SPECIAL ARTICLE

Invited Article: Lost in a jungle of evidence

We need a compass

Jacqueline French, MD, FAAN
Gary Gronseth, MD, FAAN

Address correspondence and reprint requests to the American Academy of Neurology, 1080 Montreal Avenue, St. Paul, MN 55116 guidelines@aan.com

A 65-year-old woman presents with her third small subcortical stroke. This occurred despite treatment with antiplatelet medication for secondary stroke prevention. You have maximized risk-factor reduction and the patient has no high-grade carotid stenosis or source of cardiac embolism. You are considering warfarin to prevent another stroke. Looking for direction, you find one study that advocates this decision, while another seems to indicate that warfarin would be of no value in this situation.[1,2] Which study should you believe? If these studies have been reviewed and classified as part of an evidence-based guideline, the answer will be clear: The study that has been rated higher should be given more weight.

One of the essential features distinguishing guidelines from the American Academy of Neurology (AAN) is that the quality and risk of bias in a study is measured using a four-tiered classification-of-evidence scheme (table). It is important for clinicians to understand the classification scheme so that when they see a particular designation they understand its implications. In this scheme, studies graded Class I are judged to have high quality and low risk of bias; studies graded Class II are judged to have moderate quality and risk of bias; studies graded Class III are judged to have a high risk of bias; and studies graded Class IV are judged to have a very high risk of bias.

These classification schemes have been developed using empirically validated criteria for study strength. Although there are differences between the AAN classification scheme and those of other professional societies, they all essentially contain the same elements.[3] For example, based upon the empirical observation that a randomized controlled trial is much more likely to give the correct answer than an observational study, all schemes require randomization for studies to be considered the highest class of evidence.[4] The classification should not be considered a "grade" indicating passing or failing or whether the study is good or bad. Even within a class designation, some studies may be stronger than others. In many circumstances, the only feasible study is at best Class III or IV, and some evidence may be better than none at all. Consider, for example, a study of surgical outcome for a procedure that is considered to be standard of care. As randomization may not be feasible, it will be impossible to carry out a Class I or II study.

The classification scheme employed by the AAN accounts for bias (systematic error) only. Standard statistical measures of random error (*p* values and CIs, for example) discussed in the study itself provide measures of a study's random error.

It is important to recognize that the evidence classification does not pertain to a study, but rather to a question. Thus, the same study could contain a high level of evidence (Class I) for one question, but a low level (Class III or IV) for another. Therefore, a classification should be determined for each of the main clinical questions addressed in the study, if there is more than one. For clarity, the questions should be formulated and listed along with their derived classification.

There are several classification schemes available, depending on whether the question is therapeutic, diagnostic, or prognostic. (The diagnostic and prognostic schemes can be found on the AAN Web site [www.aan.com/go/practice/guidelines/development].)

A therapeutic question appropriate for classification should be as follows: For <patient population A> does <intervention B> compared to <intervention C> improve <outcome D>? For the therapeutic dilemma faced by the clinician in the example above, the question might be posed as follows: For patients with recurrent non-cardioembolic ischemic strokes, is warfarin superior to antiplatelet therapy to prevent recurrent strokes? Note that a more general

See page 1639

From New York University Comprehensive Epilepsy Center (J.F.), New York; and University of Kansas (G.G.), Kansas City.

Disclosure: The authors report the following conflicts of interest: Dr. French holds financial interests in Jazz, Eisai, Valeant, Marinus, Pfizer, and UCB. She has received research funding from the Epilepsy Therapy Development Project, FACES, UCB, Eisai, Johnson and Johnson, and Merck. Dr. French estimates that 30% of her time is spent in outpatient epilepsy practice. Dr. Gronseth has received speaker honoraria from Pfizer, GlaxoSmithKline, and Boehringer Ingelheim and served on the IDMC Committee of Ortho-McNeil. Dr. Gronseth estimates that <2% of his time is spent on EMG and EEG.

