</td>
<td>22 (73%)</td>
<td>46 (77%)</td>
</tr>
<tr>
<td>Age (yr)</td>
<td>54.5 (SD, 10.4)</td>
<td>51.1 (SD, 13.0)</td>
<td>52.8 (SD, 11.8)</td>
</tr>
<tr>
<td>Ruptured aneurysms</td>
<td>15 (50%)</td>
<td>20 (67%)</td>
<td>35 (58%)</td>
</tr>
<tr>
<td>Retreated aneurysms</td>
<td>4 (13%)</td>
<td>6 (20%)</td>
<td>10 (17%)</td>
</tr>
<tr>
<td>Anterior circulation</td>
<td>28 (93%)</td>
<td>19 (63%)</td>
<td>47 (78%)</td>
</tr>
<tr>
<td>Aneurysm size (mm)</td>
<td>11.0 (2–27)</td>
<td>9.7 (3–30)</td>
<td>10.4 (2–30)</td>
</tr>
<tr>
<td>Wide neck (&gt;4 mm)</td>
<td>17 (56%)</td>
<td>14 (47%)</td>
<td>31 (52%)</td>
</tr>
</tbody>
</table>

Data are (No.) (%), mean (SD), or mean (minimum–maximum).

FIG 1. Illustrative cases from the portfolio of CTA clipped or MRA coiled aneurysms for each grade. The left MCA (A), anterior communicating artery (B), and right MCA (C) are CTA examples of clipped aneurysms, while the left superior cerebellar artery (D), basilar tip (E), and left MCA (F) are MRA examples of coiled aneurysms. Grade 1 (A and D): complete occlusion; grade 2 (B and E): residual neck (≤2 mm using visual estimation only); grade 3 (C and F): residual aneurysm.

are used differently depending on treatments: Surgically managed patients are often followed by CTA, while patients treated with coils are more often followed by MRA. When one judges the comparative success of therapy in clinical reports, it would seem that comparing CTA results of clipping with MRA results of coiling would be problematic, given the different diagnostic accuracies of the 2 modalities. This problem is particularly relevant for clinical trials. We cannot require a catheter angiogram solely for the purposes of the study when safer, noninvasive tests clinically suffice for most patients. The problem is compounded by the proliferation of grading scales, many of which are tailored to various devices and treatment modalities.11,12 A standardized method of reporting angiographic results that would facilitate comparisons between treatments and imaging modalities is needed.

A simple, 3-grade classification for the adjudication of results of clipped and coiled aneurysms has previously been shown to be reproducible when applied to conventional angiography. The questions that remained unanswered after a systematic review and that we sought to address in this work were the following: 1) Can the same angiographic classification system be used to evaluate CTA results of clipping and MRA results of coiling? 2) Are the results repeatable when judged by various raters? 3) Does the grade of occlusion obtained by clipping and assessed by CTA or by coiling and assessed by MRA have the same meaning in terms of clinical management?

MATERIALS AND METHODS

This article was written in accordance with the Guidelines for Reporting Reliability and Agreement Studies.16

Patient Selection

An electronic portfolio of 30 clipped and 30 coiled aneurysms was constructed. For each treatment technique, we aimed to include a wide spectrum of patients, with a balanced ratio (1:1:1) of completely occluded, residual, and gray-zone aneurysm cases to minimize the paradoxes of \( \kappa \) statistics.17,18 The number of patients per treatment group was estimated to be sufficient (>24) according to recommendations.19,20 For each patient, a high-definition video of the axial MRA or CTA sequences of the clipped or coiled aneurysm was provided. Patients with ruptured and unruptured aneurysms were selected from the clinical series of 1 tertiary care center (Center Hospitalier de l’Université de Montréal). Patient and aneurysm characteristics are summarized in Table 1.

Grading Scale

The grading scale is a variant of the Raymond-Roy classification.21 Categories of the standardized 3-grade classification system included the following: 1, complete occlusion; 2, residual neck (defined as ≤2 mm using visual estimation); and 3, residual aneurysm (Fig 1).15 Raters were not trained in the use of this classification system before the assessment.

Raters

Thirty-two clinicians were invited to participate: 24 (75%) raters (11 interventional neuroradiologists, 7 neurosurgeons, 4 interventional neurologists, and 2 diagnostic neuroradiologists) from 4 different countries accepted. Twenty raters agreed to perform a second evaluation of the cases in a permuted order at least 1 month later. There were 10 senior raters with >10 years of experience. Two of the interventional neuroradiologists had experience as core lab reviewers. Rater characteristics are shown in the Online Supplemental Data.

Agreement Study

An electronic survey was created and sent to the raters using the REDCap online data base manager (https://www.project-redcap.org/) hosted at the Center Hospitalier de l’Université de Montréal. Rater characteristics are shown in the Online Supplemental Data.
For each of the 60 cases, raters were asked to assess the grade of occlusion and to choose the most appropriate clinical management, assuming all angiographic results concerned a ruptured aneurysm in a 65-year-old patient with a good outcome and no other medical problems. Possible options were the following: follow-up imaging in 3–5 years (or none at all); close follow-up (6–18 months); and immediate retreatment by endovascular means; or immediate retreatment by surgical means. The last 2 choices were then merged as immediate retreatment (either by surgical or endovascular means). Clinically meaningful differences were also assessed for all cases and were defined in accordance with McDonald et al.24 as cases for which at least one rater recommended follow-up (close or delayed) and another rater recommended retreatment (surgical or endovascular).

