Randomized Controlled Trials Are Not Appropriate for Imaging Technology Evaluation

The randomized controlled trial (RCT) is the only means of evaluating the effectiveness of a new therapeutic modality that can be expected to provide a definitive conclusion. Because 1 patient cannot be treated by 2 different methods at the same time, 2 patient populations are required for comparison of a new treatment and reference method. Randomization is used to avoid bias in the allocation of patients for treatment by 1 method or the other, and appropriately large patient numbers are accrued to eliminate the effects of population differences that occur randomly. The number of patients required depends on the expected magnitude of the difference between the 2 methods under investigation, compared with the magnitude of random differences.

Determining the accuracy of a noninvasive diagnostic technology does not present the same problems, because 1 patient can undergo 2 different tests, thereby eliminating problems that arise from comparing 2 populations. When a single study population is used, there can be no difference in severity of disease that the 2 technologies are used to evaluate, and both technologies are necessarily compared with the same reference standard. By this means, the number of patients that is needed to reach a valid conclusion is greatly reduced. The difference in diagnostic results between the 2 tests can then be used to determine the difference in patient treatment that this would have produced and to assess the effect on treatment cost of using the new modality.

Received May 25, 1999; revision accepted Nov. 10, 1999.

For correspondence or reprints contact: Peter E. Valk, MD, Northern California PET Imaging Center, 3195 Folsom Blvd., Sacramento, CA 95816.

Use of the RCT to compare the accuracy of 2 diagnostic modalities is inappropriate, because the RCT has no advantage in terms of validity and is more difficult and expensive to perform. On the other hand, to measure directly the effect of a diagnostic technology on patient survival, an RCT does become necessary, because survival must be evaluated separately for each technology under evaluation. Unfortunately, such trials are frequently not possible in practice. In cancer management, for example, it is rarely possible to initiate an RCT in which the only difference between the 2 arms of the protocol is a single diagnostic test. Even when this can be done, the independent contribution of imaging to survival is likely to be small, thereby requiring very large sample sizes.

Even if a sufficiently large RCT of 2 diagnostic imaging modalities were performed, unbiased results would be difficult to achieve. Double-blind evaluations of imaging modalities are not really possible, because the imaging modality used is apparent from the images produced. Also, the effect of a therapeutic procedure in a blinded study is independent of the physician, but this is not the case for an imaging modality in which the management impact is dependent on the physician's diagnostic thinking. Can one expect a clinician to use the result of a new and unfamiliar test in exactly the same way as the result of an established test with which he or she is fully familiar? If the technology cannot be disguised, physician bias will be difficult to avoid.

Unfortunately, these differences in evaluation of diagnostic and therapeutic modalities have frequently been lost in the general enthusiasm for the RCT. Individual experts in technology evaluation, technology assessment organizations, and national government entities have all referred to the unique validity of the RCT, often regarding with suspicion results that have been obtained through a different approach in experimental design.

Most often quoted in support of such views is an article by Fryback and Thornbury (*1*), in which the authors presented a conceptual model for efficacy assessment of diagnostic imaging that included 6 levels of efficacy, with the highest being the determination of patient outcome by means of the RCT. Fryback and Thornsbury also concluded that such RCTs were rarely possible and described effective alternatives, but these observations appear to have been ignored by many of their readers. The authors made no reference to the head-to-head comparison of technologies in a single population in their general discussion of these issues (*1*), but a subsequent article on increasing the scientific quality of efficacy studies was based on a head-to-head comparison of MRI and CT (*2*).

The early work of Fineberg (*3*) and the more recent work of Fryback and Thornbury (*1*) and Thornbury et al. (*2*) were based on a concept of a continuum of management efficacy, whereby imaging was embedded in a global clinical process. This process included the steps of diagnostic accuracy efficacy, therapeutic efficacy, and patient outcome efficacy, all of which were regarded as being descriptive of the imaging procedure. However, one may alternately view diagnostic efficacy, expressed as sensitivity and specificity, as a characteristic of the imaging procedure and view therapeutic and patient outcome efficacy as characteristics of the clinical situation. The question may be divided into 2 parts: How good is the test at making the diagnosis, and how important is the diagnosis to patient outcome? Rather than attempting to answer both questions with an RCT, one may determine sensitivity

and specificity in a head-to-head comparison and separately address the issue of whether the diagnosis will have an impact on management and survival by means of decision analysis. This is, in effect, the approach that was proposed by Thornbury et al. (*2*) in their discussion of the scientific quality of efficacy studies.

