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

M. Gaha, C. Roy, L. Estrade, G. Gevry, A. Weill, D. Roy, M. Chagnon, and J. Raymond

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 κ 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 ± 0.07, respectively) or to the full angiographic data ($\kappa = 0.54 \pm 0.06$ and 0.59 ± 0.07, respectively). Intraobserver agreement was "fair to moderate" for both methods. Interobserver 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

Current therapies of acute stroke aim at rapid restoration of blood flow or revascularization of the occluded territory to salvage ischemic brain tissue. A gamut of methods and devices has been introduced to accomplish revascularization.[1-4] While all may agree that the well-being of the patient at the end of treatment is the most important outcome,[5] we also need surrogate markers of mechanistic efficacy, directly linked to the effect we are aiming for, to more expediently determine which method or device should be selected to be tested in a more rigorous fashion, because the heterogeneity of presentations ensures that large trials will be needed to show differences in clinical outcomes. In addition, regulatory agencies approve devices according to their ability to restore blood flow.[6] Thus angiographic scoring systems and a new vocabulary (such as Thrombolysis in Myocardial Infarction [TIMI], TICI, arterial occlusive lesion [AOL], described below) are now used to adjudicate and compare angiographic results of acute stroke therapies.[7-12]

The precision of outcome scales must be assessed before their widespread use. Testing can be accomplished by asking various individuals to repeatedly but independently categorize the angiographic results of the same patients and by studying intra- and interobserver agreement of the resulting verdicts. Despite notes of concern[13,14] and except for small studies limited to 2–3 observers introducing unusual scales[15,16] or comparing 2 scoring systems obtained from consensus reading,[17] the inter- and intraobserver agreement among multiple observers for commonly used systems

Received July 19, 2013; accepted after revision October 17.

From the Department of Radiology (M.G., C.R., G.G., A.W., D.R., M.C., J.R.), Centre Hospitalier de l'Université de Montréal Notre-Dame Hospital, Montreal, Quebec, Canada; Service de Radiologie (L.E.), Hôpital Maison Blanche, CHU Reims, France; and Department of Mathematics and Statistics (M.C.), Université de Montréal, Montreal, Quebec, Canada.

Please address correspondence to Jean Raymond, MD, CHUM-Notre-Dame Hospital, Interventional Neuroradiology, 1560 Sherbrooke East, Pavilion Simard, Room Z12909, Montreal, Quebec, Canada H2L 4M1; e-mail: jean.raymond@umontreal.ca

Indicates article with supplemental on-line table.

http://dx.doi.org/10.3174/ajnr.A3828

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 be analyzed in part 3, to assess intra- and interobserver agreement independently twice, 12 months apart, by 3 expert "internal" readers; and 3) an intra- and interobserver study of the same 32 patients by the same 3 expert readers having access to the full set of angiographic data, directly on the hospital PACS, independently adjudicating results twice, 3–12 months apart, to be compared with the survey of part 2.

Part 1: Electronic Survey with 9 Observers

To minimize variability due to different angiographic equipment, number and type of projections, and selection of final images from various series and to ease the participation of external readers, we assembled a portfolio of 148 images that could be sent electronically to and easily assessed by multiple observers. All anonymized images were retrieved by one author (L.E.) from the PACS of one institution. The portfolio comprised paired (postero-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 κ 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 scale[8] 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 al[7]: grade 0, no perfusion, no antegrade flow beyond the point of occlusion; grade 1, penetration 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 score[22]: 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 κ 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 κ 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 al,[27] measuring the agreement when the number of raters is >2. This method also allowed identifying, for each scale, the categories in

Table 1: Interobserver agreement using the TICI reperfusion scale

	Ob2	Ob3	Ob4	Ob5	Ob6	Ob6a	Ob7	Ob8	Ob9
Ob1	0.497 ± 0.098	0.411 ± 0.102	0.478 ± 0.103	0.544 ± 0.101	0.508 ± 0.104	0.315 ± 0.114	0.517 ± 0.100	0.580 ± 0.100	0.519 ± 0.094
Ob2	0.419 ± 0.100	0.286 ± 0.102	0.576 ± 0.096	0.506 ± 0.102	0.320 ± 0.115	0.458 ± 0.096	0.538 ± 0.099	0.330 ± 0.089	0.330 ± 0.089
Ob3			0.197 ± 0.088	0.284 ± 0.100	0.513 ± 0.103	0.191 ± 0.096	0.339 ± 0.103	0.345 ± 0.105	0.404 ± 0.103
Ob4				0.510 ± 0.100	0.343 ± 0.103	0.352 ± 0.103	0.384 ± 0.098	0.583 ± 0.098	0.297 ± 0.087
Ob5					**0.602 ± 0.101**	0.594 ± 0.102	**0.712 ± 0.091**	**0.752 ± 0.082**	0.397 ± 0.094
Ob6						0.525 ± 0.107	0.465 ± 0.108	**0.610 ± 0.101**	0.425 ± 0.093
Ob6a							0.542 ± 0.096	0.421 ± 0.106	0.283 ± 0.085
Ob7								0.511 ± 0.102	0.442 ± 0.102
Ob8									0.423 ± 0.095
All observers	$\kappa = 0.44570 \pm 0.013176$; $P < 0.001$								

