AJNR Am J Neuroradiol 20:1584–1596, October 1999

Study Design and Analysis in Neuroradiology: A Practical Approach

L. Santiago Medina

"Medicine is a science of uncertainty and an art of probability."

Sir William Osler

As neuroradiologists, we often emphasize the importance of new imaging technologies as worthwhile and appropriate diagnostic tests for our patients. Through advanced imaging methods, we are able to study in depth the anatomic and physiological changes that affect the body as a result of disease. Quite often, however, we have a tendency to lose our perspective of how efficacious such neuroimaging actually is. As we become more heavily focused on the technical, anatomic, and physiological significance of imaging and its interpretation, it is easy to neglect the assessment of its relative clinical value through the use of appropriately designed studies that could determine diagnostic performance and efficacy. Although mastering the technical and interpretive aspects of the various imaging techniques is important, maintaining a proper clinical perspective through the use of sound studies aimed at determining efficacy and the effect imaging has on clinical outcome is fundamental to the advancement of our field. In the following essay, the necessary building blocks of an appropriate study design and analysis to assess the diagnostic performance and efficacy of a new test are introduced. To better understand these concepts, a hypothetical index case, which illustrates common features in any practical approach to study design and analysis in neuroradiology, is incorporated into the presentation.

Assessing the Research Field

Critical Review of the Literature

Research in imaging, as with all research in medicine, begins with an idea; however, we have no way of knowing if this idea is practical or even new without a thorough review of the available literature. A review of the literature can never be too broad, and should encompass numerous disciplines (eg, neurology, neurosurgery), multiple languages, past decades (eg, pre- and post-Medline search), and medical as well as nonmedical fields (eg, physics, engineering, veterinarian references). Extensive literature searches are essential if we are to avoid the mistake of reinventing the wheel. A thorough literature review allows us to determine what diagnostic tests are available for a specific disease, the limitations of these examinations, and the scientific problems encountered during their development. The literature uncovered should be read and analyzed critically so that potential pitfalls and weaknesses can be found. An in-depth understanding of the disease process and of the state-of-the-art imaging methods used previously and of those proposed are crucial to ensure the right scientific questions are being asked.

Having a Focused Question and Concise Hypothesis

The initial idea should lead to focused questions. Too many questions, or questions that are too broad, serve only to dilute the intent of the scientific quest, and thus drown out any useful or practical information. All major solutions in medicine begin with such focused questions, upon which subsequent discovery can be built.

In addition, it is of vital importance to frame any well-defined scientific question(s) through the use of a well-tailored hypothesis. Remember to answer one hypothesis at a time. Trying to tackle several different hypotheses at once is confusing, time-consuming, and may be misleading. A single fundamental hypothesis should be answered fully before others are even considered. Remember, the eventual scientific quest is likely to be more successful if it consists of individual building blocks (eg, the answers to subsequent questions) stacked one on top of another.

Hypothetical Index Case

1. The idea: Might MR angiography replace conventional carotid angiography in the evaluation of carotid atherosclerotic disease in older patients,

Received December 1, 1998; accepted after revision April 21, 1999.

From the Sections of Health Services and Policy and Pediatric Neuroradiology, Department of Radiology, Children's Hospital Medical Center, 3333 Burnet Ave, Cincinnati, OH 45229.

Address reprint requests to L. Santiago Medina, MD, MPH.

in whom the risk of digital angiography is increased?

2. Literature review: Using angiography as the diagnostic test, the North American Symptomatic Carotid Endarterectomy Trial revealed an absolute risk reduction of 17% in stroke outcome at 2 years after surgical treatment in patients with ischemic symptoms and at least a 70% diameter stenosis of the cervical carotid artery (1, 2). This conclusion was supported by data from conventional carotid angiography and not from MR angiography.

3. Question: Can we replace invasive carotid angiography with a noninvasive diagnostic test, such as contrast-enhanced MR angiography?

Is the Research Project Worthwhile?

Does the hypothesis have clinical relevance? Any scientific evaluation of efficacy in neuroimaging should first address important questions of common diseases. This is important for two reasons. First, time spent in academic pursuits is at a premium in most university centers, and is thus better spent addressing common health issues rather than obscure diseases. Second, and equally important in the academic process, it is easier to get funding if you are pursuing important questions of prevalent disorders. Studying efficacy of neuroimaging in common disorders is also likely to have greater beneficial economic impact and to affect a larger population with respect to health care. That is not to say that the evaluation of neuroimaging in obscure disorders should not follow the same process, just not as a first step in this developing field.

A first step is often to identify teams in your institution that are currently working on various aspects of common disorders. By piggybacking an imaging project onto a major ongoing institutional clinical or basic sciences study, the radiologic investigator may be able to enhance the overall scientific quality of the research study, thereby ensuring its success. For example, a successful multidisciplinary stroke team, which may be conducting a pharmaceutical trial investigating a new drug that prevents stroke, may be using conventional carotid angiography as the standard of reference. Such a study group may openly support the inclusion of a new arm to the existing study design that incorporates a less invasive diagnostic test, such as contrast-enhanced carotid MR angiography, which eventually may prove to be a more cost-effective and lower-risk diagnostic alternative.

Hypothetical Index Case

4. Hypothesis: Contrast-enhanced MR angiography has the same diagnostic performance as conventional carotid angiography.

Does the Scientific Quest Contribute to Medical Science and Patient Care?

All too often, academic neuroradiologists spend enormous amounts of time and effort addressing the use of neuroimaging in unusual diseases or asking irrelevant questions in the evaluation of common diseases. Small irrelevant academic bridges based on individual case reports and limited series describing some neuroimaging findings do not solve major imaging issues in health care. Study design in radiology has also been deficient. In one review in 1994 of over 3125 MR imaging articles, Kent and colleagues (3) reported that only 29 studies were based on sound design and analysis. While one could argue the sound basis of even this study, there is little doubt that, to date, with respect to health services research, sound study designs in radiology have been lacking. Both private and public funding institutions have good reason to scrutinize closely our neuroimaging research endeavors and to support only those that are going to improve the outcome of our patients and the overall health of our population.

This issue also goes back to the value of focused research efforts. All too often, a promising academic career strays off course when certain fundamental rules are not followed. The process of developing a successful academic career often begins as a single idea, which, when answered, leads to new ideas and the answers to new questions. Thus, the bridge is successfully built over time by hard work, perseverance, and concentrated effort. While contributions to the literature are an important aspect of academic life, they alone do not create a strong bridge or a strong career. Investigators must be encouraged by mentors and by example to focus on issues, to find answers to clinical questions, and, more important, to follow those issues to wherever they might eventually lead for the sake of patient care not just for publication. Only by doing so will the radiologic researcher achieve his or her rightful place in the academic community.

