Scales, Agreement, Outcome Measures, and Progress in Aneurysm Therapy

Jean Raymond, Philip M. White and Andrew J. Molyneux

AJNR Am J Neuroradiol 2007, 28 (3) 501-502
http://www.ajnr.org/content/28/3/501

COMMENTARY

Scales, Agreement, Outcome Measures, and Progress in Aneurysm Therapy

Real progress will often be a graduated ascent along a scale of values. If observations can be repeated and can permit reliable predictions, then research is worthwhile and important, and if the predictions withstand critical assessment by many observers, then we enter the scientific world. Scientific methods are necessary to prevent errors and misjudgments, which in medicine, translate into patient morbidity and mortality. Cloft et al,[1] in this issue of *AJNR*, provide us with important reflections and results on methods to promote scientific progress in the treatment of intracranial aneurysms.

They address the main question: How can we assess results of endovascular aneurysm therapy, and more specifically, is the measurement scale we use reproducible? They rightly point out that if the goal is to prevent intracranial bleeding, the most appropriate outcome to measure should be clinical: the frequency of aneurysm ruptures during a certain observation period after treatment. However, most series suggest that bleeding after treatment is too rare an event to serve as the primary outcome of a trial that would be feasible in terms of size or duration of any trial.[2-5]

The main drawback of coiling is the potential for recurrences in 20%–33% of cases on follow-up imaging.[4] This finding carries the implicit threat of future bleeding, though the incidence of delayed bleeding after treatment is, based on current evidence, too low (0.1%–0.3%/year) to jeopardize the initial benefit of coiling.[2-4] Thus, we are forced to design trials that rely on a surrogate end point if we want to assess the efficacy of our new devices. In this instance, the angiographic results are the surrogate end points. Surrogate end points are commonly thought by professional trialists to provide weak evidence. In a worst case scenario, they can be misleading. However, this surrogate end point is our only realistic hope of obtaining objective evidence. Besides, it has a significant clinical implication: The angiographic recurrence rate is a common pretext for recommending further treatment, either surgical clipping or endovascular coiling, both carrying risks for the patient. When one is faced with the introduction of a new device with regulatory approval, the research questions become: 1) whether the new device can improve angiographic results, and 2) what we are willing to pay in terms of morbidity for the patient to improve an imaging finding?

If the angiographic outcome of the procedure and the recurrence phenomenon are to be analyzable and comparable, they require a set of measurements/assessments or "variables". The value of any measurement scale is judged according to accuracy, validity, precision (incorporating reproducibility), sensitivity, specificity, and appropriateness.

Variables come in various types, and in decreasing content of information, are classified as continuous, finite, or categoric; and within the latter, there are ordinal and nominal variables. Although a good general rule is to prefer continuous variables that provide more information, there are many exceptions. In this particular field, the authors are right when they state that the degree of occlusion or magnitude of a recurrence are continuous variables and their representation by ordinal classes involves difficulties that are reminiscent of the famous (or infamous) logical "paradox of the Sorites" (ie, how many grains of wheat are needed to constitute a heap) that gave so much trouble to Chrysippus.[6] However, given the extreme variability in aneurysm shape and size (let alone the variability of the measurement), it would make little sense to rely on millimeters or percentages. The outcome of the procedure should capture what is common and pertinent to all these varied cases. A conclusion such as "the new device resulted in a mean decrease of 2.3 $\pm$ 0.1 mm or *x*% in aneurysm opacification" would carry little intuitive or practical meaning. Here we are concerned about the risk of bleeding or rebleeding, and our scaling system should reflect our intuition and experience with the type of results that might expose the patient to significant hemorrhagic risks. In that vein, a 2- to 3-mm recurrence is more significant near the fundus of a 3-mm aneurysm than at the neck of a 14-mm aneurysm.

The value of a measurement scale is sometimes qualified in terms of accuracy. In the absence of a reference (or gold) standard, such as a pathologic confirmation of a "cured" aneurysm or of a large recurrence, it is difficult to make sense of the accuracy of angiographic results of coiling because this value refers to the degree to which a variable actually represents what it is supposed to represent! We recall that our ordinal grading systems are based on the assumption that completely or near completely occluded aneurysms (both often qualified as satisfactory results) carry minimal if any risk compared with residual or recurrent opacification of any portion of the aneurysmal sac. This assumption is based on the purported fact that aneurysms most often bleed from the sac and rarely from the neck.

Thus we must examine "validity," a more feasible concept when dealing with abstract outcomes and a term that refers to vague but important values concerning the link between the measure and the phenomenon of interest, such as whether the measurement makes intuitive sense (face validity), whether it incorporates most aspects of the phenomenon under study (sample validity), whether it conforms to theory (construct validity), and, most important, whether it allows the prediction of the occurrence of a defined external event (predictive validity), rupture in this case.[7]

Here we could argue that the 3-response scale, though shown to be less precise (more discordant) than the 2-response scale, is more valid because the distinction between complete occlusion and residual neck on initial angiograms was shown to have a predictive value on the incidence of recurrences on follow-up angiograms.[4]

Having conceded the appropriateness and validity of using an ordinal scale in scoring angiographic results, Cloft et al[1] proceeded with an analysis of another important value of measurement scales, precision, and they correctly used the classic methods of assessing the consistency of repeated measurements by different observers or by the same observer on different occasions.

