Inter- and Intraobserver Agreement in Scoring Angiographic Results of Intra-Arterial Stroke Therapy


ABSTRACT

BACKGROUND AND PURPOSE: Angiographic results are commonly used as surrogate markers of the success of intra-arterial therapies for acute stroke. Inter- and intraobserver agreement in judging angiographic results remain poorly characterized. Our goal was to assess 2 commonly used revascularization scales.

MATERIALS AND METHODS: A portfolio of 148 pre- and post treatment images of 37 cases of proximal anterior circulation occlusions was electronically sent to 12 expert observers who were asked to grade treatment outcomes according to recanalization (of arterial occlusive lesion) or reperfusion (TICI) scales. Three expert observers had to score treatment outcomes by using a similar portfolio of 32 patients or when they had full access to all angiographic data, twice for each method 3–12 months apart. Results were analyzed by using \( \kappa \) statistics.

RESULTS: Agreement among 9 responding observers was moderate for both the TICI (\( \kappa = 0.45 \pm 0.01 \)) and arterial occlusive lesion (\( \kappa = 0.39 \pm 0.16 \)) scales. Agreement was similar (moderate) when 3 observers had access to a portfolio (\( \kappa = 0.59 \pm 0.06 \) and 0.49 \( \pm 0.07 \), respectively) or to the full angiographic data (\( \kappa = 0.54 \pm 0.06 \) and 0.59 \( \pm 0.07 \), respectively). Intraobserver agreement became “substantial” (>0.6) when outcomes were dichotomized into “success” (TICI 2b, 3; arterial occlusive lesion II, III or “failure”; the results were judged more favorably when the arterial occlusive lesion rather than the TICI scale was used.

CONCLUSIONS: There is an important variability in the assessment of angiographic outcomes of endovascular treatments, invalidating comparisons among publications. A simple dichotomous judgment can be used as a surrogate outcome when treatments are assessed by the same observers in randomized trials.

ABBREVIATIONS: AOL — arterial occlusive lesion; IMS — Interventional Management of Stroke; TIMI — Thrombolysis in Myocardial Infarction

Copyright 2014 by American Society of Neuroradiology.
has not been rigorously assessed. The aim of the present work was to assess the precision and reproducibility of 2 angiographic outcome scales of intra-arterial therapies, one for recanalization and one for reperfusion: The primary arterial occlusive lesion recanalization scoring method, initially proposed for the Interventional Management of Stroke (IMS) I analyses,17 and the Thrombolysis in Cerebral Infarction perfusion categories, proposed by the Technology Assessment Committees of the American Society of Interventional and Therapeutic Neuroradiology and the Society of Interventional Radiology.7 These scales (with or without some modifications) are being used in recent trials on intra-arterial stroke therapy, such as IMS II and III,18 the Mechanical Retrieval and Recanalization of Stroke Clots Using Embolectomy (MR Rescue) trial,19 and the Endovascular Treatment for Small Core and Anterior Circulation Proximal Occlusion with Emphasis on Minimizing CT to Recanalization Time (ESCAPE) trial (M. Hill, personal communication; May 2013) and others.20

MATERIALS AND METHODS
The primary aim of this work was to evaluate the intra- and interobserver variability in adjudicating outcomes of treatment according to 2 ordinal scales commonly used to assess angiographic results of intra-arterial thrombectomy. The evaluation comprised 3 parts, 2 by electronic surveys; the third evaluation was designed to resemble clinical work and to validate the results obtained by the electronic surveys: 1) an electronic survey to assess interobserver agreement among 9 different expert “external” readers regarding the angiographic outcomes of 37 cases; 2) a similar electronic survey, modified and reduced to match the 32 patients to appear in explanatory boxes beside the answering boxes for each case. All anonymized images were retrieved by one author (L.E.) from the PACS of one institution. The portfolio comprised paired (pos-tero-anterior and lateral projections) selected pre- and post treatment late arterial phase angiograms of 37 cases. Cases included 32 consecutive patients who had been treated endovascularly for acute anterior circulation strokes in a single institution during 9 months (January to September 2011). For part 1, 5 additional cases were constructed by using intermediate-phase results of complex interventions in 5 patients already included, in an attempt to better balance the proportions of the marginal sums of the contingency tables and hopefully minimize paradoxes of $\kappa$ statistics.21 On each page of the electronic version sent to reviewers, 2 pretreatment and 2 post treatment images were displayed side by side. No clinical information was provided. There was no training of observers. The part 1 electronic portfolio was sent to 12 expert interventional neuroradiologists, selected because they had designed studies or trials on transarterial stroke therapy. Nine, with 5–27 years of clinical experience, answered, working in 6 different centers; 4 were from Canada; 3, from the United States; and 2, from France. One observer answered the questionnaire twice 3 months apart. Observers were given the task of grading each pair of images according to the 4-value AOL scale6 and the 5-value TICI scale.7 The explicit definitions of the 2 scales appeared in explanatory boxes beside the answering boxes for each case.