Designing the Study

As indicated earlier, the first step in designing a successful study is to put together a multidisciplinary team that understands the problems. Different thoughts and perspectives allow for greater creativity. The team should balance the trade-offs between the scientific depth of the project on one side and the feasibility and costs on the other (4). Different points of view and backgrounds may help simplify the study question and its hypothesis by focusing on efforts to resolve uncertainties that significantly influence clinical and public health policies. A sound study design that balances science with feasibility avoids deadlock and frustration.

Study Population

Inclusion and exclusion criteria are important in determining the study population. Quite often, the study population is too large to allow a feasible scientific study. Therefore, sampling the study population becomes a key issue. For example, studying

the overall role of neuroimaging in children with headache is a broad and difficult task because the prevalence is as high as 38% (5, 6). However, if the inclusion criteria specify only those children with headaches of less than 6 months' duration and with abnormal findings on a neurologic examination, the study population becomes better defined and the ability to answer specific scientific questions becomes more feasible (7). Without a focus that defines a specific high-risk population to be studied, thousands of children would have to be imaged, making the study difficult to interpret and expensive. In general, more rigorous inclusion and exclusion criteria make the sample population more manageable at a lower study cost but with less generalizable results, while less rigorous criteria would have the opposite effect.

Power analysis plays an important role in determining what is an adequate sample size so that meaningful results can be obtained. Power analysis is the probability of observing an effect in a sample of patients if the specified effect size, or greater, is found in the population (4). Mathematically, power is defined as 1 minus β, where β is the probability of having a type II error. Type II errors are commonly referred to as false-negatives in a study population. The other type of error is type I (α), also known as false-positives in a study population (4). For example, if β is set at 0.05, then the researchers acknowledge they are willing to accept a 5% chance of missing a correlation between an abnormal MR angiographic finding and the diagnosis of carotid artery stenosis, to use the hypothetical index case. This represents a power of 1 minus .05, or 0.95, which represents a 95% probability of finding a correlation of this magnitude.

Ideally, the power should be 100% by setting β at 0. In addition, ideally α should also be 0. By doing this, false-negative and false-positive results are eliminated, respectively. In practice, though, powers near 100% are seldom achievable, so, at best, a study should be designed to reduce the false negatives (β) and false positives (α) to a minimum. Achieving an acceptable reduction of false negatives and false positives requires a large subject sample size. Optimal power, β and α, settings are based on scientific need versus the issues of feasibility and cost. For example, assuming an α error of 0.10, your sample size increases from 96 to 118 subjects per study arm (eg, diseased and nondiseased arms) if you change your power from 85% to 90% (8), respectively. If you are budgeted for only 96 subjects per study arm, you may need to accept a higher rate of false positives and false negatives in order to complete the study within budget. Remember, when estimating the sample size, your study population should achieve a 95% confidence interval (CI) for the test's diagnostic performance (ie, sensitivity and specificity), which is reasonably precise for carotid atherosclerotic disease. In general, the more subjects studied per arm, the less data variability and the tighter the 95% CI obtained.

Hypothetical Index Case

5. Population: all patients older than 60 years.

Study sample: patients older than 60 years with carotid atherosclerotic disease symptoms.

6. Power analysis: 85% power with 96 subjects per study arm (carotid and noncarotid atherosclerotic disease arms) and a budget of $192,000; or 90% power with 118 subjects per study arm and a budget of $236,000.

Type of Study

Clinical studies are divided into three main categories: observational, analytical, and experimental (4). The observational study is often called descriptive. These studies observe and describe the different disease processes as seen by imaging. Descriptive studies are usually followed by analytical studies in which case and control groups are selected to determine the diagnostic performance of a test for diseased and disease-free populations. The final study type is often referred to as experimental studies or clinical trials, in which a specific intervention is introduced and the effect of the intervention is measured by using a control group treated with a placebo, an alternative mode of therapy, or a diagnostic test. Clinical trials are epidemiologic designs that can provide data of such high quality that it most closely resembles the controlled experiment done by basic science investigators (9). Clinical trials may be used to assess new diagnostic tests (eg, the impact of adding MR angiography to the work-up of carotid disease) or new radiologic interventional procedures (eg, carotid arterial percutaneous stenting).

Studies are also traditionally divided into retrospective and prospective categories. These refer more to the way the data are gathered than to the specific type of study design. Retrospective studies are usually done to assess rare disorders or to gather information for pilot studies. If the disease process is considered rare, retrospective studies allow the collection of enough subjects to deliver meaningful data. For a pilot project, retrospective studies allow for the collection of preliminary data that can be used to improve the study design of future prospective studies. The major drawback of a retrospective study is the difficulty in obtaining complete clinical data. Tracking missing film jackets and charts is frustrating and time-consuming. Even after hundreds of hours of work, retrospective studies must confront the reality of inherent bias and often the lack of a sound conclusion introduced by the inability to obtain all pertinent data.

Prospective studies are preferred, as they provide better control of the study design and of the quality of data acquired. Prospective studies, even when very large, can be performed efficiently and in a timely fashion if they focus on common diseases, are done at major institutions, and include an ade-

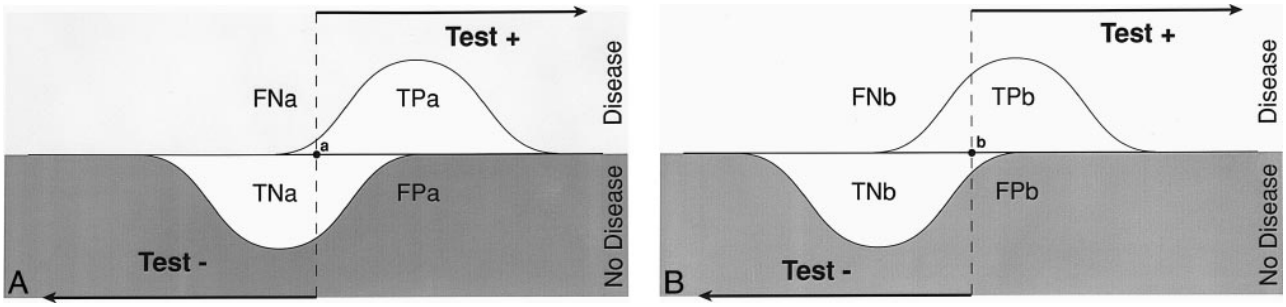

FIG 1. *A* and *B*, Test with a low (*A*) and high (*B*) threshold. The sensitivity and specificity of a test changes according to the threshold selected; hence, these diagnostic performance parameters are threshold-dependent. Sensitivity with a low threshold (*TPa*/diseased patients) is greater than sensitivity with a higher threshold (*TPb*/diseased patients). Specificity with a low threshold (*TNa*/nondiseased patients) is less than specificity with a high threshold (*TNb*/nondiseased patients). *TP* indicates true positive; *TN,* true negative; *FP,* false positive; *FN,* false negative.

quate study population. The major drawback of a prospective study is that it is difficult to ensure that the institutions and personnel comply with strict rules concerning consent, protocol, and data acquisition. Persistence to the point of irritation is crucial to completing a prospective study.

