Observer Agreement in the Angiographic Assessment of Arteriovenous Malformations of the Brain

Rustam Al-Shahi, MA, MRCP(UK); Nandita Pal, BSc; Steff C. Lewis, PhD; Jo J. Bhattacharya, FRCR; Robin J. Sellar, FRCR; Charles P. Warlow, MD, FRCP; and the AVM Observer Agreement Study Group

Purpose—We aimed to determine intraobserver and interobserver agreement in the characterization of brain arteriovenous malformation (AVM) angioarchitecture on intra-arterial digital subtraction angiograms.

Methods—Five experienced interventional neuroradiologists independently reviewed 40 anonymized angiograms obtained at the time of first-ever AVM diagnosis. The allocation of the films to observers was balanced for AVM size and complexity. Every observer was compared with himself and all the others by distributing the films in 2 batches 3 months apart. The observers used standard forms to collect both quantitative and categorized qualitative angiographic data. To measure agreement we used the kappa statistic ($\kappa$) for nominal data, weighted $\kappa$ for ordinal and discrete interval data, and Bland & Altman analysis for continuous data.

Results—Intraobserver agreement was generally moderate to substantial, with 95% confidence intervals ranging from fair to almost perfect. However, for every characteristic, interobserver agreement was less than intraobserver agreement. Interobserver agreement was generally slight to moderate, with 95% confidence intervals ranging from less than chance to almost perfect.

Conclusion—This study demonstrates the need for robust and generalizable definitions of AVM angioarchitecture and methods of nidus size measurement—with proof of good intraobserver and interobserver agreement—for future efforts to understand the prognosis and best treatment of AVMs. (Stroke. 2002;33:1501-1509.)

Key Words: cerebral angiography ■ cerebral arteriovenous malformations ■ observer variation

The vascular anatomy (angioarchitecture) of brain arteriovenous malformations (AVMs) demonstrated on intra-arterial digital subtraction angiograms is thought to influence their prognosis\(^1\) and is therefore often used to decide whether and how to treat an AVM.\(^2,3\) However, accurate description of angioarchitecture is inevitably affected by the inherent complexity of AVMs, neuroradiologists’ personal interpretations, and human error. These sources of observer variation are likely to have been compounded by inconsistencies in angioarchitecture terminology used in the literature, varying perceptions of what is abnormal, and—until very recently—the lack of structured definitions intended for widespread use.\(^4\)

A recent systematic review revealed no published studies on MEDLINE and Embase of observer agreement in the angiographic assessment of AVMs.\(^1\) We therefore sought to evaluate observer agreement using an approach that reflected day-to-day practice, minimized bias, and would not impose too great a burden on observers despite being powered adequately.

Materials and Methods

Angiograms

To be representative of everyday practice, we used 40 angiograms obtained at the time of first-ever AVM diagnosis from 1 year of a prospective population-based cohort of adults with a first-in-a-lifetime diagnosis of AVM (Scottish Intracranial Vascular Malformation Study [SIVMS], www.dcn.ed.ac.uk/ivm, accessed February 10, 2002). We used identical copies of the entire run of the 4-vessel angiogram that led to the first-ever AVM diagnosis. The angiograms were performed at the 4 Neuroscience Centers in Scotland; they included all the anteroposterior, lateral, and oblique views and vascular territories necessary to visualize the AVM and only omitted normal vascular territories and frames from superselective vascular catheterization performed before embolization. Because of the different facilities for angiography, the spectrum of experience of the radiologists, and the variable resolution of copy films, the 40 angiograms inevitably varied in quality. To explore the influence of film quality, the nonparticipating study neuroradiologists (J.J.B. and R.J.S.) rated the quality of the films before starting the study (12...
TABLE 1. Characteristics of the 5 Observers (All Consultant Interventional Neuroradiologists)*

<table>
<thead>
<tr>
<th>Observer</th>
<th>Length of Practice as a Consultant Specialist (y)</th>
<th>Special Interest in AVMs?</th>
<th>AVM Intervventional Procedures Performed (per y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>10</td>
<td>✓</td>
<td>15–30</td>
</tr>
<tr>
<td>B</td>
<td>12</td>
<td>✓</td>
<td>40</td>
</tr>
<tr>
<td>C</td>
<td>8</td>
<td>✓</td>
<td>6</td>
</tr>
<tr>
<td>D</td>
<td>7</td>
<td>✓</td>
<td>15</td>
</tr>
<tr>
<td>E</td>
<td>21</td>
<td>✓</td>
<td>30</td>
</tr>
</tbody>
</table>

AVM indicates arteriovenous malformation.