When all study patients undergo both imaging procedures, it becomes important to avoid interpretation bias by ensuring that both sets of images are read independently, without knowledge of the result of the other test. Ideally, the order in which the tests are performed should be randomized, to avoid any possibility of bias at the initial reading. For this reading, all relevant clinical information should be available to the reader, to permit evaluation of how the modalities perform in the actual clinical situation. A second, fully blinded reading in a controlled environment could then be used to evaluate the tests alone, without clinical information, and to assess intra- and interobserver variability (*4*).

Another problem that is encountered in diagnostic technology evaluation is failure to focus on the clinical purpose of the test that is being evaluated, so that making a management decision becomes lost in a statistical exercise. For example, it has been pointed out that the greatest statistical power in the determination of sensitivity and specificity is achieved by a study population in which the prevalence of disease is 50% (*5,6*). For maximum statistical validity, the study population should also include a full range of disease severity, both treated and nontreated patients and patients with commonly confused disorders (*5,6*). However, it is likely that this statistically ideal population will never be encountered in clinical practice; therefore, the clinical use of this exercise is not always clear. What is needed for the purposes of patient management is an assessment of diagnostic accuracy for specific indications, in which the prevalence and severity of disease are determined by the patients' usual clinical presentation.

A review of published data in oncologic PET that was conducted by a technology evaluation organization on behalf of a government agency provided examples of these problems (*6*). This review confused diagnostic and therapeutic evaluations by "grading" evidence from each study on the basis of criteria that had been developed for evaluation of treatment efficacy (*7*). These criteria, which formed the basis of the review, were focused on showing "statistically significant treatment effect" and were inappropriate for the purpose of evaluating a diagnostic technology. The review also criticized investigators for selecting patients according to actual clinical indications, because this approach failed to produce statistically oriented study populations.

Technology evaluation has been a recognized field of study since the 1970s but has had little impact on clinical technology use. Since 1980, both CT and MRI have become incorporated in standard medical practice without undergoing meaningful evaluation. In part, this may have resulted from the impracticality of focusing on the RCT as the principal evaluation tool, because investigators did not have access to the resources that would have been required for evaluation of a useful number of clinical indications by this means. Onerous requirements were defined for the RCT, and other, more practical methodologies were criticized for not meeting these requirements.

It might be more productive in the future to focus on studies that directly compare new and reference diagnostic technologies in the same patient population, which is selected for study on the basis of usual clinical indications. Such studies would require far less time and other resources than RCTs and could provide the guidance that physicians and insurers need in making utilization decisions. As a general tool of diagnostic technology evaluation, the RCT is inappropriate for the purpose and is too difficult, expensive, and time consuming to perform; it is not cost effective for evaluation of imaging technology.

Peter E. Valk
Northern California PET Imaging Center
Sacramento, California
University of California, Los Angeles,
School of Medicine
Los Angeles, California

REFERENCES

1. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. *Med Decis Making.* 1991;11:88–94.
2. Thornbury JR, Kido DK, Mushlin AI, Phelps CE, Mooney C, Fryback DG. Increasing the scientific quality of clinical efficacy studies of magnetic resonance imaging. *Invest Radiol.* 1991;26:829–833.
3. Fineberg HV. Evaluation of computed tomography: achievement and challenge. *AJR.* 1978;131:1–4.
4. Begg CB, McNeil BJ. Assessment of radiologic tests: control of bias and other design considerations. *Radiology.* 1988;167:565–569.
5. Department of Clinical Epidemiology and Biostatistics, McMaster University Health Sciences Center. How to read journals: II. To learn about a diagnostic test. *Can Med Assoc J.* 1981;124:703–710.
6. Management Decision and Research Center Technology Assessment Program. *Positron Emission Tomography: Descriptive Analysis of Experience with PET in VA and Systemic Reviews: FDG PET as a Diagnostic Test for Cancer and Alzheimer's Disease.* Washington, DC: Veterans Health Administration, Department of Veterans Affairs; 1996.
7. Cook DK, Guyatt GH, Laupacis A, Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. *Chest.* 1991; 102(suppl):305S–311S.