Other specific types of study designs include case-control and cohort studies. Case-control studies consist of groups defined by disease status whereas cohort studies are defined by risk factor status (9). Case-control studies are usually retrospective, in which subjects in a case group (eg, patients who sustained a cerebrovascular accident [CVA]) are compared with subjects in a control group (eg, without CVA) to determine a possible cause (eg, degree of carotid artery stenosis). Cohort studies are usually prospective, in which the risk factor of degree of carotid artery stenosis is correlated with the outcome of CVA.

Hypothetical Index Case

7. Prospective analytical study assessing the diagnostic performance of contrast-enhanced MR angiography in carotid atherosclerotic disease.

Prospective experimental study assessing the therapeutic outcome of endarterectomy versus percutaneous stent placement in patients with carotid atherosclerotic disease.

Mathematical Analysis and Statistics

Diagnostic Performance Evaluation

Statistical analysis of the study population is crucial for determining the scientific validity of the project results. However, statistical analysis should be tailored to each specific study design and hypothesis if it is to have scientific merit. Remember, the statistical results may be wrongfully significant if the study design is not rigorous.

In evaluating diagnostic tests, we rely on the statistical calculations of sensitivity and specificity. Sensitivity refers to the proportion of subjects with the disease who have a positive test (Fig 1); therefore, it indicates how well a test identifies the subjects with disease (4). Specificity is defined as the proportion of subjects without the disease who have a negative test (Fig 1); therefore, it indicates how well a test identifies the subjects with no disease (4). It is important to note that the sensitivity and specificity are characteristics of the test being evaluated and are independent of prevalence (proportion of individuals in a population who have the disease at a specific time), since sensitivity only deals with the diseased subjects while specificity only deals with the nondiseased subjects. However, sensitivity and specificity both depend on a cut-off point set by the investigator, and therefore may change according to which threshold is selected (10) (Fig 1). Given exactly the same diagnostic test, sensitivity with a low threshold is greater than sensitivity with a high threshold. Conversely, specificity with a low threshold is less than specificity with a high threshold.

Defining the presence or absence of an outcome (eg, disease or no disease) is based on a standard of reference. While a perfect standard of reference, or so-called ''gold standard,'' often cannot be obtained, careful attention should be paid to the selection of the standard, which should be widely believed to offer the best approximation to the truth (11). Lack of selection of an adequate reference standard hampers the validity of the study.

The description of test performance as the relation between the true-positive rate and the false-positive rate is called a receiver-operating characteristic (ROC) curve (10) (Fig 2). The ROC curve is used to indicate the trade-offs between sensitivity and specificity for a particular diagnostic test, and hence, describes the discrimination capacity of that test (12). If the thresholds for sensitivity and specificity are variable, a ROC curve can be generated (12). The diagnostic performance of a test is determined by the area under the ROC curve. The higher the ROC curve is skewed toward the left upper corner, the better the discriminatory capacity of the test (Fig 2). A perfect test has an area of 1.0, while a useless test has an area of 0.5 (Fig 2). The

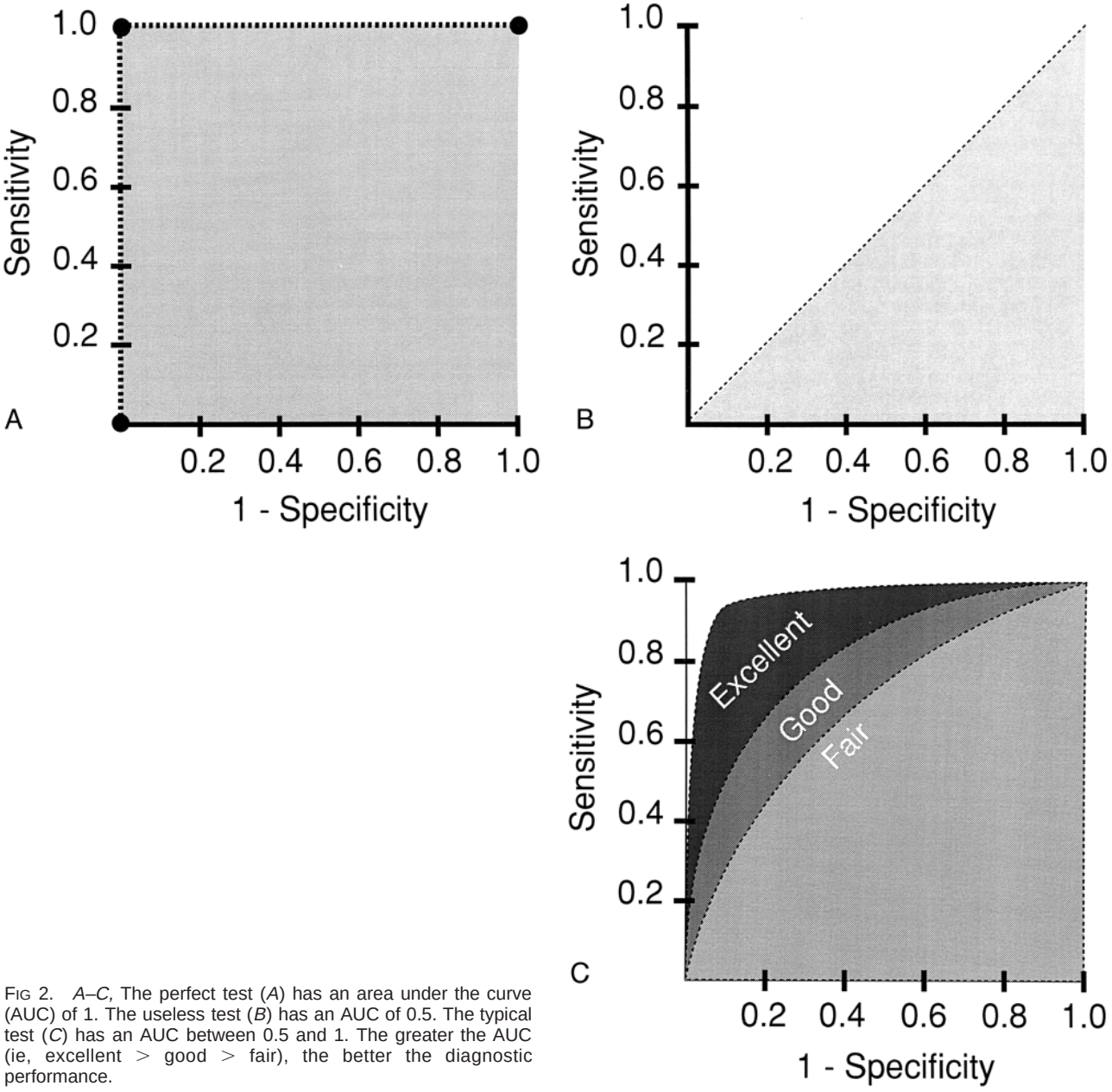

FIG 2. *A–C,* The perfect test (*A*) has an area under the curve (AUC) of 1. The useless test (*B*) has an AUC of 0.5. The typical test (*C*) has an AUC between 0.5 and 1. The greater the AUC (ie, excellent > good > fair), the better the diagnostic performance.