*See Appendix for the list of participants.

were “excellent,” 18 were “good,” 8 were “average,” and 2 were “poor” but none were “terrible”). Axial CT or MR imaging from the presentation that led to diagnosis was included to aid localization. Size markers on CT and MR imaging were obscured to force observers to estimate nidus dimensions from the angiograms.

Observers

The 5 observers were practicing consultant interventional neuroradiologists in the UK (Table 1). They worked in separate cities in the UK and had never been involved in the management of any of the patients in this study. They interpreted the angiograms without knowledge of the original findings or each other’s results. The observers were not presented with definitions of any of the angiographic architectural features under investigation, and—as often occurs in everyday practice—they used the diameter of the genu of the petrosal portion of the internal carotid artery (5 mm) as a reference for sizing a nidus, either using calipers or a customized scale on paper.5

Angiogram Distribution

The 40 angiograms were distributed and reviewed in 2 batches between January and May 2001. Every angiogram was reviewed by 2 of the 5 observers for the interobserver study, and 38 of the 40 angiograms were reviewed by the same observer on 2 separate occasions for the intraobserver agreement study. The angiograms were divided among the observers so that each neuroradiologist was allocated a similar spectrum of AVMs according to crude indicators of their nidus diameter (2 large [>6cm], 14 medium, 24 small [<3cm]) and vascular complexity (18 simple [≤2 feeders or draining veins] and 22 complex [>3 feeders or draining veins]).

Data Collection

A standard data collection form was distributed with each angiogram. The following data were collected (forced categories are described in parentheses): depth (deep, superficial); nidus diameter (mm) in each of 3 dimensions (anteroposterior, transverse, and vertical); number of feeding arteries; feeding artery angiopathy (yes, no), and if present whether abnormally dilated, stenosed, or both dilated and stenosed; angiogenesis (yes, no); collateral supply (yes, no), and if present whether dural, pial/leptomeningeal, or both dural and leptomeningeal; nidus border (compact/diffuse); discernible fistula in the nidus (yes, no); number of draining veins/nidus compartments; Spetzler-Martin surgical grade, calculated by an observer summing scores from the scale provided for AVM nidus size, pattern of venous drainage, and eloquence of adjacent brain (although constituent scores on each of these 3 items were not collected)11; venous varices (yes, no)12; venous ectasia (yes, no)13; venous stenosis (yes, no)14; aneurysm(s) (yes, no), and aneurysm type if identified (feeding artery, nidal, pseudo-aneurysm, remote).15

Data Processing

All data were double-punched to ensure accuracy of data entry. Because the same observer reviewed some scans on 2 separate occasions for the intraobserver agreement study, only data from the earliest date that an observer reviewed a particular scan were used to evaluate interobserver agreement.

Statistical Analysis

The primary outcomes in this study were observer variation quantified by the kappa statistic (κ) for nominal data (eg, dichotomous yes/no answers),16 the weighted κ statistic for ranked ordinal data (eg, Spetzler-Martin grade) and discrete interval data (eg, number of feeders),17 and Bland & Altman analysis for continuous data (eg, nidus dimensions in millimeters).18,19 Percentage agreement between observers is not a good measure, because—unlike κ (Figure 1)—it does not discriminate between actual agreement and agreement that arises as a result of chance.