Table Classification scheme requirements for therapeutic questions

Class I. A randomized, controlled clinical trial of the intervention of interest with masked or objective outcome assessment, in a representative population. Relevant baseline characteristics are presented and substantially equivalent among treatment groups or there is appropriate statistical adjustment for differences.
The following are also required:
a. Concealed allocation
b. Primary outcome(s) clearly defined
c. Exclusion/inclusion criteria clearly defined
d. Adequate accounting for dropouts (with at least 80% of enrolled subjects completing the study) and crossovers with numbers sufficiently low to have minimal potential for bias
e. For noninferiority or equivalence trials claiming to prove efficacy for one or both drugs, the following are also required*
1. The standard treatment used in the study is substantially similar to that used in previous studies establishing efficacy of the standard treatment (e.g., for a drug, the mode of administration, dose, and dosage adjustments are similar to those previously shown to be effective).
2. The inclusion and exclusion criteria for patient selection and the outcomes of patients on the standard treatment are substantially equivalent to those of previous studies establishing efficacy of the standard treatment.
3. The interpretation of the results of the study is based on an observed-cases analysis.
Class II. A randomized controlled clinical trial of the intervention of interest in a representative population with masked or objective outcome assessment that lacks one criteria a–e above or a prospective matched cohort study with masked or objective outcome assessment in a representative population that meets b–e above. Relevant baseline characteristics are presented and substantially equivalent among treatment groups or there is appropriate statistical adjustment for differences.
Class III. All other controlled trials (including well-defined natural history controls or patients serving as their own controls) in a representative population, where outcome is independently assessed, or independently derived by objective outcome measurement.
Class IV. Studies not meeting Class I, II, or III criteria including consensus or expert opinion.

*Note that numbers 1–3 in Class Ie are required for Class II in equivalence trials. If any one of the three is missing, the class is automatically downgraded to a Class III.

question, such as "Is warfarin better than aspirin?" is not appropriate for classifying evidence, as the answer could differ depending on the scenario, patient population, and outcome of interest. This is why clinical questions must be specific in their formulation. The two studies selected by the clinician to answer the question are relevant because they include the population in question, study the intervention of interest, and measure the pertinent outcomes. Given the therapeutic question for this patient, how would the two studies under consideration, or any study, be classified? Following is a compilation of elements that impact classification in therapeutic articles. These elements will aid in determining which study is more likely to be helpful.

COMPARISON (CONTROL) GROUP A comparison (or control) group in a therapeutic study consists of a group of patients who did not receive the treatment of interest or received an alternative treatment (active control). Studies without a comparison group (or a pre-post treatment comparison) have a higher risk of bias and are graded Class IV. To be graded Class II or higher, studies should have concurrent controls. For example, extracranial-intracranial bypass was considered an effective therapy based on case series. When a randomized controlled trial was undertaken, the procedure was found to be unhelpful.[5] Including patients who were treated medically in the randomized study demonstrated that this cohort had a better outcome than was previously suspected. Studies using non-concurrent controls, such as those using patients as their own controls (e.g., a before-after design) or those using external controls (e.g., historical controls), would at best be Class III. Both of our studies in question include patients who were treated with warfarin or antiplatelet therapy. Thus, both have a comparison group and meet this criterion.

TREATMENT ALLOCATION AND CONFOUNDING A study that ensures that treated and untreated patient groups are similar in every way other than the intervention of interest has the least risk of bias. In other words, known and unknown confounding differences between the treated and untreated groups have been minimized.

Randomized allocation to treatment and comparison groups is the best way to minimize these confounding differences. Thus, a therapeutic study that randomizes patients to treatment arms would be eligible for Class I designation. For example, nonrandomized studies of thymectomy vs medical therapy are confounded by bias in selection of younger patients for surgery, among other things. These patients may have a better prognosis irrespective of treatment.[6] A randomized study to determine whether thymectomy is beneficial is under way.

Additionally, to be graded Class I, the randomization scheme must have effectively balanced the treatment and comparison groups for confounding baseline differences. Unfortunately, a study that has randomly allocated patients in a proper fashion may end up with a Class II designation just by ill luck if substantial differences in baseline characteristics that a reasonable person might presume would impact the results are found after completion. Under some circumstances the study can maintain a Class I designation even with confounding baseline differences if appropriate statistical adjustments can be made.