Part 2: Electronic Survey with 3 Readers
The electronic questionnaire, modified to include only the 32 real patients (excluding the 5 “constructed cases” added to part 1), was administered twice, 12 months apart, to the 3 internal interventional neuroradiologists involved in the treatment of acute stroke who participated in part 3 of the study.

Part 3: Intra- and Interobserver Agreement Using All Angiographic Images
To verify that the artificial conditions imposed by the electronic survey did not affect results and to better assess intraobserver agreement, the same 3 observers were asked to grade the angiographic outcomes of the same 32 patients, by using the full set of angiographic data presented by 3 authors not participating in the evaluation of cases (L.E., C.R., M.G.) directly on the PACS, independently twice, 3–12 months apart.

Scores and Dichotomies
To assess intracranial reperfusion, readers were asked to use the TICI score as described by Higashida et al7: grade 0, no perfusion, no antegrade flow beyond the point of occlusion; grade 1, pene-tration with minimal perfusion; grade 2, partial perfusion; grade 2a, only partial filling (less than two-thirds) of the entire vascular territory visualized; grade 2b, complete filling of all of the expected vascular territory visualized but filling more slowly than normal; and grade 3, complete perfusion.

To assess arterial recanalization, readers were asked to use the AOL score22: score 0, no recanalization of the primary occlusive lesion; score I, incomplete or partial recanalization of the primary occlusive lesion with no distal flow; score II, incomplete or partial recanalization of the primary occlusive lesion with any distal flow; and score III, complete recanalization of the primary occlusion with any distal flow.

Because many reports providing results of treatments have used dichotomous “success-versus-failure” end points, we repeated $\kappa$ statistics, lumping categories into “success” defined as TICI 2b, 3 or AOL II, III, versus “failure,” defined as TICI 0, 1, 2a or AOL 0, I.4,23-25

Statistical Analyses
The multirater $\kappa$ statistics were computed by using the macro MAGREE with SAS, Version 9.3 (SAS Institute, Cary, North Carolina). This macro implements the methodology of Fleiss et al27 measuring the agreement when the number of raters is $>2$. This method also allowed identifying, for each scale, the categories in
which the most frequent disagreements occurred. \( \kappa \) values were interpreted according to Landis and Koch\( ^{27} \) with \( \kappa \) coefficients of 0 = poor; 0.01–0.20 = slight; 0.21–0.40 = fair; 0.41–0.60 = moderate; 0.61–0.80 = substantial; and 0.81–1.0 = almost-perfect agreement.

**RESULTS**

**Patients**

The portfolio included 32 consecutive patients (17 women; mean age, 63 ± 12). In addition to intra-arterial therapy, patients received IV-rtPA in 61% of the cases. The mean delay between symptoms and thrombectomy was 199 ± 47 minutes. The most frequent occlusions were located on the M1 segment of the middle cerebral artery (\( n = 19; 60% \)) or on the distal internal carotid artery (T-occlusion; \( n = 10; 32\% \)). The most frequent thrombectomy methods used during this period were an aspiration system (\( n = 13; 41\% \)) or a Stentriever (Trevo; Stryker, Kalamazoo, Michigan) system (\( n = 14; 43\% \)). Characteristics of patients are summarized in the On-line Table.

**Survey with 9 Observers**

There were large discrepancies in the adjudication of angiographic outcomes, with, for example, complete perfusion (TICI 3) being attributed to a wide range (8%–49%) of patients or complete recanalization (AOL III) in 22%–65% of patients, depending on observers.