The raw data were recoded for some variables to enable a simple assessment of observer agreement using the κ statistic, although we also examined cross tabulations of the raw data and Bland & Altman plots for these variables where appropriate (eg, nidus dimensions, see below). For example, nidus size in each of the 3 dimensions was recoded into a dichotomous variable according to whether an observer found the size to be ≥30 mm or <30 mm (the size threshold for stereotactic radiotherapy). The numbers of feeding arteries and draining veins were recoded as 1, 2, or ≥3 because different radiologists had different thresholds for deeming there to be “multiple” vessels when they were too numerous to count, which made Bland & Altman analysis of these variables impossible.

All analyses were performed in Statistical Product for the Social Sciences (SPSS) version 10.0.5 except confidence intervals for κ, which were calculated using Confidence Interval Analysis software,20 and weighted κ tests and their confidence intervals, which were calculated in Statistical Analysis Software (SAS) version 8.

Sample Size Calculation

The study was designed with 87% power to detect a greater than fair agreement (κ=0.4) at the P=0.05 level of significance, assuming the level of agreement for the characteristics would be substantial (κ=0.7).21

Ethics

Every film was carefully anonymized to protect patient confidentiality and comply with UK Data Protection legislation. This study is covered under the approval given for SIVMS by the Multicentre Research Ethics Committee for Scotland (MREC/98/0/48).

Results

Complete responses were received from all 5 observers, and the median time between an observer reporting the 2 batches of angiograms was 5 (range, 4 to 6) months. The interpretation of intraobserver agreement requires a prior appreciation of intraobserver agreement because between-observer variation is inevitably affected by the extent of within-observer variation.

Figure 2 demonstrates that for every characteristic, intraobserver agreement was greater than interobserver agreement. Intraobserver agreement was moderate to substantial, with 95% confidence intervals ranging from fair to almost perfect. Interobserver agreement was slight to moderate, with 95%
Comparisons of raw continuous data explains why about the mean for interobserver as opposed to intraobserver nidus dimension as an example (Figure 3), the greater scatter size increases, especially above 20 mm. Using the transverse intraobserver and interobserver variation to increase as nidus weighted (asterisked) /H9260

server (filled boxes) agreement, measured by the unweighted or /H11350

AVM nidus was perfect. Confidence intervals ranging from less than chance to almost perfect; 0.6 to 0.8, substantial; 0.4 to 0.6, moderate; 0.2 to 0.4, fair; 0 to 0.2, slight; <0, agree less than chance).

The extent of intraobserver (open boxes) and interobserver agreements measured by kappa estimates with 95% confidence intervals. Qualitative ranges for the extent of agreement, measured by κ are marked with dotted lines (0.8 to 1, almost perfect; 0.6 to 0.8, substantial; 0.4 to 0.6, moderate; 0.2 to 0.4, fair; 0 to 0.2, slight; <0, agree less than chance).

Our study is a pragmatic effort to understand both intraobserver and interobserver agreement in neuroradiologists’ interpretations of AVM angioarchitecture in day-to-day decision making. The study sample reflected everyday practice – the study sample. The results should be considered in the light of important statistical caveats common to all studies using κ, and they have some implications for routine practice. The study sample reflected everyday practice – the study sample. The results should be considered in the light of important statistical caveats common to all studies using κ, and they have some implications for routine practice.

Figure 2. The extent of intraobserver (open boxes) and interobserver (filled boxes) agreement, measured by the unweighted or weighted (asterisked) κ, shown as point estimates with 95% confidence intervals. Qualitative ranges for the extent of agreement measured by κ are marked with dotted lines (0.8 to 1, almost perfect; 0.6 to 0.8, substantial; 0.4 to 0.6, moderate; 0.2 to 0.4, fair; 0 to 0.2, slight; <0, agree less than chance).

Intraobserver agreement about whether the diameter of an AVM nidus was ≥30 mm or <30 mm ranged from substantial to almost perfect, whereas interobserver agreement was somewhat worse. Plotting the raw, continuous nidus size data on Bland & Altman plots reveals a tendency for both intraobserver and interobserver variation to increase as nidus size increases, especially above 20 mm. Using the transverse nidus dimension as an example (Figure 3), the greater scatter about the mean for interobserver as opposed to intraobserver comparisons of raw continuous data explains why κ=0.36 for interobserver comparisons as opposed to κ=0.78 for intraobserver comparisons using categorical data with a 30-mm size threshold (Figure 2).