**Statistical Analysis**

All calculations were performed using R 3.5.3 statistical and computing software (http://www.r-project.org/).

Inter- and intrarater agreement for the grading scale and for the clinical management choices was estimated using the Krippendorff α (κK) statistic, and the 95% confidence intervals were determined using 1000 bootstrap iterations. Interpretation of κK values was given in accordance with Landis and Koch.25 Comparisons of proportions of ratings between prespecified aneurysm and rater subgroups as well as the strength of the association between the raters’ angiographic verdict and the management of the patient were evaluated using the Fisher exact test followed by a Cramer V test, with a significance threshold of .05.

**RESULTS**

**Grading Scale**

The number of aneurysms judged to be completely occluded (grade 1) by various raters varied between 10 (17%) and 36 (60%). Similarly, residual aneurysms (grade 3) were judged to be present in 11 (18%) to 36 (60%) patients. Perfect agreement among all 24 raters was found in 7/60 (12%) patients or in 22/60 (37%) after dichotomization of the scale into absence or presence of a residual aneurysm (grades 1 + 2 versus 3). The distribution of angiographic verdicts differed between clipped and coiled aneurysms (P = .01): Clipped aneurysms were more often judged to be completely occluded, and coiled aneurysms were more often judged to have residual necks, while residual aneurysms were similarly allocated (Online Supplemental Data).

Overall interrater agreement of the grading scale for all raters and all patients was substantial (κK = 0.63; 95% CI, 0.55–0.70). When treatment and imaging modalities were considered separately, agreement was substantial for clipped aneurysms followed by MRA (κK = 0.69; 95% CI, 0.56–0.85) and moderate for coiled aneurysms followed by CTA (κK = 0.58; 95% CI, 0.44–0.69), yet with overlapping confidence intervals. Better agreement for coiled cases assessed by MRA than clipped aneurysms assessed by CTA was also a trend for all rater subgroups. Senior raters performed no better than juniors, and training background had no effect (Table 2).

Individual intrarater agreement was at least substantial for all raters and varied between 0.66 and 0.89. There were no significant differences between the mean intrarater agreement of the subgroups defined according to experience or training background. (Online Supplemental Data).

**Clinical Management**

After we dichotomized the clinical management into follow-up or immediate retreatment (options 1–2 versus 3), perfect agreement among all 24 raters was reached in 23/60 (38%) cases. The number of cases for which delayed follow-up imaging was chosen ranged from 0 (0%) to 36 (60%); those with close follow-up, from 12 (20%) to 53 (88%); and those with immediate retreatment, from 7 (12%) to 26 (43%) of 60 cases, depending on the rater. Sixty-five percent of cases (39/60) had a clinically meaningful difference in the choice of recommended management. Retreatment (n = 345 judgments of 1440) was more frequently endovascular than surgical (92%: n = 317/345), a preference shown by both clippers (88%) and coilers (95%).

Interrater agreement for the 3 clinical options (1, delayed follow-up; 2, close follow-up; 3, immediate retreatment) was moderate for both clipped (κK = 0.49; 95% CI, 0.32–0.61) and coiled aneurysm subgroups (κK = 0.47; 95% CI, 0.34–0.54) and was so regardless of years of experience. Raters with core lab experience (n = 2) were the only raters who substantially agreed for clipped (κK = 0.75; 95% CI, 0.55–0.88) and coiled (κK = 0.67; 95% CI, 0.38–0.87) subgroups (Online Supplemental Data).

Individual intrarater agreement ranged from fair (κK = 0.31; 95% CI, 0.06–0.51) to almost perfect (κK = 0.88; 95% CI, 0.78–0.96). The mean overall interrater agreement was substantial (κK = 0.68 [SD, 0.14]) and did not change significantly according to rater background or experience (Online Supplemental Data).