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

Table 2: Interobserver agreement using the dichotomized TICI scale

	Ob2	Ob3	Ob4	Ob5	Ob6	Ob6a	Ob7	Ob8	Ob9
Ob1	**0.773 ± 0.104**	**0.634 ± 0.116**	0.546 ± 0.138	**0.674 ± 0.121**	**0.679 ± 0.117**	0.329 ± 0.156	**0.628 ± 0.122**	**0.727 ± 0.113**	0.578 ± 0.125
Ob2		0.452 ± 0.116	0.507 ± 0.151	**0.673 ± 0.116**	0.580 ± 0.119	0.291 ± 0.162	0.432 ± 0.128	**0.722 ± 0.111**	0.493 ± 0.118
Ob3			0.452 ± 0.116	0.517 ± 0.135	**0.724 ± 0.113**	0.337 ± 0.133	**0.665 ± 0.124**	0.576 ± 0.126	**0.717 ± 0.117**
Ob4				**0.673 ± 0.116**	0.580 ± 0.119	**0.645 ± 0.131**	0.536 ± 0.119	**0.722 ± 0.111**	0.493 ± 0.118
Ob5					**0.784 ± 0.101**	**0.674 ± 0.119**	**0.73 ± 0.111**	**0.946 ± 0.053**	**0.677 ± 0.118**
Ob6						0.575 ± 0.126	**0.727 ± 0.113**	**0.731 ± 0.110**	**0.780 ± 0.103**
Ob6a							**0.634 ± 0.116**	**0.614 ± 0.129**	0.483 ± 0.128
Ob7								**0.679 ± 0.117**	**0.723 ± 0.115**
Ob8									**0.734 ± 0.107**
All observers	$\kappa = 0.61569 \pm 0.024507$; $P < 0.001$								

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

Table 3: Interobserver agreement using the AOL recanalization scale

	Ob2	Ob3	Ob4	Ob5	Ob6	Ob6a	Ob7	Ob8	Ob9
Ob1	0.132 ± 0.066	0.241 ± 0.103	0.288 ± 0.118	0.196 ± 0.096	0.425 ± 0.116	**0.607 ± 0.115**	0.199 ± 0.091	0.265 ± 0.086	0.382 ± 0.098
Ob2		0.420 ± 0.111	0.272 ± 0.084	**0.787 ± 0.103**	0.336 ± 0.093	0.065 ± 0.070	0.376 ± 0.110	**0.648 ± 0.101**	0.352 ± 0.106
Ob3			0.304 ± 0.102	0.553 ± 0.105	0.575 ± 0.106	0.258 ± 0.109	0.474 ± 0.108	0.535 ± 0.109	0.595 ± 0.104
Ob4				0.370 ± 0.095	0.424 ± 0.109	0.340 ± 0.121	0.366 ± 0.090	0.305 ± 0.089	0.467 ± 0.103
Ob5					0.514 ± 0.105	0.263 ± 0.095	0.461 ± 0.110	**0.733 ± 0.093**	0.533 ± 0.107
Ob6						0.290 ± 0.124	0.419 ± 0.111	0.509 ± 0.102	**0.633 ± 0.101**
Ob6a							0.264 ± 0.091	0.265 ± 0.086	0.445 ± 0.098
Ob7								**0.653 ± 0.110**	**0.623 ± 0.116**
Ob8									0.550 ± 0.111
All observers	$\kappa = 0.39434 \pm 0.015957$; $P < 0.001$								

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

which the most frequent disagreements occurred. κ values were interpreted according to Landis and Koch,[27] with κ 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 κ values obtained when the 9 observers scored angiographic outcomes according to the TICI reperfusion categories (overall agreement, $\kappa = 0.446 \pm 0.013$). Table 2 summarizes the κ values when the categories were dichotomized as success (2b, 3) versus failure (0, 1, 2a) (overall agreement, $\kappa = 0.616 \pm 0.025$). κ 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 ($\kappa = 0.242 \pm 0.025$).