Interobserver agreement was greatest for characteristics such as determining whether nidus diameter was ≥30 mm or <30 mm (eg, vertical dimension κ=0.62 [95% CI, 0.37 to 0.88]) and whether there were venous varices or not (κ=0.56 [95% CI, 0.31 to 0.81]). In general there was the greatest overall agreement between observers for AVMs with simple angioarchitecture lacking many of the features of interest (Figure 4). Interobserver agreement was worst for characteristics such as venous stenosis (κ=0.14 [95% CI, −0.33 to 0.60]), angio genesis (κ=0.18 [95% CI, −0.13 to 0.49]), and the type of nidus border (κ=0.22 [95% CI, −0.20 to 0.64])—in other words, AVMs with more complex angioarchitecture (Figure 5). Alarmingly, interobserver agreement was only moderate for the variables with the greatest importance in routine practice: Spetzler-Martin grade, which influences predictions of morbidity from surgery (weighted κ=0.47 [95% CI, 0.30 to 0.64]), and the presence of aneurysms, which are thought to confer a greater risk of subsequent hemorrhage and may also influence management decisions (κ=0.40 [95% CI, 0.11 to 0.68]).

We asked observers to assess film quality with the intention of exploring whether it influenced observer agreement. However, intraobserver agreement about film quality was κ=0.43 (95% CI, 0.20 to 0.67), and interobserver agreement was κ=0.19 (95% CI, −0.04 to 0.42), making stratification of kappas by quality for each angioarchitectural feature susceptible to observer variation in the determination of quality. Moreover, the small number of observations in each quality category resulted in even wider confidence intervals around the kappa estimates, and we found no consistent trend toward a better level of agreement for higher quality films.

**Discussion**

This is the first study of its kind. However, there has been an abstract reporting agreement about AVM size and morphology using World Wide Web–based joint photographic expert group (JPEG) format magnetic resonance and angiographic images.22 This study’s response rate was 63%, only 2 of 19 participants were neuroradiologists, and the images used were in a different format from ours.

Our study is a pragmatic effort to understand both intraobserver and interobserver agreement in neuroradiologists’ interpretations of AVM angioarchitecture in day-to-day decision making. The study sample reflected everyday practice by being drawn from a population-based cohort. The angiograms were of adequate quality, although it will always be difficult to reflect the dynamic nature of angiography in hard copy format. Data collection was complete. Our main findings are that there was greater intraobserver than interobserver agreement, and agreement ranged from less than chance to almost perfect. These findings should be regarded as a baseline measure of observer agreement for future studies; the results should be considered in the light of important statistical caveats common to all studies using κ, and they have some implications for routine practice.

**Statistical Considerations**

**Bias, Confounding, and Chance**

We avoided bias by using angiograms that the study neuroradiologists had never seen before, by anonymizing films, and by leaving a median of 5 months between an observer re-reviewing the same angiogram (to lessen recognition effects). By distributing angiograms evenly according to image quality and AVM complexity, we sought to minimize confounding. Chance effects were minimized by ensuring the study was adequately powered to detect a difference from
only fair agreement ($\kappa=0.4$), assuming agreement would be substantial ($\kappa=0.7$) for any characteristic. Indeed we confirmed our suspicion that intraobserver agreement was of this order for most angioarchitectural features (Figure 2).

**Precision**
To establish the extent of observer agreement with greater precision, a larger study will be required. This could be achieved by increasing the number of observers and/or angiograms, which would also enable an analysis of the bias of any individual observer (see below). A greater number of angiograms for which there are 2 observer comparisons would certainly narrow the 95% confidence intervals around estimates of $\kappa$ (Figure 2), but having more than 2 observers per angiogram would complicate the statistical analysis.