Finally, for a study to be designated Class I the allocation process must have been sufficiently concealed so that it is clear there could have been no treatment assignment manipulation. Acceptable methods include sequentially numbered opaque envelopes, independent data centers, or telephone call centers. Systematic allocation (e.g., every other patient, even vs odd days) is considered an unacceptable method and would potentially lead to downgrading. Studies have indicated that inadequate treatment allocation schemes can dramatically increase apparent treatment effects.[7]

Nonrandomized studies that successfully match each treated patient with an untreated, comparison patient with similar baseline characteristics (matched cohort design) are eligible for Class II designation. It will be up to the investigator to carefully consider what those characteristics would be.

For the studies in our example, WARS[2] used concealed randomized treatment allocation to ensure that patients receiving warfarin were substantially similar to those receiving antiplatelet therapy, hence the study remains eligible for a Class I designation. In contrast, the nonrandomized study[1] was an observational study, with potential confounding differences between patients receiving warfarin and antiplatelet therapy, hence at best it is eligible for Class III designation.

COMPLETENESS OF FOLLOW-UP Patients enrolled in studies are sometimes lost to follow-up. Losses to follow-up occur for non-random reasons. Such losses may introduce confounding differences between the treated and untreated groups. Thus, more than 80% of patients within the study must have complete follow-up for the study to receive a Class I designation. Studies that do not meet this number are downgraded by one class. An example of the impact of loss to follow-up can be seen in a 6-month comparative trial of rivastigmine vs placebo.[8] More patients failed to complete the study due to side effects in the rivastigmine arm. Thus, since the last cognitive testing observation (obtained earlier in drop-outs) was used as the final outcome, the patients in the rivastigmine arm would appear to have less cognitive decline. This might make rivastigmine look spuriously better than placebo.

For various reasons, sometimes patients initially assigned to the treatment group do not receive treatment and patients assigned to the comparison group do receive treatment. Patients cross over from the treated group to the comparison group for nonrandom reasons. Consequently, if enough patients cross over, the similarity in the characteristics of the treatment and comparison groups attained by the initial randomized treatment allocation may be lost and result in bias.

When patients cross over or drop out, it is important that the investigators analyze the results using intent-to-treat principles. Put simply, this means the investigators analyze the results according to whichever group (treated or comparison) the patient was originally assigned. This ensures that the similarity in characteristics of patients in each group is maintained. The downside of the intent-to-treat analysis is that it can bias the study to show no treatment effect. The upside is that if a difference in outcomes between treated and comparison patients is observed, it is more likely to be related to a treatment effect than to confounding differences between treated and comparison patients.

Failure to use intention-to-treat would lead to downgrading by one class. The only exception would be an active-control equivalence trial (see below). For such trials, if patients drop out, it would be important to demonstrate that completers behave similarly to dropouts, or downgrading may be necessary.

In our studies, >80% of patients in the WARS trial completed the study, and they used intention-to-treat principles for their analysis. Thus, this study remains eligible for Class I designation. More than 80% of the patients in the nonrandomized study completed the study, so no downgrading is necessary.

PRIMARY OUTCOME VARIABLE Many potential outcomes can be derived from a single study. If one primary outcome variable is not selected in advance, it would be possible to select the most favorable outcome post hoc. Therefore, all studies that have not selected a primary outcome variable in advance of the performance of the study would be downgraded by one level. For example, in a study of botulinum toxin use for headache, the primary outcome variable selected was change in the frequency of migraines per month. The study was negative for this outcome, but positive for a secondary outcome of reduction of 2 or more headaches/month.[9] Based on the secondary outcome, some investigators concluded that the

treatment was effective for the treatment of migraine. Subsequent studies did not confirm this.[10]

INCLUSION/EXCLUSION CRITERIA In a well-designed study, the investigator will have considered the characteristics of the individuals who will comprise the best study population without producing confounders. Thus, one would expect such a study to provide a list of inclusion and exclusion criteria for the population under investigation. Studies that do not provide such a list are downgraded by one level.