area under the ROC curve also determines the overall diagnostic performance of the test independent of the threshold selected (10, 12). The ROC curve is threshold-independent because it is generated by using variable thresholds of sensitivity and specificity. Therefore, when evaluating a new imaging test, in addition to the sensitivity and specificity, an ROC curve analysis should be done so the threshold-dependent and threshold-independent diagnostic performance can be fully determined, respectively.

The value of the diagnostic test depends not only on the characteristics of the test (ie, sensitivity and specificity, or test information) but also on the prevalence (pretest probability) of the disease in the test population. As the prevalence of a specific disease decreases, it becomes less likely that someone with a positive test actually has the disease and more likely that the test represents a false-positive finding. The relationship between the sensitivity and specificity of the test and the prevalence (pretest probability) can be expressed through the use of Bayes' theorem (10) (point 8 of the Hypothetical Index Case illustrates this relationship).

The positive predictive value of a positive test refers to the probability that a person with a positive test result actually does have the disease. The negative predictive value of a negative test refers to the probability that a person with a negative test result does not have the disease. Since the predic-

tive value is determined once the test results are known (ie, sensitivity and specificity), it actually represents a post-test probability; therefore, the post-test probability is determined by both the prevalence (pretest probability) and the test information (ie, sensitivity and specificity). Thus, the predictive values are affected by the prevalence of disease in the study population. A practical understanding of this concept is shown in point 9 of the Hypothetical Index Case. The example shows an increase in the positive predictive value of a positive test from 67% to 98% when the prevalence is increased from 16% to 82%, respectively. Note that the sensitivity and specificity of 83% and 92%, respectively, are unchanged. If the test information is kept constant (same sensitivity and specificity), the pretest probability (prevalence) affects the post-test probability (predictive value) results. One should be cautious in using the sample prevalence for the general patient population, since it may be very different. Ideally, the patient population prevalence should be estimated from a random, unbiased sample or should be based on a priori estimates of the general patient population.

Hypothetical Index Case

8. Probability revision of Bayes' theorem: Information before test × information from test = information after test; pretest probability (prevalence) × sensitivity / 1 − specificity = post-test probability (predictive value).

9. Predictive value: The predictive value (post-test probability) changes according to the differences in prevalence (pretest probability), although the diagnostic performance of the test (ie, sensitivity and specificity) is unchanged. This example illustrates how the prevalence (pretest probability) can affect the predictive value (post-test probability) with the same test information in three different study groups.

Sample 1: Low Prevalence of Carotid Atherosclerotic Disease

	Disease (Carotid Atherosclerotic Disease)	No Disease (No Carotid Atherosclerotic Disease)	Total
Test positive (positive MR angiography)	20	10	30
Test negative (negative MR angiography)	4	120	124
Total	24	130	154

Results: Sensitivity[a] = 83%; specificity[b] = 92%; prevalence[c] = 16%. Positive predictive value of a positive test[e] = 67%; negative predictive value of a negative test[f] = 98%.

Sample 2: Intermediate Prevalence of Carotid Atherosclerotic Disease

	Disease (Carotid Atherosclerotic Disease)	No Disease (No Carotid Atherosclerotic Disease)	Total
Test positive (positive MR angiography)	100	10	110
Test negative (negative MR angiography)	20	120	140
Total	120	130	250

Results: Sensitivity[a] = 83%; specificity[b] = 92%; prevalence[c] = 48%. Positive predictive value of a positive test[e] = 91%; negative predictive value of a negative test[f] = 86%.

Sample 3: High Prevalence of Carotid Atherosclerotic Disease

	Disease (Carotid Atherosclerotic Disease)	No Disease (No Carotid Atherosclerotic Disease)	Total
Test positive (positive MR angiography)	500	10	510
Test negative (negative MR angiography)	100	120	220
Total	600	130	730

Results: Sensitivity[a] = 83%; specificity[b] = 92%; prevalence[c] = 82%. Positive predictive value of a positive test[e] = 98%; negative predictive value of a negative test[f] = 55%. For equations a, b, c, e, and f, see Appendix.

Note.—As the prevalence of carotid atherosclerotic disease increases from 16% (low) to 48% (intermediate) to 82% (high), the positive predictive value of a positive contrast-enhanced MR angiogram increases to 67%, 91%, and 98%, respectively. Note that the sensitivity and specificity remain unchanged at 83% and 92%, respectively. This example also illustrates that diagnostic performance (ie, sensitivity and specificity) is a characteristic of the test and hence, it is independent of the prevalence (pretest probability).

Accuracy

Accuracy as a statistical term is defined as the weighted average of the sensitivity and specificity. As indicated earlier, the sensitivity and specificity give very different information about the diagnostic performance of a test. In some cases, the information provided by the accuracy value may be misleading because one is weighing together two

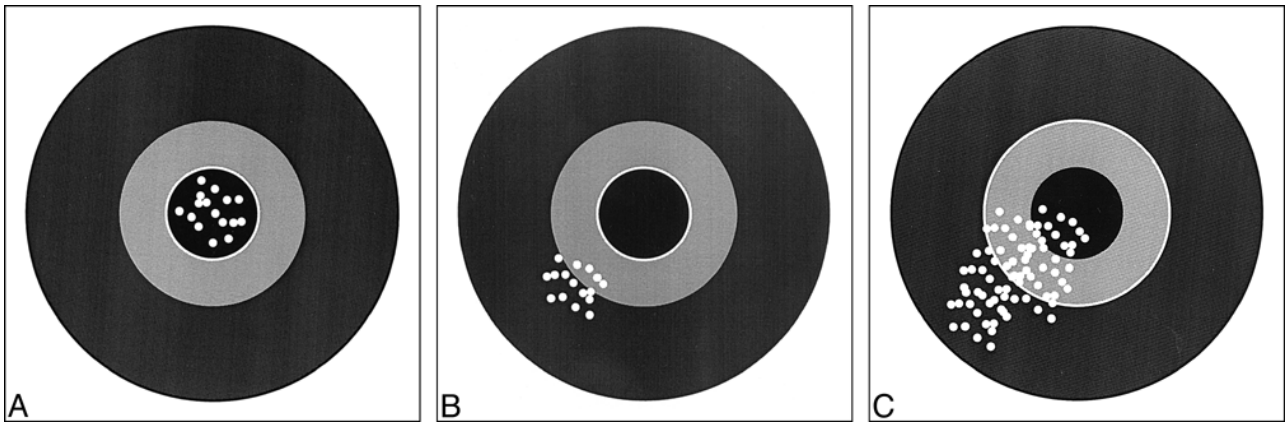

FIG 3. Validity and reliability of a test.
A, Good validity and reliability means the center of the target is always hit by different gunmen.
B, Good reliability but poor validity means the same area of the target is hit by different gunmen but off center.
C, Poor reliability of the test means the bullets are spread out, all over the target, with a few hitting the center. In the hands of a few gunmen a test can be valid, since the center of the target is hit.