**Kappa Caveats**
Although there are several well-rehearsed caveats to the use of $\kappa$, experts debate whether the intraclass correlation coefficient is a better measure, $\kappa$ is nevertheless the most frequently used index of agreement. $\kappa$ relies on both the subjects under study and the observers being independent and that the categories in the scale are independent, mutually exclusive, and exhaustive. These assumptions held for this study, although varying beliefs about angioarchitecture between research groups (a form of global observer variation) will affect whether readers perceive overlap between categories for some characteristics.

The greater the number of scale categories, the lower $\kappa$ will inevitably be, so agreement will tend to appear better with a dichotomous scale. For example, $\kappa$ for interobserver agreement falls from 0.29 to 0.19 when subdividing angiopathy into more than yes/no categories, and it falls from 0.41 to 0.30 when subdividing collateral supply into more than yes/no categories. Conversely, using only 3 categories for the numbers of feeding and draining vessels masked variation; we would have used Bland & Altman analysis were it not for different thresholds between observers for declaring vessels “multiple.”

An artifact of $\kappa$ is that it is affected in complex ways by both the prevalence of abnormality among the subjects used and also by observer bias. Firstly, most characteristics in this study were unevenly distributed (the most extreme being venous stenosis and nidus border), although the distribution of some features such as aneurysms was more even (Table 2). The nature of our sample means this probably reflects the population distribution of these abnormalities. But when marginal totals are unbalanced, or expected levels of agreement are high because of a high underlying prevalence, $\kappa$ is...
When there are several ordinal scale categories (eg, Spetzler-Martin grade) between which large disagreements are more serious but would be treated as equally serious by $\kappa$, the weighted $\kappa$ can be used when the relative seriousness of disagreements is specified.\(^\text{17}\) In this study we allocated weights evenly (eg, weights of 1, 0.75, 0.5, 0.25, and 0 for Spetzler-Martin grades 1 to 5), but whether and how uneven weighting should be used to reflect clinically important thresholds (such as $\leq 2$ and $\geq 3$ on the Spetzler-Martin scale) is debatable.

Fragile and examination of the influences of prevalence effects will be essential when comparing studies.\(^\text{25}\) Secondly, agreement is only one aspect of variation between observers, the other being biases between them (eg, a tendency for one observer to systematically overestimate nidus size).\(^\text{23}\) We avoided bias by ensuring each neuroradiologist was observer 1 or observer 2 a comparable number of times in the interobserver study, and each neuroradiologist was equally represented in the intraobserver study.

---

**Figure 4.** Anteroposterior (A) and lateral (B) projections of a right carotid intra-arterial digital subtraction angiogram that showed a right frontoparietal arteriovenous malformation (AVM). There was agreement that this AVM was a simple fistula of Spetzler-Martin grade II, lacking many angioarchitectural features. However, there was disagreement about whether there were 1 or 2 feeders (dashed arrow) and whether there was a nidal aneurysm or venous varix (solid arrow).

**Figure 5.** Anteroposterior (A) and lateral (B) projections of a right carotid intra-arterial digital subtraction angiogram that showed a right frontal arteriovenous malformation (AVM) (arrow). There was disagreement about whether the nidus was greater or less than 30 mm in all dimensions, number of feeders, type of feeding artery angiopathy, presence of angiogenesis, existence of collaterals, nidus border, discernible fistula in the nidus, number of draining veins, Spetzler-Martin grade, venous ectasia, and existence of nidal aneurysms.
Implications for Routine Practice

The high level of intraobserver agreement shows that experienced interventional neuroradiologists are consistent, but the poor interobserver agreement shows that assessment and interpretation differ between them. This argues for caution in interpreting prognosis and basing treatment decisions on angioarchitectural features with less than adequate interobserver agreement.