Table 3 summarizes the κ values when angiographic outcomes were categorized according to the AOL recanalization categories (overall agreement, $\kappa = 0.394 \pm 0.016$). Table 4 results were obtained when they were analyzed as success (II, III) or failure (0, I) (overall agreement, $\kappa = 0.762 \pm 0.025$). κ 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 ($\kappa = 0.188 \pm 0.025$).

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

Table 4: Interobserver agreement using the dichotomized AOL scale

	Ob2	Ob3	Ob4	Ob5	Ob6	Ob6a	Ob7	Ob8	Ob9
Ob1	**0.646 ± 0.144**	0.593 ± 0.154	**0.613 ± 0.141**	**0.788 ± 0.117**	**0.64 ± 0.144**	**0.773 ± 0.122**	**0.726 ± 0.128**	**0.726 ± 0.128**	**0.726 ± 0.128**
Ob2		**0.65 ± 0.153**	**0.802 ± 0.107**	**0.853 ± 0.101**	**0.85 ± 0.101**	**0.682 ± 0.146**	**0.788 ± 0.117**	**0.788 ± 0.117**	**0.929 ± 0.070**
Ob3			0.491 ± 0.149	**0.654 ± 0.153**	**0.65 ± 0.153**	**0.802 ± 0.133**	0.593 ± 0.154	0.593 ± 0.154	0.593 ± 0.154
Ob4				**0.802 ± 0.107**	**0.802 ± 0.107**	**0.654 ± 0.135**	**0.742 ± 0.120**	**0.742 ± 0.120**	**0.871 ± 0.088**
Ob5					**0.853 ± 0.101**	**0.841 ± 0.108**	**0.929 ± 0.070**	**0.929 ± 0.070**	**0.929 ± 0.070**
Ob6						**0.682 ± 0.146**	**0.788 ± 0.117**	**0.788 ± 0.117**	**0.929 ± 0.070**
Ob6a							**0.773 ± 0.122**	**0.773 ± 0.122**	**0.773 ± 0.122**
Ob7								**0.863 ± 0.094**	**0.863 ± 0.094**
Ob8									**0.863 ± 0.094**
All observers	$\kappa = 0.76197 \pm 0.024507$; $P < 0.001$								

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

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

	Intraobserver (All Images)			Interobserver Agreement		
	Observer 1	Observer 2	Observer 3	Questionnaire	All Images Session 1	All Images Session 2
Full TICI scale	0.488 ± 0.119	0.158 ± 0.122	0.095 ± 0.105	0.590 ± 0.060	0.364 ± 0.060	0.538 ± 0.060
Dichotomized TICI	**0.716 ± 0.132**	0.216 ± 0.179	0.246 ± 0.176	**0.699 ± 0.100**	0.489 ± 0.110	**0.678 ± 0.110**

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

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

	Intraobserver (All Images)			Interobserver Agreement		
	Observer 1	Observer 2	Observer 1	Observer 2	Observer 1	Observer 2
Full AOL scale	**0.611 ± 0.126**	0.233 ± 0.149	0.302 ± 0.143	0.492 ± 0.070	0.238 ± 0.070	0.590 ± 0.070
Dichotomized AOL	**0.731 ± 0.145**	**0.790 ± 0.142**	**0.626 ± 0.168**	**0.787 ± 0.100**	0.561 ± 0.110	**0.870 ± 0.110**

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

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 κ 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 >0.6 with dichotomization of the results of the electronic survey for both scales and for one of the PACS sessions, but not the other.

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) κ 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 reidentifying 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 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 the link between the measure and the phenomenon of interest, such as whether the scale makes intuitive sense (face validity), whether it conforms to theory (construct validity), and whether it allows the prediction of an important clinical outcome (predictive validity).[30] These considerations were taken into account when scales were designed. Revascularization can be conceived as angiographic recanalization of the primary arterial occlusive lesion (what the AOL attempts to capture) or by reperfusion in the arterial bed distal to the occlusion.[22] Reperfusion can be assessed by TIMI, originally used to estimate coronary blood flow after percutaneous angioplasty[31] and used in Prolyse in Acute Cerebral Thromboembolism II,[32] or by TICI, introduced by Higashida et al[7] to intuitively adapt the TIMI scheme to the cerebral circulation and used in IMS-II and III.[18] The results obtained with the AOL classification recanalization system in one study should not be compared with those obtained by using the TICI reperfusion scale in another, the former being more frequently associated with a verdict of success than the latter, at least in this study.