Table 1 summarizes the \( \kappa \) values obtained when the 9 observers scored angiographic outcomes according to the TICI reperfusion categories (overall agreement, \( \kappa = 0.446 ± 0.013 \)). Table 2 summarizes the \( \kappa \) values when the categories were dichotomized as success (2b, 3) versus failure (0, 1, 2a) (overall agreement, \( \kappa = 0.616 ± 0.025 \)). \( \kappa \) coefficients of pairs of observers that reached “substantial agreement” (\( \kappa > 0.6 \)) increased from 9% to 60% with dichotomization. The TICI category that was the subject of most disagreements was 2b (\( \kappa = 0.242 ± 0.025 \)).

Table 3 summarizes the \( \kappa \) values when angiographic outcomes were categorized according to the AOL recanalization categories (overall agreement, \( \kappa = 0.394 ± 0.016 \)). Table 4 results were obtained when they were analyzed as success (II, III) or failure (0, I) (overall agreement, \( \kappa = 0.762 ± 0.025 \)). \( \kappa \) coefficients of pairs of observers that reached substantial agreement (\( \kappa > 0.6 \)) increased from 16% to 91% with dichotomization. The AOL category that was the subject of most disagreements was II (\( \kappa = 0.188 ± 0.025 \)).

The endovascular intervention was successful in 68%–87% of patients according to various observers when success was defined in terms of recanalization (AOL II or III) but in only 32%–62% of
patients when success was defined in terms of reperfusion (TICI 2b or 3).

**Intra-and Interobserver Agreement with Electronic or Full Datasets**

The results of parts 2 and 3 are summarized in Tables 5 and 6.

Intraobserver agreement between 2 sessions for experts having access to the full set of angiographic data on 32 patients was only slight to fair (0–0.4) in most cases, with only 1 observer reaching substantial agreement for the AOL scores. The $k$ values of the interobserver agreement obtained by comparing answers to the electronic questionnaire and the verdicts of the first and second reading sessions when observers had access to all images were similar and always below 0.6 (less than substantial).

When results were dichotomized, intraobserver agreement remained fair for 2 of 3 observers assessing reperfusion with the TICI scale but reached substantial agreement for all 3 readers when they assessed recanalization with the AOL scale; the interobserver agreement improved to substantial or better for 2 of 3 observers assessing reperfusion with the TICI scale but reached substantial agreement for all 3 readers assessing recanalization with the AOL scale.

The endovascular intervention was successful in 72%–79% of patients according to 3 observers having access to all images when success was defined in terms of recanalization (AOL II or III) but in only 35%–59% of patients when success was defined in terms of reperfusion (TICI 2b or 3).

**DISCUSSION**

The salient features of this work are the following: 1) Agreement in adjudicating angiographic results of endovascular interventions among multiple observers is, at best, fair to moderate; 2) this problem is not limited to divergent interpretations of the definitions of the various categories by various experts because intraobserver agreement was similarly poor when the same experts re-evaluated the same results twice on 2 independent occasions; 3) difficulties were not caused by the artificial and limited information provided by the survey we used to assess interobserver variations because similarly poor results were obtained when 3 observers were given access to all angiographic images; 4) $k$ coefficients reached more reassuring substantial agreement values when results were dichotomized into successes and failures; 5) the AOL recanalization scoring system seemed more repeatable than the TICI reperfusion scheme; and 6) the AOL II, III recanalization categories provided a more frequent verdict of success compared with the 2b, 3 TICI categories when dichotomization was used.

To evaluate our interventions, we had no choice but to reduce the variety and heterogeneity naturally found in clinical results to a (preferably small) number of categories and terms to name these categories that will determine what counts as a success or failure, in a common language that will allow tabulation of results and both valid comparisons between groups and communication of results among clinicians. When new categories (such as the TICI scale) are proposed, definitions can be provided as a sort of manual of translation, rules to translate the concrete results obtained in each particular case to a common language. As with any language, translation and communication by using our new terms may fail. If the meaning of such terms as TICI 2b or AOL II can be intentionally defined by explicit descriptions, whether these definitions and categories succeed in fixing the referents (ie, in re-identifying the same angiographic outcomes when they are seen by different observers or by the same observers at different time points) must be empirically tested to ensure that the new language does what it was designed to do, to allow valid comparisons and unambiguous communication. Both TICI and AOL scales had poor concordance when the same results were judged by different observers.