Evaluate Nidus Size With Standardized Calibration Markers

There are several barriers to accurately sizing an AVM on an angiogram, which are reflected by the scatter of size estimates in Figure 3. The definition of a nidus as the area toward which multiple feeding arteries converge and from which enlarged veins drain is somewhat arbitrary. This is especially problematic when nidus morphology is diffuse and when the AVM is a simple fistula (Figure 4, which elicited nidus dimensions of 0 to 20 mm in this study). The nidus is often not imaged in its entirety when catheterizing single vascular territories during angiography, making the maximum linear diameter in any dimension difficult to gauge. Without reference markers, the widespread use of the diameter of the genu of the petrous portion of the internal carotid artery—as in this study—can be inaccurate; for example, it is larger than 5 mm if feeding ipsilateral, large, high-flow brain AVMs. Potential solutions to these sources of measurement error include consistent angiographic magnification factors, widespread use of simple calibration markers such as rulers, washers, and coins (which were not in frequent use a decade ago and do not seem to be now), and magnification/minification rulers. The routine use of standardized calibration markers on angiograms is therefore essential to reduce variation in which patients are considered—and then further evaluated with stereotactic sizing—for stereotactic radiotherapy.

Perform Angiography That Is Adequate to Characterize Subtle Angioarchitecture

The reliable characterization of angioarchitectural features of interest usually requires imaging of suspected abnormalities in at least 2 planes (eg, identification of aneurysms and venous stenosis). This should be the required standard of routine clinical practice. Superselective angiography further assists expert neuroradiologists in their interpretation of arterial supply (eg, the number of feeders) and the nidus (eg, whether there is a single fistula within it). But the balance between the risks and benefits of the routine use of superselective angiography has not been established, and there are no data on whether it carries greater risks than routine catheter angiography.

Future Research

The results of this study argue that the further development of acceptable radiological definitions and future studies of observer agreement should be prioritized in certain areas.

TABLE 2. Supplementary Data: % Agreement About the Different Variables in the Intraobserver and Interobserver Agreement Studies

<table>
<thead>
<tr>
<th>Variable</th>
<th>% Prevalence of Agreement (Intraobserver Study)</th>
<th>% Prevalence of Agreement (Interobserver Study)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVM depth</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deep</td>
<td>24</td>
<td>23</td>
</tr>
<tr>
<td>Superficial</td>
<td>58</td>
<td>53</td>
</tr>
<tr>
<td>Anteroposterior diameter</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥3 cm</td>
<td>34</td>
<td>23</td>
</tr>
<tr>
<td>&lt;3 cm</td>
<td>55</td>
<td>58</td>
</tr>
<tr>
<td>Transverse diameter</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥3 cm</td>
<td>34</td>
<td>23</td>
</tr>
<tr>
<td>&lt;3 cm</td>
<td>55</td>
<td>48</td>
</tr>
<tr>
<td>Vertical diameter</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥3 cm</td>
<td>37</td>
<td>28</td>
</tr>
<tr>
<td>&lt;3 cm</td>
<td>53</td>
<td>55</td>
</tr>
<tr>
<td>Number of feeders</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>16</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
<td>10</td>
</tr>
<tr>
<td>≥3</td>
<td>45</td>
<td>45</td>
</tr>
<tr>
<td>Angiopathy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>50</td>
<td>40</td>
</tr>
<tr>
<td>No</td>
<td>26</td>
<td>15</td>
</tr>
<tr>
<td>Angiogenesis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>32</td>
<td>18</td>
</tr>
<tr>
<td>No</td>
<td>58</td>
<td>43</td>
</tr>
<tr>
<td>Collateral supply</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>29</td>
<td>23</td>
</tr>
<tr>
<td>No</td>
<td>58</td>
<td>50</td>
</tr>
<tr>
<td>Nidus border</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Compact</td>
<td>68</td>
<td>68</td>
</tr>
<tr>
<td>Diffuse</td>
<td>18</td>
<td>8</td>
</tr>
<tr>
<td>Discernible fistula</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>29</td>
<td>20</td>
</tr>
<tr>
<td>No</td>
<td>58</td>
<td>48</td>
</tr>
<tr>
<td>Number of veins</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>32</td>
<td>33</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td>≥3</td>
<td>34</td>
<td>28</td>
</tr>
<tr>
<td>Spetzler-Martin grade</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>13</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>18</td>
</tr>
<tr>
<td>3</td>
<td>16</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Venous varices</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>37</td>
<td>30</td>
</tr>
<tr>
<td>No</td>
<td>55</td>
<td>48</td>
</tr>
<tr>
<td>Venous ectasia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>55</td>
<td>55</td>
</tr>
<tr>
<td>No</td>
<td>24</td>
<td>18</td>
</tr>
<tr>
<td>Venous stenosis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>13</td>
<td>5</td>
</tr>
<tr>
<td>No</td>
<td>74</td>
<td>70</td>
</tr>
<tr>
<td>Aneurysms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>29</td>
<td>30</td>
</tr>
<tr>
<td>No</td>
<td>58</td>
<td>40</td>
</tr>
</tbody>
</table>