The primary outcome variables and inclusion/exclusion criteria in both the WARS and nonrandomized trial were prespecified, so downgrading is not necessary.

MASKING A study does not need to be double-blinded to be Class I. For a study to be graded Class I or II, an investigator who is unaware of the patient's original treatment assignment must determine the outcome. To wit, only the outcome assessor needs be masked to treatment assignment. A study could still meet criteria for Class I or II if other members of the treatment team and even the patient are aware of treatment assignment. For an unmasked study to be eligible for Class III, an investigator who is not part of the treatment team (i.e., independent) must determine the outcome. This is necessary because it would be easy to imagine someone from the treating team having a bias when assessing the outcome of a treated patient based on his or her own expectation.[11] A survey that is completed entirely by the patient out of the presence of the treating team could be considered independent. The requirement for masked or independent assessment can be waived if the outcome measure is objective. An objective outcome is one that is unlikely to be affected by expectation bias (e.g., survival or a laboratory assay).

The WARS trial was double-blind, and therefore remains eligible for Class I designation. In contrast, the nonrandomized trial was an observational trial without masking. Arguably, however, the method of determining the presence of recurrent stroke in this study could be determined objectively, so no downgrading is necessary.

ACTIVE CONTROL EQUIVALENCE TRIALS Active control "equivalence" or "noninferiority" trials are considered to represent a special circumstance that requires extra scrutiny. These trials are performed to demonstrate that a new treatment is "as good as" (equivalent) or "not worse than" (noninferior) a standard treatment. Such trials assume that the standard treatment is indeed effective. While the standard treatment may be effective for some populations and under some circumstances, there is no guarantee that it would be effective using the population, methodology, dose, et cetera, of the current trial. Additionally, including patients who did not actually have the disease or who were noncompliant would actually favor the no-difference outcome. Thus, for noninferiority or equivalence trials claiming to prove efficacy for one or both drugs there are additional requirements. For each of the following requirements not met, the study should be downgraded by one level.

1. The standard treatment used in the study should be substantially similar to that used in previous studies establishing efficacy of the standard treatment (e.g., for a drug, the mode of administration, dose, and dosage adjustments should be similar to those previously shown to be effective).[12]
2. The inclusion and exclusion criteria for patient selection and the outcomes of patients on the standard treatment should be substantially equivalent to those previously establishing efficacy of the standard treatment.
3. The interpretation of the results of the study should be based on an observed-cases analysis. For an active control equivalence trial, an intention-to treat analysis could bias the study in favor of equivalence. Imagine, for example, a study where the results for subjects who are noncompliant in the "standard therapy" arm are included on an intent-to-treat basis in the outcome. This could make standard therapy appear less effective and favor equivalence.

Both the WARS and nonrandomized trials are active control equivalence studies, and both meet the preceding criteria. Both used a standard treatment (aspirin) that had previously been shown to be effective. Also, the inclusion and exclusion criteria and outcomes of patients on aspirin were similar to those in previous studies that had proven the efficacy of aspirin. An observed-case analysis was included in the presentation of results. Thus, no downgrading is necessary.

After considering all the elements of study design and execution that impact a study's quality, we conclude that the WARS study has a low risk of bias and attains a Class I designation. The results of the nonrandomized study might have been confounded by variables that could not be accounted for. The study has a moderately high risk of bias, and is designated Class III. Our clinician considers the results of the WARS trial more believable than the nonrandomized trial.

APPLYING THE EVIDENCE TO YOUR PATIENT

The clinician must now apply the evidence from these studies to the patient in front of him or her. After carefully reviewing the studies mentioned above, he or she realizes that the patients included in those studies are not exactly the same as the current patient. Both studies looked at the risk of recurrent stroke in patients with a history of stroke. However, the nonrandomized trial was limited to patients with intracranial artery stenosis. Additionally, the studies were not limited to patients who had a stroke despite already being on antiplatelet therapy. Most likely, such patients were included in both studies but were not the focus of either study. The relative effectiveness of warfarin was not studied in the "antiplatelet-failure" subgroup of patients. Given this, the clinician realizes that he or she must exercise considerable clinical judgments in the selection of antithrombotic therapy for the patient who has "failed" antiplatelet therapy. Realizing the established danger of warfarin and the lack of evidence of superiority over antiplatelet therapy even in patients with recurrent stroke on antiplatelet drugs, he or she reasonably opts to continue the patient on antiplatelet therapy. Given the limitations in the available evidence, another clinician might have looked at this same evidence and reasonably opted for warfarin.