very different diagnostic performance parameters. Point 10 of the Hypothetical Index Case illustrates this important point. Two different diagnostic tests (eg, MR angiography and CT angiography) may show an accuracy of 91.9% with very different sensitivity and specificity, of 85.7% and 92.3%, respectively, for diagnostic test 1 (MR angiography) versus 95.2% and 91.3%, respectively, for diagnostic test 2 (CT angiography). Therefore, if the accuracy of a test is to be reported, it should always be accompanied by the sensitivity and specificity.

Hypothetical Index Case

10. Accuracy: The following example illustrates how different the sensitivity and specificity of two different tests are although the accuracy is exactly the same. Hence, accuracy may be misleading if the sensitivity and specificity are not given.

Diagnostic Test 1 (MR Angiography)

	Disease (Carotid Atherosclerotic Disease)	No Disease (No Carotid Atherosclerotic Disease)	Total
Test positive (positive MR angiography)	6	10	16
Test negative (negative MR angiography)	1	119	120
Total	7	129	136

Results: Accuracy[d] = 91.9%; sensitivity[a] = 85.7%; specificity[b] = 92.3%.

Diagnostic Test 2 (CT Angiography)

	Disease (Carotid Atherosclerotic Disease)	No Disease (No Carotid Atherosclerotic Disease)	Total
Test positive (positive CT angiography)	20	10	30
Test negative (negative CT angiography)	1	105	106
Total	21	115	136

Results: Accuracy[d] = 91.9%; sensitivity[a] = 95.2%; specificity[b] = 91.3%. For equations a, b, and d, see Appendix.

Note.—The accuracy of these two noninvasive diagnostic tests, MR angiography and CT angiography, is the same at 91.9%. The sensitivity and specificity, however, are very different for MR angiography and CT angiography, at 85.7%, 92.3% and 95.2%, 91.3%, respectively.

Validity and Reliability

Validity indicates whether the diagnostic test is measuring what was intended. For example, the degree of stenosis measured by MR angiography reflects the actual amount of carotid artery stenosis. Reliability indicates that repeated measurements by the same observer (intraobserver reliability) or different observers (interobserver reliability) produce the same or similar results (13).

Validity and reliability are better understood by using the target example (Fig 3). The validity of the test is measured by the number of times the test hits the center of the target. The reliability of the test is defined as the number of times the same

specific target area is repeatedly hit. A valid test is one with high sensitivity and specificity, thus allowing us to hit the center of the target often. On the other hand, reliability is measured as the degree of intraobserver or interobserver agreement, and hence demonstrates how often the same area of the target is hit by the same reader or different readers, respectively. For example, carotid sonography may be a valid test for assessing atherosclerotic disease because of its high sensitivity and specificity. However, its reliability may vary according to the operator (low intraobserver agreement) or operators (low interobserver agreement); therefore, how often the center of the target is hit (validity) depends on the expertise (diagnostic performance) of the operator(s). Point 11 of the Hypothetical Index Case further emphasizes this point by illustrating interobserver agreement between two sets of readers. In this example, the number of disagreements between the readers is the same, at six; interobserver reliability varies according to the agreement between the readers, from 0.43 to 0.75, as the number of agreeable cores increases. Remember, the kappa statistic assesses the strength of agreement by demonstrating statistically the extent to which observer agreement exceeds that expected purely by chance (14, 15).

Validity is further divided into internal and external categories (4). Internal validity applies to the diagnostic test for the study sample; external validity applies to the general population, which is defined as the world outside the study sample. Determining how the study results (internal validity) may be generalized to the overall population (external validity) is an important study design challenge.

Hypothetical Index Case

11. Interobserver agreement (reliability) for readers 1 and 2 using MR angiography and CT angiography in patients with carotid atherosclerotic disease.

Sample 1: Interobserver Agreement (Reliability) for Readers 1 and 2 using MR Angiography

	Reader 1: Positive MR Angiography	Reader 1: Negative MR Angiography	Total
Reader 2: Positive MR angiography	13	6	19
Reader 2: Negative MR angiography	6	18	24
Total	19	24	43

Results: MR angiography κ statistic = .43[g]. This denotes a satisfactory interobserver agreement (reliability) for MR angiography (14).

Sample 2: Interobserver Agreement (Reliability) for Readers 1 and 2 using CT Angiography

	Reader 1: Positive CT Angiography	Reader 1: Negative CT Angiography	Total
Reader 2: Positive CT angiography	40	6	46
Reader 2: Negative CT angiography	6	45	51
Total	46	51	97

Results: CT angiography κ statistic = .75[g]. This denotes an excellent interobserver agreement (reliability) for CT angiography (14). For equation 8, see Appendix.

Note.—MR angiography (sample 1) and CT angiography (sample 2) had the same number of disagreements between the readers; however, the interobserver agreement was higher with CT angiography than MR angiography as illustrated by a higher agreement core and hence κ statistic.

Limitations

Errors

Diagnostic tests are prone to errors caused by bias and chance. The design and analysis phases are important in dealing with these errors (9).

Systemic Error.—Systemic error is an erroneous result due to bias (4). Biases are sources of variation that distort the study results in one direction. The most common are sampling and measurement biases.

Sampling bias: This bias occurs when the study sample is not representative of the target population for which the test will be used. The disease subjects may come from a referral center; hence, the test's sensitivity for the population in general may be overestimated. The best way to deal with this bias is by selecting a sample similar to the population in which the test is actually going to be performed.

In addition, sampling bias may be introduced into the study if the diagnostic performance of the test is disclosed before the end of the study. This may change the referral pattern and thus affect the exact understanding of the test performance and validity. Ethical issues may press for early disclosure of the study results, but this should only be done if there is strong evidence that the diagnostic performance of the test is directly linked to a significantly better outcome for the study population.

