AJNR

Assessing the usefulness of diagnostic tests.

J G Jarvick, D R Haynor and W T Longstreth, Jr

AJNR Am J Neuroradiol 1996, 17 (7) 1255-1258
http://www.ajnr.org/content/17/7/1255.citation

Commentary

Assessing the Usefulness of Diagnostic Tests

Jeffrey G. Jarvik, *Assistant Professor of Radiology, University of Washington Medical Center, Seattle,*
David R. Haynor, *Associate Professor of Radiology, Veterans Affairs Medical Center, Seattle,* and
William T. Longstreth, Jr, *Professor of Neurology, Harborview Medical Center, Seattle*

Ideally, to ascertain the usefulness of a given medical treatment, investigators organize broad multicenter trials, such as the North American Symptomatic Carotid Endarterectomy Trial (NASCET) (1), in which they can compare outcomes in a sufficiently large number of patients. One could argue that a comparable method should be used for the evaluation of diagnostic tools. After all, it ultimately matters only whether a new test, like a new therapy, helps or hurts patients. But despite the difficulty and expense of the trials needed to evaluate new therapies, such trials are still more straightforward than those that would be necessary to determine whether a *diagnostic* modality is useful. The problem is that a diagnostic technology is several steps removed from patient outcome. Interposed between the making of a diagnosis and the outcome for a patient are several factors, including how the clinician uses the diagnostic information and the effectiveness of the therapy. Thus, a perfectly good diagnostic technology, if evaluated solely by patient outcomes, might look worse than it actually is because of problems "downstream." Additionally, as therapies change, the effectiveness of diagnostic techniques may need to be reassessed. One way out of this difficulty is to analyze separately and sequentially the various components that lead to patient outcome.

In 1977, Fineberg et al (2) outlined a hierarchical scheme that first consisted of four and was later revised to five levels of efficacy (Table 1). Other authors have presented similar schemes for the evaluation of diagnostic technologies (3–5). Each level of efficacy depends on the preceding level (hence the hierarchical arrangement). Thus, in order for a technology to provide useful information for diagnostic decision making (diagnostic impact) it must be an accurate test. Similarly, in order for a test to improve patient outcome, it must have a positive therapeutic impact.

Not only are radiologists the logical group to design and implement the studies that evaluate these various aspects of diagnostic technologies, to do so is in their own best interest. However, the cost of these studies cannot be borne solely by radiologists. Instead, the health care system as a whole must agree on a mechanism to fund this type of research. The Society of Magnetic Resonance recently published a report that suggested several approaches to funding, including using a cooperative group to seek support from government, industry, payers, and providers (6).

Two main options are open to the investigator evaluating the usefulness of a diagnostic technology. The first is a decision-analysis approach, in which the researcher constructs a model combining known values for the test characteristics (sensitivity and specificity) with estimates of disease prevalence and of the outcomes of treatment. While estimates of test accuracy can be extracted from the literature, there will still be uncertainty. With sensitivity analysis, a crucial aspect of decision analysis, one would substitute the range of accuracy values that could reasonably be expected from each test. If the conclusions of the model are unchanged, then the model can be regarded as insensitive to changes of the variable in question for the range tested. Unfortunately, such models are only as valid as the probability estimates from which they are constructed. Since these estimates are culled from a literature that is frequently biased and incomplete, the usefulness of such models in drawing conclusions is limited. Perhaps their most important function is to indicate the critical bits of knowledge that are

Address reprint requests to Jeffrey G. Jarvik, MD, Department of Radiology, University of Washington Medical Center, Box 357115, Seattle, WA 98195-7115.

Index terms: Commentaries; Brain, magnetic resonance; Efficacy studies; Magnetic resonance, in treatment planning

AJNR 17:1255–1258, Aug 1996 0195-6108/96/1707–1255

TABLE 1: Hierarchical scheme for technology assessment*

Level	Description
Technical capacity	Reliability and image quality
Diagnostic accuracy	Sensitivity, specificity, true-positive ratios, false-positive ratios, receiver operator characteristics
Diagnostic impact	Ability of a diagnostic test to affect diagnostic work-up
Therapeutic impact	Ability of diagnostic test to affect therapeutic choices
Patient outcome	Ability of a diagnostic test to increase length and/or quality of life

* According to Fineberg et al (2).