Interobserver disagreement in adjudicating treatment results may be caused by multiple problems: intrinsic ambiguities in the definitions of the classifications; discrepancies in the various ways the definitions are interpreted by various readers; and even if the definitions were understood in the same way, discrepancies in applying the definitions to individual cases. The current study cannot disentangle these various reasons for the discordance among observers. One way to improve agreement would be to modify the proposed classification and retest in a trial-and-error fashion the same portfolio to progressively improve repeatability. It was not feasible to independently test all classification systems and their various modifications in the same study by using the same portfolio and the same observers. Other scales could have been more repeatable than the ones we tested.

The TICI system has been criticized for internal inconsistencies, particularly regarding the 2a, 2b, 3 categories,[13] a problem clearly exposed by the present work, which confirms 2b as the category with the most frequent disagreements. The AOL recanalization categories, however, were also subject to discrepancies in interpretation; the contentious category was AOL II. It may be impossible to obtain consensus for these "gray zone" cases.

Modifications of the TICI scale have been proposed, for example, to get rid of the difficult 2b, 3 distinction between "complete" and "complete-but-slow" perfusion.[33] Others have proposed entirely different classifications.[16]

The interobserver agreement between 2 radiologists assessing angiographic results according to the TIMI classification scheme in 38 patients has shown 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 described[15,16,36] 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.

Disclosures: Daniel Roy and Jean Raymond—*UNRELATED*: *Grants/Grants Pending*: MicroVention,* *Comments*: funding of the Patients Prone to Recurrence after Endovascular Treatment study. Miguel Chagnon—*RELATED*: *Consulting Fee or Honorarium*: Statistical Consulting Service, University of Montreal.* *Money paid to the institution.

REFERENCES

1. Penumbra Stroke Trial Investigators. **The Penumbra pivotal stroke trial: safety and effectiveness of a new generation of mechanical devices for clot removal in intracranial large vessel occlusive disease.** *Stroke* 2009;40:2761–68
2. Mourand I, Brunel H, Costalat V, et al. **Mechanical thrombectomy in acute ischemic stroke: catch device.** *AJNR Am J Neuroradiol* 2011;32:1381–85
3. Smith WS, Sung G, Saver J, et al. **Mechanical thrombectomy for acute ischemic stroke: final results of the multi Merci trial.** *Stroke* 2008;39:1205–12
4. Costalat V, Machi P, Lobotesis K, et al. **Rescue, combined, and standalone thrombectomy in the management of large vessel occlusion stroke using the Solitaire device: a prospective 50-patient single-center study: timing, safety, and efficacy.** *Stroke* 2011;42:1929–35
5. Kent TA, Mandava P. **Recanalization rates can be misleading.** *Stroke* 2007;38:e103, author reply e104
6. Felten RP, Ogden NR, Pena C, et al. **The Food and Drug Administration medical device review process: clearance of a clot retriever for use in ischemic stroke.** *Stroke* 2005;36:404–06
7. Higashida RT, Furlan AJ, Roberts H, et al. **Trial design and reporting standards for intra-arterial cerebral thrombolysis for acute ischemic stroke.** *Stroke* 2003;34:e109–137
8. Khatri P, Hill MD, Palesch YY, et al. **Methodology of the Interventional Management of Stroke III trial.** *Int J Stroke* 2008;3:130–37
9. Rha JH, Saver JL. **The impact of recanalization on ischemic stroke outcome: a meta-analysis.** *Stroke* 2007;38:967–73
10. Goyal M, Fargen KM, Turk AS, et al. **2c or not 2c: defining an improved revascularization grading scale and the need for standardization of angiography outcomes in stroke trials.** *J Neurointerv Surg* 2014;6:83–86
11. Almekhlafi MA, Menon BK, Freiheit EA, et al. **A meta-analysis of observational intra-arterial stroke therapy studies using the Merci device, Penumbra system, and retrievable stents.** *AJNR Am J Neuroradiol* 2013;34:140–45
12. Jauch EC, Saver JL, Adams HP Jr, et al. **Guidelines for the early management of patients with acute ischemic stroke: a guideline for healthcare professionals from the American Heart Association/American Stroke Association.** *Stroke* 2013;44:870–947
13. Kallmes DF. **TICI: if you are not confused, then you are not paying attention.** *AJNR Am J Neuroradiol* 2012;33:975–76
14. Tomsick T. **TIMI, TIBI, TICI: I came, I saw, I got confused.** *AJNR Am J Neuroradiol* 2007;28:382–84
15. Bar M, Mikulik R, Jonszta T, et al. **Diagnosis of recanalization of the intracranial artery has poor inter-rater reliability.** *AJNR Am J Neuroradiol* 2012;33:972–74
16. Qureshi AI. **New grading system for angiographic evaluation of arterial occlusions and recanalization response to intra-arterial thrombolysis in acute ischemic stroke.** *Neurosurgery* 2002;50:1405–14, discussion 1414–15
17. Khatri P, Neff J, Broderick JP, et al. **Revascularization end points in stroke interventional trials: recanalization versus reperfusion in IMS I.** *Stroke* 2005;36:2400–03
18. Broderick JP, Palesch YY, Demchuk AM, et al, for the Interventional Management of Stroke (IMS) III Investigators. **Endovascular therapy after intravenous t-PA versus t-PA alone for stroke.** *N Engl J Med* 2013;368:893–903.
19. Kidwell CS, Jahan R, Gornbein J, et al. **MR RESCUE Investigators: a trial of imaging selection and endovascular treatment for ischemic stroke.** *N Engl J Med* 2013;368:914–23
20. Broderick JP, Schroth G. **What the SWIFT and TREVO II trials tell us about the role of endovascular therapy for acute stroke.** *Stroke* 2013;44:1761–64
21. Feinstein AR, Cicchetti DV. **High agreement but low kappa. I. The problems of two paradoxes.** *J Clin Epidemiol* 1990;43:543–49
22. IMS II Trial Investigators. **The Interventional Management of Stroke (IMS) II study.** *Stroke* 2007;38:2127–35
23. Castaño C, Dorado L, Guerrero C, et al. **Mechanical thrombectomy with the Solitaire AB device in large artery occlusions of the anterior circulation: a pilot study.** *Stroke* 2010;41:1836–40
24. Roth C, Papanagiotou P, Behnke S, et al. **Stent-assisted mechanical recanalization for treatment of acute intracerebral artery occlusions.** *Stroke* 2010;41:2559–67
25. Stampfl S, Hartmann M, Ringleb PA, et al. **Stent placement for flow restoration in acute ischemic stroke: a single-center experience with the Solitaire stent system.** *AJNR Am J Neuroradiol* 2011;32:1245–48
26. Fleiss J, Levin B, Paik M. *Statistical Methods for Rates and Proportions.* New York: Wiley & Sons; 2003
27. Landis JR, Koch GG. **The measurement of observer agreement for categorical data.** *Biometrics* 1977;33:159–74
28. Saver JL, Liebeskind DS, Nogueira RG, et al. **Need to clarify thrombolysis in myocardial ischemia (TIMI) scale scoring method in the Penumbra pivotal stroke trial.** *Stroke* 2010;41:e115–16
29. Cohen J. **A coefficient of agreement for nominal scales.** *Educ Psychol Meas* 1960;20:27–46