Another common problem during the execution of a study is noncompliance of some subjects. Noncompliance should be reported because it is

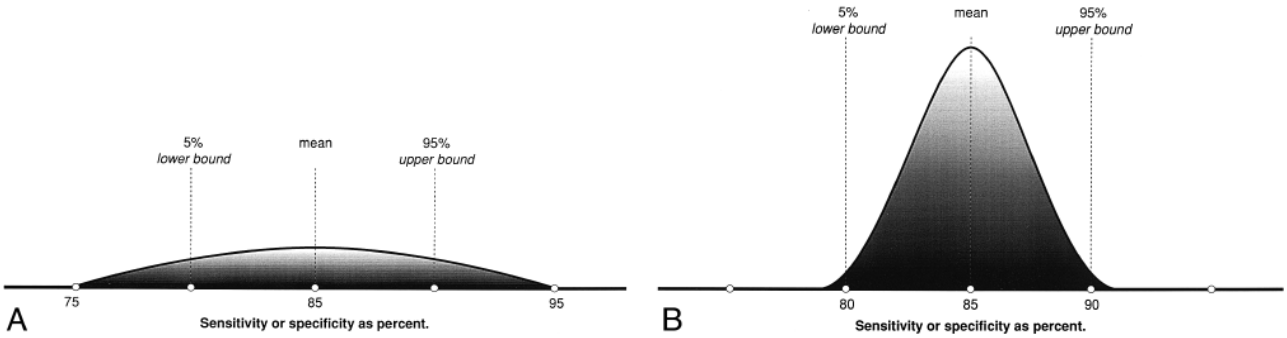

FIG 4. *A* and *B,* Although the mean is the same for study populations of both small (*A*) and large (*B*) patient samples, the CI narrows as the number of subjects increases.

another source of sampling bias that may over- or underestimate the diagnostic performance of a test. Noncompliance must be minimized in order to decrease this important bias.

Another type of sampling bias is based on the prevalence (pretest probability) of disease in the study population. As we indicated earlier, positive and negative predictive values of the test are highly dependent on the prevalence of disease. Ideally, estimates of prevalence should be based on large prospective studies or on prior estimates of prevalence for each risk group (9, 10). Prevalence obtained from sample data may not reflect the prevalence for the specific group studied. Point 9 of the Hypothetical Index Case illustrated this point. The sensitivity and specificity are kept constant at 83% and 92%, respectively. However, as the prevalence (pretest probability) is increased from 16% to 48%, the positive predictive value of a positive test increases from 67% to 91%, respectively. Furthermore, if the prevalence increases to 82%, the positive predictive value of a positive test becomes 98%. Showing different predictive value calculations according to the prevalence (pretest probability) helps to illustrate the usefulness of the test according to different risk groups found in clinical practice. However, these prevalences should be independent estimates based on a priori assessments rather than on the sample.

Measurement bias: Study design and analysis of diagnostic tests are prone to various types of measurement bias (4). If the final outcome is known to the reader interpreting the test, the risk of measurement bias is increased. These biases are also referred to as test-review bias and diagnostic-review bias (16). It is easy to imagine that if the researchers interpreting the findings on contrast-enhanced MR angiograms obtained to assess carotid atherosclerotic disease know the results of the conventional angiogram, their interpretations may be significantly influenced. In order to obtain objective scientific assessment of a new imaging test, all readers should be blinded to other diagnostic tests and final diagnosis; and all patient-identifying marks on the test should be masked.

Another common type of measurement bias arises when the interpretation criteria and test results classification are not clearly set at the onset of the research study. Such tests often produce borderline or technically inadequate results (4). It is fundamental to decide in advance how these results are going to be treated. Disregarding these test results when the diagnostic performance is being determined may lead to misleading findings because the sensitivity and specificity may look better than they actually are.

The best way to confront all types of measurement bias is by determining a priori the criteria for a positive diagnostic test and disease state. In addition, diagnostic test and final outcome results should be determined in a blinded manner.

Random Error.—Random error is an erroneous result due to chance (4, 9). By chance alone, some patients with disease may have a diagnostic test with normal results. Random error is unavoidable, but it may be quantifiable. The best way to quantify random error is by using a 95% CI (Fig 4). CIs allow the reader to see the range of diagnostic performance around the mean sensitivity and specificity (Fig 4). The larger the number of subjects studied, the tighter the 95% CI. One way of estimating the optimal sample size for a test being evaluated is by calculating the mean and 95% CI for different numbers of subjects within the study. The Hypothetical Index Case (point 8) demonstrates that as you increase your sample size from 50 to 250 subjects, your 95% CI interval for sensitivity and specificity becomes tighter around the mean. The tighter the 95% CI, the more consistent is the diagnostic performance of the test.

Hypothetical Index Case

12. 95% CI: In the following example, the sensitivity, specificity, and prevalence are kept the same for the two study sample sizes. Note that the 95% CI becomes tighter as the study population increases.

Small Sample Population of 50 Patients

	Disease (Carotid Atherosclerotic Disease)	No Disease (No Carotid Atherosclerotic Disease)	Total
Test positive (positive MR angiography)	20	2	22
Test negative (negative MR angiography)	4	24	28
Total	24	26	50

Results: Sensitivity[a] = 83%; specificity[b] = 92%; prevalence[c] = 48%. 95% CI[h] sensitivity = 73%, 94%; 95% CI[h] specificity = 85%, 100%.

Large Sample Population of 250 Patients

	Disease (Carotid Atherosclerotic Disease)	No Disease (No Carotid Atherosclerotic Disease)	Total
Test positive (positive MR angiography)	100	10	110
Test negative (negative MR angiography)	20	120	140
Total	120	130	250

Results: Sensitivity[a] = 83%; specificity[b] = 92%; prevalence[c] = 48%. 95% CI[h] sensitivity = 79%, 88%; 95% CI[h] specificity = 89%, 96%. For equations a, b, c, and h, see Appendix.

Note.—As the number of patients is increased from 50 to 250, the sensitivity's 95% CI becomes tighter from 73%, 94% to 79%, 88%. Similarly, the specificity's 95% CI goes from 85%, 100% to 89%, 96%.

Diagnostic Technology and Technique

It is important to master the technology and protocol(s) used for a specific test (eg, contrast-enhanced carotid MR angiography) in order to determine its diagnostic performance. Using a relatively mature technique allows more meaningful long-term results. Newer experimental technologies or techniques are, in the short term, more prone to changes in diagnostic performance as they become perfected. Once a specific technology and protocol are selected, the same yardstick should be used to measure all the outcomes. Changing the yardstick (ie, protocol) halfway through the study creates measurement bias, since we may be using tests with different diagnostic performance.

Hypothetical Index Case

13. Three different noninvasive diagnostic tests are being evaluated for carotid atherosclerotic disease: 1) gray-scale and Doppler sonography, 2) CT angiography, and 3) contrast-enhanced MR angiography.