% Disagreement = 100 - ∑(prevalence of agreement in each category).
Agreement about nidus size should be re-evaluated when standardized calibration markers are in widespread use because of its importance in determining who is eligible for stereotactic radiotherapy.

The Spetzler-Martin grading system is in widespread use to predict morbidity from surgical excision, but the observed levels of intraobserver (weighted $\kappa=0.63$ [95% CI, 0.48 to 0.79]) and interobserver (weighted $\kappa=0.47$ [95% CI, 0.30 to 0.64]) agreement are cause for some concern. In addition to nidus size, the other 2 components of the Spetzler-Martin grading system (eloquence of adjacent brain and pattern of venous drainage) should be explored. It will be interesting to discover which of the 3 components is the most responsible for less than perfect observer agreement in the use of this scale. In this study, observers assumed left hemisphere dominance for the determination of eloquence. To assist with determination of eloquence, future studies might benefit from a brief clinical history accompanying each angiogram. However, this benefit might be offset by indirectly encouraging overinterpretation of some angioarchitectural features in cases with a more severe presentation, thereby introducing bias.

Since a worse prognosis for the first occurrence of hemorrhage seems to be conferred by the identification of aneurysms in conjunction with unruptured AVMs (especially those in the nidus, not thought to be pseudoaneurysms), efforts should be made to understand why interobserver agreement about the very presence of aneurysms was only $\kappa=0.40$ (95% CI, 0.11 to 0.68). This is likely to be only partly explained by neuroradiologists in this study reviewing hard copies of the angiograms, rather than performing them, and superselective studies not being available. It will be important to further evaluate agreement about the presence of aneurysms (distinct from infundibula), their number, their locations, and whether they should be treated.

In this study individual observers’ thresholds for declaring feeding or draining vessels multiple seemed to differ. Clearly, the ability to correctly define feeding vessel anatomy is to some extent dependent on the use of superselective angiography—neuroradiologists do occasionally discover unsuspected feeders only at the time of embolization. Therefore, the only way for us to interpret agreement about feeding and draining vessel anatomy was to group the raw, continuous data about numbers of vessels into 3 simple ordinal categories (1, 2, or $\geq$3 vessels), although some authors have reservations about this statistical approach. Although there are no clear data about the absolute number of feeding or draining vessels carrying particular prognostic importance, future studies might profit from assessing the agreement about the original vascular territories of feeding vessels, because this may determine outcome.

Particular emphasis should be placed on developing internationally agreed-upon definitions for the characteristics above and for those features that had the greatest interobserver variation in this study (angiopathy, angio genesis, collateral supply, nidus border, discernible fistula, and venous stenosis). Thereafter, using our study as a baseline measure of observer agreement before to publication of the Joint Writing Group’s definitions, it will be important to reassess observer agreement in larger studies, among neuroradiologists from different countries, and using emerging techniques (such as MR and digital angiography, image manipulation, and reconstruction). Furthermore, research workers have a similar need for clear definitions of clinical events affecting people with AVMs. These should focus on the nature of the clinical event (eg, transient, persistent and progressive focal neurological deficits in the absence of a seizure or hemorrhage), its timing in relation to the inception point of first presentation (eg, first occurrence, first recurrence, second recurrence), and whether it is attributable to an AVM or not (degree of certainty).