RECOMMENDATIONS FOR THE FUTURE

As shown here, the systematic classification of evidence is extremely helpful to clinicians as they assess the results of different and possibly conflicting trials. It would be a benefit to readers of *Neurology*® to ask authors of therapeutic studies to classify the strength of their study using the AAN therapeutic scheme. Before classification can occur, the authors will have to define the question that their study is meant to address. This in itself is a service, since at times the conclusions of a study can imply that the study results pertain to a larger population than was addressed by the investigation. The specific study question and class of evidence could be indicated in a section at the end of the structured abstract entitled "Strength of evidence."

This section should include the question the investigation was designed to answer, specifically identifying the patient population, intervention of interest, and relevant outcomes. The section would also include the class of evidence as determined by AAN criteria and a brief statement of the results of the study. Hence, the last section of the abstract from a study pertinent to our clinician's patient might read as follows:

Strength of evidence: This study provides Class I evidence that warfarin (target INR 1.7 to 2.5) is equivalent to aspirin 81 mg daily in preventing recurrent strokes during an average of 3 years of follow-up in patients aged 20 to 70 with a history of stroke (relative risk of stroke warfarin vs aspirin 0.98, 95% CIs 0.81 to 1.10).

Received July 9, 2008. Accepted in final form August 13, 2008.

REFERENCES

1. Chimowitz MI, Kokkinos J, Strong J, et al. The Warfarin-Aspirin Symptomatic Intracranial Disease Study. Neurology 1995;45:1488–1493.
2. Mohr JP, Thompson JL, Lazar RM, et al. A comparison of warfarin and aspirin for the prevention of recurrent ischemic stroke. N Engl J Med 2001;345:1444–1451.
3. Lohr KN. Rating the strength of scientific evidence: relevance for quality improvement programs. Int J Qual Health Care 2004;16:9–18.
4. Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. BMJ 1998;317:1185–1190.
5. Failure of extracranial-intracranial arterial bypass to reduce the risk of ischemic stroke: results of an international randomized trial. The EC/IC Bypass Study Group. N Engl J Med 1985;313:1191–1200.
6. Gronseth GS, Barohn RJ. Practice parameter: thymectomy for autoimmune myasthenia gravis (an evidence-based review): report of the Quality Standards Subcommittee of the American Academy of Neurology. Neurology 2000;55:7–15.
7. Schulz KF. Subverting randomization in controlled trials. JAMA 1995;274:1456–1458.
8. Rosler M, Anand R, Cicin-Sain A, et al. Efficacy and safety of rivastigmine in patients with Alzheimer's disease: international randomized controlled trial. BMJ 1999;318:633–638.
9. Silberstein S, Mathew N, Saper J, Jenkins S, for the BOTOX Migraine Clinical Research Group. Botulinum toxin type A as a migraine preventive treatment. Headache 2000;40:445–450.
10. Naumann M, So Y, Argoff C, et al. Assessment: Botulinum neurotoxin in the treatment of autonomic disorders and pain (an evidence-based review): Report of the Therapeutics and Technology assessment Subcommittee of the American Academy of Neurology. Neurology 2008;70:1707–1714.
11. Viera AJ, Bangdiwala SI. Eliminating bias in randomized controlled trials: importance of allocation concealment and masking. Fam Med 2007;39:132–137.
12. Temple R, Ellenberg SS. Placebo-controlled trials and active-control trials in the evaluation of new treatments: part 1: ethical and scientific issues. Ann Intern Med 2000;133:455–463.

Invited Article: Lost in a jungle of evidence: We need a compass

Jacqueline French and Gary Gronseth

Neurology 2008;71;1634-1638
DOI 10.1212/01.wnl.0000336533.19610.1b