The best diagnostic strategy for the diagnosis of carotid atherosclerotic disease may be a combination of the different noninvasive diagnostic tests rather than a single one. Having an open mind to other diagnostic tests and different test combinations usually yields the best results.

How Much Does the Research Study Cost?

Clinical research can be very expensive. Using good judgment and common sense is critical. Trade-offs between cost and feasibility should be considered (4). There is no need to solve the whole puzzle at once. Therefore, the most important and feasible scientific hypotheses should be answered in the least costly manner.

Hypothetical Index Case

14. Selecting the diagnostic test with the potentially highest yield can improve the study's feasibility and decrease the total study cost. The study with the potentially highest yield may be determined from the results of less expensive pilot studies.

Following are two proposals the research team is considering: 1) study 100 patients with unenhanced MR angiography and contrast-enhanced MR angiography, for a total cost of $240,000; or 2) study 100 patients with only contrast-enhanced MR angiography, for a total cost of $195,000.

The research team is probably better off with proposal 2, since it is less expensive, more feasible, and most likely to provide all the relevant clinical information.

Starting the Research Project

The pilot study is the miniature model for the whole project. Ideally, the pilot study should be performed in one or a few institutions so the research team can become familiar with the hypothesis, study design, sample population, imaging protocols, potential limitations, and costs. The preliminary results from the pilot study should be presented to the whole research team and, ideally, at national meetings in order to get constructive criticism. Modifications to the study design and imaging protocols should be done at this stage. Changes should be reviewed over and over again until the project is running smoothly and ready for a multicenter trial, in which adequate patient numbers and statistical power can be achieved.

Finishing the Research Project

Once the initial framework (the pilot study) of the bridge has been constructed and tested for its

integrity, a large multi-institutional study can be launched to complete final construction. Conducting a multicenter trial is like constructing a bridge with products assembled in different geographic locations. The major advantage of multicenter studies is that they allow the possibility of obtaining a large sample size in a relatively short period of time, thus permitting rapid assessment of newly emerging diagnostic tests. Another advantage is the institutional heterogeneity that allows assessment of the diagnostic test performance in communities with different patterns of practice. Both academic and nonacademic centers, therefore, should be included in order to account for interinstitutional variability (11).

The major drawback of multicenter trials is their complexity given the variability among institutions and their geographic spread. The three key components to overcoming these problems are study managers, communication, and quality control standards.

An experienced study center headquarters with a seasoned general manager should be selected to supervise the whole research operation (11). Each participating institution should have a site database manager. Continuous communication among the managers is crucial so deadlines can be met in a timely fashion and problems can be solved early on. Modes of communication include electronic mail, telephone conferences, and periodic face-to-face meetings (11). At the same time, open communication between the managers and the rest of the team (eg, imagers, referring physicians, hospital administrators, statisticians) is fundamental in running a tight ship.

Quality in patient selection, technique, and protocol should be closely scrutinized. Inclusion and exclusion criteria should be followed rigorously to maintain the validity of the study design. Strict adherence to the imaging protocols with a high-quality technique is crucial for ensuring that the best performance is being obtained from the diagnostic test evaluated. Falling short in quality may make the study vulnerable to the criticism that the results do not reflect the maximal capability of the test (11).

Making Available the Scientific Results

Once the study is completed, the scientific data should be both presented and published in order to reach the broadest target audience and to get constructive criticism. An article should be expeditiously published in a major journal that reaches the most appropriate audience. The readership ideally should be composed of physicians and nonphysicians. It is hoped that this approach will expose your new diagnostic test to practicing physicians and policy decision makers. Broad, scientific promotion of your study, without flamboyancy, is the key to having your valuable work become important in health policy matters rather than seeing it become just another reference in the literature.

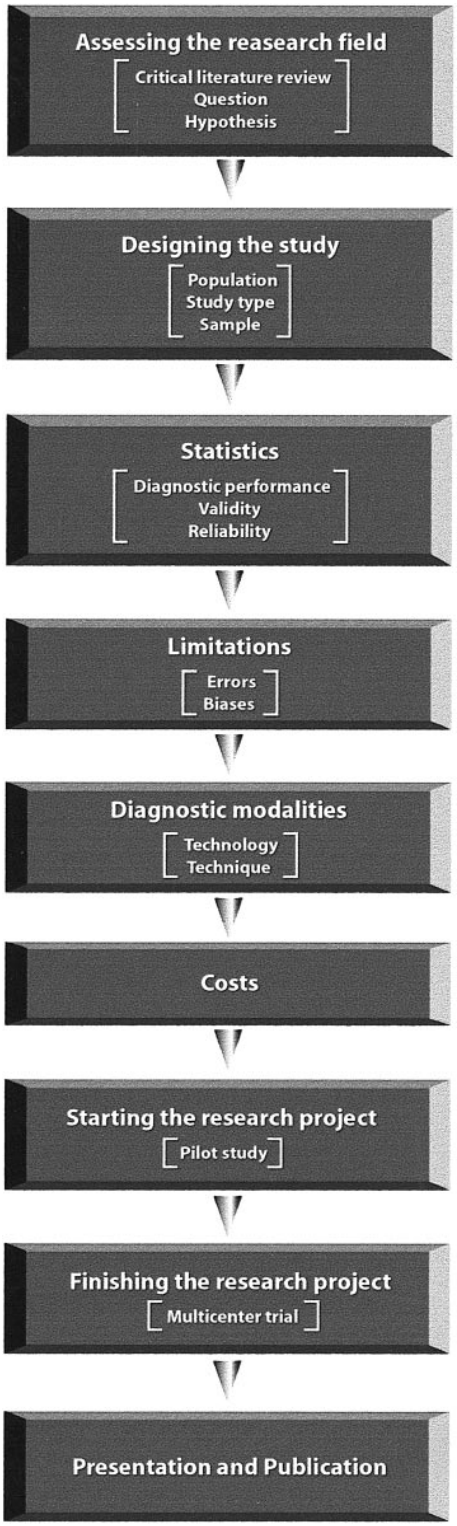

FIG 5. Flow diagram summarizes the key steps required to do sound clinical research.

Sound Research Is Not Done Overnight

Research brings out our exciting and creative being. However, major scientific breakthroughs

are unusual. Therefore, careful planning and execution over many years yield the most prosperous results. Science is a long-term investment that is based on adding small but true bits of information to the accumulated knowledge of a discipline (4).

Conclusion

Clinical research in neuroradiology should strictly follow the scientific method. Methodical planning and execution of the study design and analysis are crucial in achieving meaningful results. The key blocks for building sound research include 1) assessing the research field in depth (critical literature review, questions, and hypothesis); 2) designing the study (population, study type, and sample); 3) applying mathematical analysis and statistics (diagnostic performance evaluation, validity, and reliability); 4) accounting for limitations (errors and biases); 5) implementing diagnostic technology and technique; 6) estimating cost; 7) starting the project (pilot study); 8) finishing the research (multicenter study); and 9) presenting and publishing the results (Fig 5). We as neuroradiologists have significant scientific knowledge to offer our patients and the neuroscientific community; we just have to do it right in order to have a meaningful impact on public health.