In conclusion, it is important to further develop internationally agreed-upon, generalizeable definitions of AVM clinical features, angioarchitecture, and methods of nidus size measurement and prove they have good intraobserver and interobserver agreement. These are necessary for future efforts to understand the prognosis and best treatment of AVMs.

Appendix

The members of the AVM Observer Agreement Study Group were as follows: Andy Clifton, Department of Neuroradiology, Atkinson Morley’s Hospital, London; Anil Gholkar, Neuro X ray, Newcastle General Hospital, Newcastle-upon-Tyne; Shawn Halpin, Department of Radiology, University of Wales College of Medicine, Cardiff; John Millar, Department of Radiology, Wessex Neurological Centre, Southampton General Hospital, Southampton; and Andy Molyneux, Department of Radiology, Radcliffe Infirmary, Oxford.

Acknowledgments

Rustam Al-Shahi was funded by a United Kingdom Medical Research Council clinical training fellowship. This study was supported by the Chief Scientist Office of the Scottish Executive Health Department (CZB/4/35), the Stroke Association (TSA/04/01), and a Small Project Grant from the University of Edinburgh. We are very grateful to our collaborators and those participating in the Scottish Intracranial Vascular Malformation Study (SIVMS) who made this study possible.

References

Management of brain arteriovenous malformations (AVM) continues to evolve as evidence-based strategies strengthen clinical practice. With careful analyses of observational data derived from tertiary referral centers, information is emerging to create practice. With careful analyses of observational data derived from tertiary referral centers, information is emerging to create reliably studies.

Stroke June 2002


17. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull. 1968;70:213–220.


distribution of cases was balanced for brain AVM size and complexity. Standardized forms were used in data collection, although standardized definitions were not provided. The authors used sophisticated measures to assess agreement in the analyses.

Interobserver agreement ranged from 14% to 62% and was generally worse for characteristics associated with more complex angioarchitecture of the brain AVM. Of notable concern was low interobserver agreement involving the measurement of Spetzler-Martin grade (47%) and the presence of aneurysms (40%), 2 variables that arguably bear the greatest importance in patient management decisions. However, it would be premature at this time to consider these estimates as evidence of the low reliability inherent in the measurement of these variables. Some modifications in study methodology may allow for a better measure of data reliability.

In future studies, the use of real-time high-quality angiographic digital images or animation will eventually supplant printed static images for communicating angiographic information. A reliability estimate obtained using high-quality angiographic media may be of apparently limited generalizability because it may not reflect the quality of angiograms seen in routine practice. However, this approach is nonetheless necessary to determine the extent to which reliability can be improved by increased angiographic quality and whether this improvement is significant enough to be of clinical importance. In addition, the use of standardized definitions and scales and stratification by angioarchitectural complexity would improve reliability.

Al-Shahi and colleagues should be commended for their efforts. The article underscores the importance of measurement issues in brain AVM studies. Once it is determined that brain AVM variables can be measured with a high degree of reliability, associations with clinically significant variables should be more obvious, and prediction models and guidelines should improve. This information will ultimately help to construct ethical and efficient clinical trials to compare treatment modalities.

Alexander X. Halim, PhD, Guest Editor
William L. Young, MD, Guest Editor
Department of Anesthesia & Perioperative Care
S. Claiborne Johnston, MD, PhD, Guest Editor
Department of Neurology
University of California San Francisco
School of Medicine
San Francisco, California

References
Observer Agreement in the Angiographic Assessment of Arteriovenous Malformations of the Brain

Rustam Al-Shahi, Nandita Pal, Steff C. Lewis, Jo J. Bhattacharya, Robin J. Sellar, Charles P. Warlow and the AVM Observer Agreement Study Group

Stroke. 2002;33:1501-1508
doi: 10.1161/01.STR.000018318.83802.18

Stroke is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2002 American Heart Association, Inc. All rights reserved.
Print ISSN: 0039-2499. Online ISSN: 1524-4628

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/33/6/1501

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Stroke can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Stroke is online at:
http://stroke.ahajournals.org//subscriptions/