### Table 4: Interobserver agreement using the dichotomized AOL scale

<table>
<thead>
<tr>
<th>Ob1</th>
<th>Ob2</th>
<th>Ob3</th>
<th>Ob4</th>
<th>Ob5</th>
<th>Ob6</th>
<th>Ob6a</th>
<th>Ob7</th>
<th>Ob8</th>
<th>Ob9</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.646 ± 0.144</td>
<td>0.593 ± 0.154</td>
<td>0.613 ± 0.141</td>
<td>0.788 ± 0.117</td>
<td>0.64 ± 0.144</td>
<td>0.723 ± 0.122</td>
<td>0.726 ± 0.128</td>
<td>0.726 ± 0.128</td>
<td>0.726 ± 0.128</td>
<td></td>
</tr>
</tbody>
</table>

**Note:**—Inter-observer Kappa values ≥ 0.6 are highlighted in bold type.

### Table 5: Intra- and interobserver agreement using TICI (3 observers)

#### Intraobserver Agreement (All Images)

<table>
<thead>
<tr>
<th>Observer 1</th>
<th>Observer 2</th>
<th>Observer 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.488 ± 0.119</td>
<td>0.158 ± 0.122</td>
<td>0.095 ± 0.105</td>
</tr>
<tr>
<td><strong>0.716 ± 0.132</strong></td>
<td><strong>0.216 ± 0.179</strong></td>
<td><strong>0.246 ± 0.176</strong></td>
</tr>
</tbody>
</table>

#### Intraobserver Agreement using the dichotomized TICI scale

<table>
<thead>
<tr>
<th>Questionnaire</th>
<th>All Images Session 1</th>
<th>All Images Session 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.590 ± 0.060</td>
<td>0.364 ± 0.060</td>
<td>0.538 ± 0.060</td>
</tr>
<tr>
<td><strong>0.699 ± 0.100</strong></td>
<td><strong>0.489 ± 0.110</strong></td>
<td><strong>0.678 ± 0.110</strong></td>
</tr>
</tbody>
</table>

**Note:**—Inter-observer Kappa values ≥ 0.6 are highlighted in bold type.

### Table 6: Intra- and interobserver agreement using AOL (3 observers)

#### Intraobserver Agreement (All Images)

<table>
<thead>
<tr>
<th>Observer 1</th>
<th>Observer 2</th>
<th>Observer 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.611 ± 0.126</td>
<td>0.233 ± 0.149</td>
<td>0.302 ± 0.143</td>
</tr>
<tr>
<td><strong>0.731 ± 0.145</strong></td>
<td><strong>0.790 ± 0.142</strong></td>
<td><strong>0.626 ± 0.168</strong></td>
</tr>
</tbody>
</table>

#### Intraobserver Agreement using the dichotomized AOL scale

<table>
<thead>
<tr>
<th>Observer 2</th>
<th>Observer 1</th>
<th>Observer 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.492 ± 0.070</td>
<td>0.238 ± 0.070</td>
<td>0.590 ± 0.070</td>
</tr>
<tr>
<td><strong>0.787 ± 0.100</strong></td>
<td><strong>0.561 ± 0.110</strong></td>
<td><strong>0.870 ± 0.110</strong></td>
</tr>
</tbody>
</table>

**Note:**—Inter-observer Kappa values ≥ 0.6 are highlighted in bold type.

---

$k = 0.76197 ± 0.024507; P < 0.001$
obscures or by the same observers on 2 different occasions. This finding suggests that results of various case series or registries should not be compared when angiographic outcomes have not been adjudicated by the same observer.20,28

κ statistics provide a measure of agreement that takes into account the role of chance in the occurrence of concordant verdicts when estimating agreement between observers.29 Depending on prevalence, κ statistics are liable to paradoxes, such as high agreement but low κ values, when the distribution of the cases is imbalanced among categories.21 This problem may partly explain the low κ values of the intraobserver study on the 32-patient sample, in which tables were asymmetric. We believe that paradoxes do not explain the poor precision found in the 37-patient sample assessed by the 9 reviewers, more balanced by the introduction of 5 “intermediate cases” and when agreement was low for each category.