TABLE 2: Grading the quality of published studies*

Grade	Description
A	Broad generalizability to a variety of patients and no significant flaws in research methods: large randomized controlled trial when assessing therapeutic impact or patient outcomes
B	Narrower spectrum of generalizability than grade A studies, with only a few flaws that are well described so that their impact on conclusions can be assessed: randomized trial for therapeutic effects or patient outcomes
C	Several flaws in research methods, small sample sizes or incomplete reporting: nonrandomized comparisons for therapeutic impact or patient outcomes
D	Multiple flaws in research methods or reports of opinion unsubstantiated by data

* According to Kent and Larson (3).

lacking to build a model that validly predicts the usefulness of a given technology.

The alternative approach to modeling is primary data collection. This can take many forms, including both retrospective and prospective case-control and cohort studies, and the "gold standard" of studies, the randomized controlled trial (RCT). Time and expense dictate that RCTs be limited to a few questions that have an extremely important impact on our health care delivery system. Fortunately, however, these other methods can still produce valid and meaningful results. Moreover, primary data collection and modeling are not mutually exclusive, and can be combined to extend the range of answerable questions from a given data set.

In 1992, Kent and Larson outlined criteria for the evaluation of clinical efficacy assessment for magnetic resonance (MR) imaging (3). They proposed grading studies on the basis of the quality of the research methods (Table 2). When the article was published in 1992, there were no diagnostic impact studies that made the A grade, only 3 that made the B grade, and 6 that rated C. Forty-eight were rated D. In a later review of the clinical efficacy of MR in neuroimaging (7), almost no papers addressed therapeutic impact for stroke, carotid evaluation, intracranial hemorrhage, aneurysms, dementia, head trauma, epilepsy, or human immunodeficiency virus.

Thus, the field remains ripe for the type of study done by Hirsch et al (8) in this issue of *AJNR*. Hirsch et al focus on the intermediate outcome of diagnostic impact by measuring the clinician's estimate of the pretest and posttest probabilities for a diagnosis in order to calculate likelihood ratios. The goals of the study are well thought out and laudable in terms of assessing the usefulness of MR. However, having posed an appropriate, albeit general, question, their study has several potential problems. The following criticisms must be taken in the context that this is a pilot study, a fact that the authors acknowledge in their introduction and that likely accounts for many of the apparent shortcomings.

First, their sample lacks the size to answer definitively their very broad question. The clinical utility of MR is likely to vary with different clinical questions. As such, there may be many different answers to the question "How useful is MR?" depending on the clinical setting. The authors state that the indications for brain MR imaging were varied. Their most frequent reason for MR, a change in neurologic status, is not truly a single indication, but rather a composite of a vast array of possible symptoms and signs, affecting various parts of the nervous system.

Similar complexity applies to their third most common indication, follow-up of a previous study. The authors also observe that the presumptive clinical diagnoses, which would be another way to subdivide their cohort into more meaningful units, were as varied as the sample studied.

Kent et al (9) used 35 diseased and 35 nondiseased subjects as the minimum number needed for a high-quality study. They chose these numbers because 35 subjects is the minimum for which the lower bound of the 95% confidence interval for a true sensitivity or specificity of 1.0 would be greater than 0.9. In Hirsch et al's study, stroke and tumor were the only two diagnoses with close to 30 patients each. However, given the sizes of the effects, it is likely that, had they built confidence intervals around their likelihood ratios, they would have found important changes for many of the possible subgroups. In any case, it would have been valuable if the authors had presented analyses for at least the subsets of stroke and tumor patients.

The usefulness of any diagnostic test is likely to vary not only by patient indication and diagnosis, but also by the individual clinician. Two clinicians ordering the same test on the same patient may have quite different pretest probabilities for their presumptive diagnosis. One clinician may be a more astute observer and be able to detect a critical sign that makes a particular diagnosis highly probable, while the other clinician remains unaware of this important clue, leaving the pretest probabilities at an indistinguishable level. There may in fact be certain clinician characteristics that would enable one to predict whether a clinician would be more or less likely to have a large change in a "personal" likelihood ratio before and after the diagnostic test. For example, differences in clinical experience are likely to affect the usefulness of a test: clinicians who have been in the trenches for years rely less on diagnostic tests than do neophytes fresh out of residency, or even still in residency. Other important clinician characteristics might include specialty and subspecialty certification, training institution, and spectrum of disease in the practice. Thus, it is important in this type of study to be able to account for not only patient effects, but *physician* effects as well. One approach would have been to build a regression model, in which the dependent variable (the variable that one is trying to predict from other variables) is the difference between pretest and posttest likelihood ratio, and the independent variables include indication for the study, presumptive diagnosis, and various physician characteristics. Once again, one runs into the problems of sample size, because controlling for all these variables requires an increase in the number of patients studied.