30. Newman TB, Browner WS, Cummings SR. **Designing studies of medical tests.** In: Hulley SB, Cummings SR, Browner WS, et al, eds. *Designing Clinical Research: An Epidemiologic Approach.* 4th ed. Philadelphia: Lippincott, Williams & Wilkins; 2001:175–93
31. Chesebro JH, Knatterud G, Roberts R, et al. **Thrombolysis in myocardial infarction (TIMI) trial, phase I: a comparison between intravenous tissue plasminogen activator and intravenous streptokinase—clinical findings through hospital discharge.** *Circulation* 1987;76:142–54
32. Furlan A, Higashida R, Wechsler L, et al. **Intra-arterial prourokinase for acute ischemic stroke: the PROACT II study—a randomized controlled trial. Prolyse in Acute Cerebral Thromboembolism.** *JAMA* 1999;282:2003–11
33. Bankier AA, Levine D, Halpern EF, et al. **Consensus interpretation in imaging research: is there a better way?** *Radiology* 2010;257:14–17
34. Maclure M, Willett WC. **Misinterpretation and misuse of the kappa statistic.** *Am J Epidemiol* 1987;126:161–69
35. Noser EA, Shaltoni HM, Hall CE, et al. **Aggressive mechanical clot disruption: a safe adjunct to thrombolytic therapy in acute stroke?** *Stroke* 2005;36:292–96
36. Zaidat OO, Yoo AJ, Khatri P, et al. **Recommendations on angiographic revascularization grading standards for acute ischemic stroke: a consensus statement.** *Stroke* 2013;44:2650–63
37. Suh SH, Cloft HJ, Fugate JE, et al. **Clarifying differences among thrombolysis in cerebral infarction scale variants: is the artery half open or half closed?** *Stroke* 2013;44:1166–1168