Precision is affected by 3 main sources of random error: observer, subject, and instrument variability. Unless the study is restricted to aneurysms of a certain size or location or to centers using similar methods and equipment, little can be done to limit subject or instrument variability. This inability must have contributed to the difficulties in reading the angiographic results in this study. Thus, the variability between and

within observers reported in this article also includes subject and instrument variability. Strategies to enhance precision include standardizing angiographic projections and techniques, using an operations manual, refining criteria defining the score classes, training (and sometimes certifying) the observers, and repeating the measurements by a number of observers, with resolution of discrepancies by adjudication or consensus. All these measures are important because lack of precision will strongly affect the power of a trial and the sample size necessary to test the hypotheses.[7]

Because we have no more objective way of testing whether our ordinal angiographic scale actually represents what it is supposed to, the concepts of sensitivity and specificity do not apply here, except perhaps when one wants to report stability of angiographic results. For this purpose, a different variable is needed, and one that uses paired comparisons is an elegant way to portray a meaningful change with time, between 2 states of the same individual, despite the wide variability between different individual lesions. The proposed distinction between 2 classes of recurrence (any recurrence or major recurrences) was an attempt to capture either 1) sensitivity in detecting unstable results (any recurrence, in effect a dichotomous variable similar to the worse/no worse system), thus resisting the temptation of embellishing results, or 2) specificity, at the cost of losing sensitivity, in selecting only lesions in which the recurrence is extensive enough to render the hemorrhagic risk intuitively plausible (major recurrence).[4] The authors seem to favor a more precise dichotomous worse/no worse classification of recurrence as opposed to a 3-response scale (improved/same/worse).

An "improved" category seems, at first, helpful when observers have the task of prospectively adjudicating angiographic results as they come and when it is thought necessary for this scale to give an account or explanation for results that appear better at follow-up. We believe, however, that the "improved" category, if needed at all, should not be used as an isolated outcome to compare 2 devices because it cannot distinguish patients who improve because of added biologic or physiologic effect of the device from patients who improve spontaneously because initial results were suboptimal.

Not surprisingly, the authors found the number of categories will significantly affect agreement and a balance is necessary to allow precision without losing too much information. We have previously mentioned, for angiographic results, our preference for a 3-response scale that has more predictive validity. In the evaluation of recurrences, once a recurrence (or "worse") is found, qualifying the recurrence as major or not seems important to resist criticism that the new treatment only improves on a surrogate end point and a smaller number of minor recurrences may not justify the potential risk for increased complications that may come with the new therapy. We doubt, however, given the small number of events, it will ever be possible to predict a differential hemorrhagic risk within a multiple class system.

Most importantly, the authors found good-to-excellent agreement in assessing the angiographic outcomes of therapy. This result is crucial because it means that potential benefits of new devices and therapies can be and should be objectively tested before adoption by the neurointerventional community. This testing should be with standard scientific methods: a randomized comparison between the new treatment and standard platinum coil embolization. Assessment of angiographic outcomes by using an ordinal scale adjudicated by an independent central laboratory staffed with experienced observers masked to treatment allocation can provide appropriate primary outcome measures.

Scientific progress must be differentiated from apparent progress. With the recent multiplication of devices of unproven benefit but that can potentially increase procedural risk, the neurointerventional field has been the arena of much enthusiasm but very little science. Remembering that the danger is borne by our patients, enthusiasm must somehow be tempered. Enthusiasm literally means "having the god enter into the worshipper", and came with the Bacchic rituals. Quoting B. Russell: "Much of what is greatest in human achievement involves some element of intoxication, some sweeping away of prudence by passion. Without enthusiasm, life would be uninteresting; with it, it is dangerous. . . In the sphere of thought, sober civilization is roughly synonymous with science".[8]

We thank Drs Cloft, Kaufmann, and Kallmes for leading us closer to neurointerventional civilization.

References

1. Cloft HJ, Kaufmann T, Kallmes DF. **Observer agreement in the assessment of endovascular aneurysm therapy and aneurysm recurrence.** *AJNR Am J Neuroradiol* 2007;28:XXXXX
2. Molyneux A, Kerr R, Stratton I, et al. **International Subarachnoid Aneurysm Trial (ISAT) of neurosurgical clipping versus endovascular coiling in 2143 patients with ruptured intracranial aneurysms: a randomised trial.** *Lancet* 2002;360:1267–74
3. Molyneux AJ, Kerr RS, Yu LM, et al. **International Subarachnoid Aneurysm Trial (ISAT) of neurosurgical clipping versus endovascular coiling in 2143 patients with ruptured intracranial aneurysms: a randomised comparison of effects on survival, dependency, seizures, rebleeding, subgroups, and aneurysm occlusion.** *Lancet* 2005;366:809–17
4. Raymond J, Guilbert F, Georganos S, et al. **Long-term angiographic recurrences after selective endovascular treatment of aneurysms with detachable coils.** *Stroke* 2003;34:421–27
5. Raymond J, Leblanc P, Chagnon M, et al. **New devices designed to improve the long-term results of endovascular treatment of intracranial aneurysms: a proposition for a randomized clinical trial to assess their safety and efficacy.** *Interventional Neuroradiology* 2004;10:93–102
6. Barnes J, Bobzien S, Mignucci M. **Logic.** In: Algra K, Barnes J, Mansfeld J, Schofeld M, eds. *The Cambridge History of Hellenistic Philosophy.* Cambridge, UK: Cambridge University Press; 2005:170–76
7. Newman TB, Browner WS, Cummings SR. **Designing Studies of Medical Tests.** In: Hulley SB, Cummings SR, Browner WS, Grady D, Hearst N, Newman TB, eds *Designing Clinical Research: An Epidemiologic Approach*. 2nd ed. Philadelphia: Lippincott, Williams & Wilkins; 2001:175–93
8. Russell B: **The Rise of Greek Civilization**. In: *History of Western Philosophy*. London and New York: Routledge Classics; 2004:15–32

Jean Raymond
Centre Hospitalier de l'Université de Montréal, Canada

Philip M. White
Western General Hospital, Edinburgh, UK

Andrew J. Molyneux
Radcliffe Infirmary, Oxford, UK