Another problem in the study by Hirsch et al is that the authors rely on self-assessments of test utility. This raises the question of bias, in that an individual may be more likely to rate a test as useful if he or she ordered it. An alternative would be to look at how diagnostic and therapeutic preferences differed before and after a diagnostic test. If there was a large change in the differential diagnosis, or if the choice of therapy was altered, then one could be more confident that the test had an impact than if one simply relies on the opinion of the person who ordered the test.

Often a diagnostic test is ordered not with the expectation that it will confirm a high-probability diagnosis, but rather that it will rule out the presence of an important low-probability diagnosis. This aspect of the utility of a diagnostic test not only is important for the ordering clinician, who may be seeking reassurance that a cancer is not causing the patient's problem, but is also reassuring for the patient. The reassurance to the patient can have a direct influence on patient outcome and is embodied in the concept of the therapeutic value of a diagnostic test. Harold Sox and coauthors (10) performed an experiment in the 1970s in which patients with benign-appearing chest pain were randomly assigned to receive or not receive electrocardiograms. Patients undergoing diagnostic testing actually did better than their counterparts not receiving the tests. This is an important value of diagnostic testing that is frequently overlooked. While Hirsch et al conclude that MR imaging is useful, they are potentially not evaluating an important contribution to patient care.

One particular point that Kent and Larson (3) make is that when measuring diagnostic uncertainty, an independent reference standard is necessary to know whether reductions in uncertainty are attributable to misinformation, such as improved certainty based on false-positive findings (3). Hirsch et al make no attempt to address this issue.

The jazz musician Les McCann has a song called "Compared to What?" This article begs the same question. The implicit comparison in this study is of MR imaging to clinical examination. But this study, like so many other studies before it, lacks a control group. We do not know what the natural course of the decision-making process would be without MR imaging. Simply waiting a few days might increase the diagnostic confidence as either a disease process declares itself or a resident is able to read up on a rare constellation of signs and symptoms. Another question is how MR imaging compares with computed tomography, a less costly but, in many circumstances, nearly as accurate alternative.

Despite the potential shortcomings of their study, Hirsch et al are to be commended and encouraged for attempting to evaluate aspects of diagnostic imaging mostly ignored until now: the outcomes of diagnostic impact and therapeutic impact. This type of research is difficult because it requires clinicians to do two things. First, they must spend time thinking in a way they might not be used to about why they ordered a diagnostic study. And second, they must complete forms, usually after being hounded by a dedicated researcher. Nonetheless, the time invested for these types of studies is likely to return extensive dividends in practical knowledge about the usefulness of imaging. In short, Hirsch et al make a worthwhile contribution, first by asking difficult questions and then by helping to develop and refine the tools to answer them.

Acknowledgment

Many thanks to Peter M. Jucovy, MD, for his review of the manuscript.

References

1. North American Symptomatic Carotid Endarterectomy Trial Collaborators (NASCET). Beneficial effect of carotid endarterectomy in symptomatic patients with high-grade stenosis. *N Eng J Med* 1991;325:445–453
2. Fineberg HV, Bauman R, Sosman M. Computerized cranial tomography: effect on diagnostic and therapeutic plans. *JAMA* 1977;238:224–227
3. Kent D, Larson E. Disease, level of impact, and quality of research methods: three dimensions of clinical efficacy assessment applied to magnetic resonance imaging. *Invest Radiol* 1992;27:245–254
4. Fryback D, Thornbury J. The efficacy of diagnostic imaging. *Med Decis Making* 1991;11:88–94
5. Schwartz J. Evaluating diagnostic tests: what is done–what needs to be done. *J Gen Intern Med* 1986;1:266–267
6. Harms S, Radensky P, Sunshine J, et al. MRI efficacy and effectiveness research: who needs it and who pays for it? *J Magn Reson Imaging* 1996;1:4–6
7. Kent D, Haynor D, Longstreth W, Larson E. The clinical efficacy of magnetic resonance imaging in neuroimaging. *Ann Intern Med* 1994;120:856–871
8. Hirsch JA, Langlotz CP, Lee J, Tanio CP, Grossman RI, Schulman KA. Clinical assessment of MR of the brain in nonsurgical inpatients. *AJNR Am J Neuroradiol* 1996;17:1245–1253
9. Kent D, Haynor D, Larson E, Deyo R. Diagnosis of lumbar spinal stenosis in adults: a metaanalysis of the accuracy of CT, MR, and myelography. *AJR Am J Roentgenol* 1992;158:1135–1144
10. Sox HJ, Margulies I, Sox C. Psychologically mediated effects of diagnostic tests. *Ann Intern Med* 1981;95:680–685