Acknowledgments

I am indebted to William S. Ball, Jr, Janet L. Strife, Olga L. Villegas-Medina, and David Zurakowski for their valuable comments and editing. In addition, I am indebted to Kathleen Joiner for her constructive comments and secretarial assistance, and to Glenn Miñano for his assistance in graphics.

APPENDIX

Equations:

Nomenclature for two-by-two table:

	Disease	No Disease
Test +	a(TP)	b(FP)
Test −	c(FN)	d(TN)

a. Sensitivity: $[a/(a + c)*100$

b. Specificity: $[d/(d + b)]*100$

c. Prevalence*: $\frac{\text{Diseased}}{\text{Diseased} + \text{Nondiseased}}$

d. Accuracy: $[(a + c)*(\text{sensitivity}) + (b + d)*(\text{specificity})]/\text{Total} = \frac{a + d}{\text{Total}}$

e. Positive predictive value of a positive test*: $[a/(a + b)]*100$

f. Negative predictive value of a negative test*: $[d/(c + d)]*100$

g. kappa statistic (Intra- and Interobserver agreement) $\kappa = \frac{Po - Pe}{1 - Pe}$

Po = observed agreement
Pe = expected agreement
$\kappa > .75$ denotes excellent agreement
$.4 \leq \kappa \leq .75$ denotes good agreement
$0 \leq \kappa < .4$ denotes marginal agreement

h. 95% CI: $P \pm 1.96* \sqrt{\frac{P(1 - P)}{n}}$

P = mean
n = number of subjects

* Only correct if prevalence is estimated from a random, unbiased sample or based on an a priori estimate of the general patient population.

Note.—**TP** indicates true positive; **FP**, false positive; **FN**, false negative; **TN**, true negative.

References

1. Brant-Zawadzki M. **The roles of MR angiography, CT angiography and sonography in vascular imaging of the head and neck.** *AJNR Am J Neuroradiol* 1997;18:1820–1825
2. Executive Committee for the ACAS Study. **Endarterectomy for asymptomatic carotid arteries stenosis.** *JAMA* 1995;273:1421–1428
3. Kent DL, Haynor DR, Longstreith WT, Larson EB. **The clinical efficacy of magnetic resonance imaging in neuroimaging.** *Ann Intern Med* 1994;120:856–871
4. Hulley SB, Cummings SR. ***Designing Clinical Research: An Epidemiologic Approach.*** Baltimore: Williams & Wilkins; 1998:v–135
5. Sillanpaa ML. **Headache in children.** In: Olesen J, ed. *Headache Classification and Epidemiology.* New York: Raven; 1994:273–281
6. The Childhood Brain Tumor Consortium. **The epidemiology of headache among children with brain tumor.** *J Neurooncol* 1991;10:31–46
7. Medina LS, Pinter JD, Zurakowski D, et al. **Children with headache: clinical predictors of surgical space occupying lesions and the role of neuroimaging.** *Radiology* 1997;202:819–824
8. Donner A. **Approaches to sample size estimation in the design of clinical trials: a review.** *Stat Med* 1984;3:199–214
9. Hennekens CH, Buring JE. ***Epidemiology in Medicine.*** Boston: Little, Brown; 1987:23–212
10. Weinstein MC, Fineberg HV. ***Clinical Decision Analysis.*** Philadelphia: Saunders; 1980:75–130
11. Sunshine JH, McNeil BJ. **Rapid method for rigorous assessment of radiology imaging technologies.** *Radiology* 1997;202:549–557
12. Metz CE. **Basic principles of ROC analysis.** *Semin Nucl Med* 1978;8:283–298

13. Steiner DL, Norman GR. ***Health Measurements Scales: A Practical Guide to the Development and Use.*** 2nd ed. Oxford, England: Oxford University Press; 1995:4–14
14. Rosner B. ***Fundamentals of Biostatistics.*** New York: Duxbury Press; 1995:426
15. Agresti A. ***Categorical Data Analysis.*** New York: Wiley; 1990: 366–370
16. Begg CB, McNeil BJ. **Assessment of radiologic test: control of bias and other design considerations.** *Radiology* 1988;167:565–569

Other Suggested Readings

1. Fryback DG, Thornbury JR. **The efficacy of diagnostic imaging.** *Med Decis Making* 1991;11:88–94
2. Brismar J, Jacobsson B. **Definition of terms used to judge the efficacy of diagnostic tests: a graphic approach.** *AJR Am J Roentgenol* 1990;155:621–623
3. Black WC. **How to evaluate the radiology literature.** *AJR Am J Roentgenol* 1990;154:17–22
4. McNeil BJ, Keeler E, Adelstein SJ. **Primer on certain elements of medical decision making.** *N Engl J Med* 1975;293:211–215
5. Reid MC, Lachs MS, Feinstein AR. **Use of methodological standards in diagnostic test research: getting better but still not good.** *JAMA* 1995;274:645–651
6. Burton E, Troxclair D, Newman W. **Autopsy diagnoses of malignant neoplasms: how often are clinical diagnoses incorrect?** *JAMA* 1998;280:1245–1248
7. Black WC, Welch HG. **Advances in diagnostic imaging and over estimations of disease prevalence and the benefits of therapy.** *N Engl J Med* 1993;328:1237–1243
8. Hanley JA, McNeil BJ. **The meaning and use of the area under a receiver operating characteristic (ROC) curve.** *Radiology* 1982;143:29–36
9. Metz CE. **ROC methodology in radiologic imaging.** *Invest Radiol* 1986;21:720–733
10. Brismar J. **Understanding receiver-operating-characteristic curves: a graphic approach.** *AJR Am J Roentgenol* 1991;157: 1119–1121
11. Black WC, Dwyer AJ. **Local versus global measures of accuracy: an important distinction for diagnostic imaging.** *Med Decis Making* 1990;10:266–273
12. Sox HC, Blatt MA, Higgins MC, Marton KI. ***Medical Decision Making.*** Boston: Butterworth-Heinemann; 1988
13. Ransohoff D, Feinstein A. **Problems of spectrum bias in evaluating the efficacy of diagnostic tests.** *N Engl J Med* 1978;229: 926–930
14. Hillman BJ. **Outcome research and cost-effectiveness analysis for diagnostic imaging.** *Radiology* 1994;193:307–310
15. Black WC, Armstrong P. **Communicating the significance of radiologic test results: the likelihood ratio.** *AJR Am J Roentgenol* 1986;147:1313–1318