Difficult questions arise when one tackles the problem of agreement regarding a treatment outcome. Surely there must be some reality regarding revascularization, but in the absence of a standard criterion, truth regarding the verdict of the test, its accuracy, is impossible to capture. To construct and assess our scales, we are left with “validity,” a vague concept that attempts to secure some reality regarding revascularization, but in the absence of a standard criterion, truth regarding the verdict of the test, its accuracy, is impossible to capture. To construct and assess our scales, we are left with “validity,” a vague concept that attempts to secure

The interobserver agreement between 2 radiologists assessing angiographic results according to the TIMI classification scheme in 38 patients shows low-weighted κ values (0.4; 95% CI, 0.2–0.6),15 similar to those in our study. On the other hand, another report has previously shown better agreement among 3 observers by using the TIMI or a new Qureshi grading system in 15 randomly selected patients, with κ values in the range of 0.7.16

By adding statistical noise, variability or lack of precision may affect results of studies comparing 2 treatments. This will, by necessity, impose methodologic adjustments such as increasing the number of patients to be recruited in the study to show a difference between 2 groups. The variability we observed in judging the success of the procedure was probably underestimated because there are many other sources of discrepancies in a core lab context compared with the electronic survey: There are more images, from various series, by using various projections and diverse equipment from various centers. Legitimate strategies to enhance precision include standardization of angiographic projections and techniques, using an operation manual, refining criteria defining the score classes, and training (or even certifying) the observers. Repeating the measurements by a number of observers, with resolution of discrepancies by consensus, can succeed in achieving a precision that is artificial; it is unclear, however, if such verdicts are more valid.34

For sure, predictive validity is important; clinical outcomes have been correlated to revascularization in acute ischemic stroke,9 though many other factors (collateral circulation, penumbra, eloquence of the vascular territory, and so forth) may also impact outcomes.7 To attempt to incorporate all potential factors in a single, intuitively appealing but complex scale with multiple categories may increase the variability of interpretations and, consequently, decrease precision, when the time comes to assess interobserver agreement.35 Various other scales have been described15,16,34 in attempts to intuitively enhance validity. The real test for any scale, however, will eventually come with usage, if it is used. To propose yet another classification could only add to the confusion that already plagues this field.14

The present work suggests that it is possible to live with the current reperfusion and recanalization scales, provided a number of precautions are kept in mind. First, any comparison among revascularization methods or devices should be given credibility when performed within a randomized trial, by using predefined, simple ordinal scales adjudicated by an independent central laboratory staffed with experienced observers masked to treatment allocation. While the verdicts of various observers can be alarmingly divergent for the 2 scales we studied, κ values can reach satisfactory levels of agreement (0.6) when results are dichotomized as success or failure. Surrogate angiographic outcomes can serve a useful function if only to help explain or
understand reasons for disappointing clinical results. Clinical outcomes, typically translated into the modified Rankin Scale scores at 3 months, are probably better end points for major clinical trials.\(^5,18\)

**Limitations**

The electronic survey was designed to ease the assessment by multiple interventional neuroradiologists. Nine of 12 potential observers responded. Readers had access to only 4 selected images to evaluate results according to the TICI and AOL scores. We can only speculate what the results would be if missing responses had been available. How seriously observers worked to come to verdicts can always be questioned, and the context of the assessment is certainly different from typical clinical work or from a core lab context. It may not be realistic to expect readers to assess perfusion on a few static images. The intra- or interobserver agreement was, however, no better when 3 neuroradiologists had access to all images. Posterior circulation strokes were not included, and this feature may have decreased variability in interpretations. New recommendations on angiographic revascularization grading scales have now been published.\(^36\) They include a modified TICI scale, slightly different from the one we used, which is also subject to variability in interpretation.\(^37\)

**CONCLUSIONS**

Recanalization or reperfusion scales are interpreted differently by different observers. Rather than yet another classification scheme, we propose to dichotomize results for analyses and comparisons.

**ACKNOWLEDGMENTS**

We are grateful to the 9 international experts who took time to respond to the survey. This work would not have been possible without their generous help.


**REFERENCES**

5. Kent TA, Maniava P. Recanalization rates can be misleading. *Stroke* 2007;38:e103, author reply e104
13. Kallmes DF. TICI: if you are not confused, then you are not paying attention. *AJNR Am J Neuroradiol* 2012;33:975–76
27. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–74