### Table 2: Interrater agreement for the grading scale

<table>
<thead>
<tr>
<th></th>
<th>CTA, Surgical (n = 30)</th>
<th>MRA, Endovascular (n = 30)</th>
<th>Total (n = 60)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Overall</strong></td>
<td>0.58 (0.44–0.69)</td>
<td>0.69 (0.56–0.76)</td>
<td>0.63 (0.55–0.70)</td>
</tr>
<tr>
<td><strong>Experience</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Junior (0–10 years; n = 14)</td>
<td>0.59 (0.43–0.72)</td>
<td>0.68 (0.56–0.76)</td>
<td>0.64 (0.54–0.71)</td>
</tr>
<tr>
<td>Senior (&gt;10 years; n = 10)</td>
<td>0.56 (0.40–0.68)</td>
<td>0.69 (0.55–0.77)</td>
<td>0.62 (0.52–0.69)</td>
</tr>
<tr>
<td><strong>Background</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Core lab (n = 2)</td>
<td>0.74 (0.54–0.87)</td>
<td>0.81 (0.63–0.9)</td>
<td>0.77 (0.64–0.86)</td>
</tr>
<tr>
<td>INR (n = 1)</td>
<td>0.54 (0.38–0.68)</td>
<td>0.65 (0.53–0.74)</td>
<td>0.60 (0.50–0.68)</td>
</tr>
<tr>
<td>DNR (n = 2)</td>
<td>0.58 (0.23–0.80)</td>
<td>0.87 (0.72–0.97)</td>
<td>0.73 (0.55–0.85)</td>
</tr>
<tr>
<td>NSX (n = 7)</td>
<td>0.67 (0.54–0.77)</td>
<td>0.71 (0.56–0.81)</td>
<td>0.69 (0.59–0.76)</td>
</tr>
<tr>
<td>INL (n = 4)</td>
<td>0.53 (0.30–0.71)</td>
<td>0.72 (0.57–0.82)</td>
<td>0.63 (0.50–0.73)</td>
</tr>
</tbody>
</table>

Note: INL indicates interventional neuroradiology; DNR, diagnostic neuroradiology; NSX, neurosurgery; INL, interventional neurology.

*a* Data are κκ (95% CI) for the ordinal 3-grade classification. Classification grades: 1, complete occlusion; 2, residual neck (<2 mm); 3, residual aneurysm.
or using randomized trials. It is important if we are to learn and progress from clinical experience. Verifying the reliability of treatment outcome measures is particularly important when comparing outcomes of various aneurysm treatments assessed by different noninvasive angiographic modalities.15,26-28 Even if interrater agreement was suboptimal, the substantial level of agreement achieved overall among raters, regardless of treatment or imaging technique and experience or background, is somewhat reassuring. Furthermore, intrarater agreement was at least substantial for all raters.

**DISCUSSION**

Many scales have been proposed to grade angiographic results of various aneurysm treatments, but a previous review has shown that few have proved to be reliable and none have previously attempted to evaluate multiple treatment or imaging modalities at the same time.15 Yet, a common language is needed to share experiences and to meaningfully compare outcomes of various treatments assessed by different noninvasive angiographic modalities. Verifying the reliability of treatment outcome measures is important if we are to learn and progress from clinical experience or using randomized trials.

In the present work, we demonstrated variability within and between raters in adjudicating angiographic results of clipping or coiling using CTA or MRA. Perhaps unsurprisingly, agreement was, for many rater and aneurysm subgroups, less concordant than when results were assessed using conventional angiography.15,26-28 Even if intrarater agreement was suboptimal, the substantial level of agreement achieved overall among raters, regardless of treatment or imaging technique and experience or background, is somewhat reassuring. Furthermore, intrarater agreement was at least substantial for all raters.

The exact same angiographic result, obtained by one or the other treatment technique, may not have the same clinical significance in terms of clinical management. This is why we attempted to verify the clinical pertinence of the angiographic verdict of each rater for each case, assuming that the angiographic result concerned the same patient. Interrater agreement regarding the clinical management of the hypothetic case with various angiographic results was only moderate overall. This finding is not so surprising when one considers that this measure combines both disagreement in the assessment of angiographic results and disagreement in the clinical management of the same adjudicated residuum. When intrarater agreement is compared, more variability is introduced at the level of clinical decisions than at the time of the angiographic verdict (Online Supplemental Data).

Divergence in clinical management may be explained by the diversity of raters from different specialties, hospitals, and countries who may apply different follow-up protocols. This feature may also explain the relatively high number of cases with clinically meaningful differences in the management decisions (65%), which are comparable with the number in previous studies.24 Another important observation is the apparent different clinical meaning of a complete occlusion documented by CTA postcoiling compared with a complete occlusion demonstrated by MRA postcoiling: Raters seemed more confident to opt for delayed follow-up when aneurysms were completely clipped, while being often inclined to follow completely coiled aneurysms more closely (Online Supplemental Data). This choice has also been previously observed with conventional angiography results.15 If the clinical meaning of grade 1 angiographic result differs when it is obtained by CTA on clipped aneurysms or by MRA on coiled aneurysms, by contrast, the allocation of a residual aneurysm had a more reliable clinical meaning, at least in terms of retreatment, no matter the imaging or treatment technique (Online Supplemental Data).

Given the inherent pitfalls of using a surrogate angiographic outcome measure, the impossibility of blinding assessors to the treatment received, and the variability in the clinical significance of complete occlusion (grade 1) and residual neck (grade 2) shown in the present study, we believe that the residual aneurysm...
CONCLUSIONS
Noninvasive angiographic results of clipping or coiling of aneurysms can be reliably reported by raters of various experience and backgrounds using a standardized classification system. The proposed classification was shown to be clinically meaningful, with each grade being strongly correlated to a different management option. This classification could be used to standardize results of published randomized trials, registries, or case series.


REFERENCES
19. Donner A, Rotondi MA. Sample size requirements for interval estimation of the kappa statistic for interobserver agreement studies
with a binary outcome and multiple raters. Int J Biostat 2010;6:Article 31 CrossRef Medline
25. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74 CrossRef Medline