A framework for clinical evaluation of diagnostic technologies

Current Review

Gordon H. Guyatt,* MD
Peter X. Tugwell, MD
David H. Feeny, PhD
R. Brian Haynes, MD, PhD
Michael Drummond, DPhil

Most new diagnostic technologies have not been adequately assessed to determine whether their application improves health. Comprehensive evaluation of diagnostic technologies includes establishing technologic capability and determining the range of possible uses, diagnostic accuracy, impact on the health care provider, therapeutic impact and impact on patient outcome. Guidelines to determine whether each of these criteria have been met adequately are presented. Diagnostic technologies should be disseminated only if they are less expensive, produce fewer untoward effects and are at least as accurate as existing methods, if they eliminate the need for other investigations without loss of accuracy, or if they lead to institution of effective therapy. Establishing patient benefit often requires a randomized controlled trial in which patients receive the new test or an alternative diagnostic strategy. Other study designs are logistically less difficult but may not provide accurate assessment of benefit. Rigorous assessment of diagnostic technologies is needed for efficient use of health care resources.

Dans la plupart des cas, les nouvelles techniques de diagnostic n'ont pas été scrutées d'assez près afin de savoir si elles sont susceptibles d'améliorer la santé. Il faut les examiner complètement afin d'en déterminer la validité technologique, les applications possibles, la précision diagnostique et l'incidence sur les personnes et les organismes chargés des soins sanitaires, sur les traitements et sur le devenir des malades. On présente ici des principes devant servir à découvrir s'il est satisfait à chacun de ces critères. Une technique de diagnostic ne doit être rendue disponible que si elle est moins chère et moins nuisible, tout en étant au moins aussi précise, que les techniques déjà en place, si elle rend inutile le recours à d'autres explorations sans perte de précision, ou si elle permet la mise en route d'un traitement efficace. La preuve d'un avantage pour les malades nécessite souvent un essai comparatif, chez des sujets choisis au hasard, de la nouvelle méthode et d'une autre approche diagnostique. Si on adopte une forme d'essai moins difficile à réaliser, on risque de se tromper dans l'estimation des avantages. L'examen rigoureux de la validité de toutes les méthodes de diagnostic est indispensable à l'utilisation efficace des ressources sanitaires.

The recent introduction of sophisticated and expensive diagnostic technology into medical practice has given rise to important questions. First, do these technologies actually improve patient care? Second, if they do improve care, is their price — in terms of services to other patients forgone — in keeping with their benefit? Unfortunately, many technologies have been disseminated without adequate evaluation, and we now face a monumental backlog of technologies that need to be assessed plus a burgeoning stream of even newer machines. To make matters worse, while methodologic standards for determining the accuracy of diagnostic tests are well established,[1-3] the criteria that should be met before a new test is introduced into routine clinical practice remain controversial. We believe the standards for evaluating new techniques have not been sufficiently

From the departments of Medicine, Economics, and Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ont.

*Career scientist, Ontario Ministry of Health

Reprint requests to: Dr. Gordon H. Guyatt, Department of Clinical Epidemiology and Biostatistics, McMaster University Health Sciences Centre, Rm. 3H7, 1200 Main St. W, Hamilton, Ont. L8N 3Z5

← For prescribing information see page 661

rigorous and that inadequate evaluation has contributed to overutilization of diagnostic techniques. In this paper we present a framework and a set of practical and scientific standards for the clinical assessment of diagnostic technologies. Our framework is perhaps an idealized outline of how diagnostic technologies should be assessed; however, adherence to its principles would substantially decrease the problems we have described.

Health-care-related technology can be broadly defined as "the set of techniques, drugs, equipment and procedures used by health care professionals in delivering medical care to individuals, and the system within which such care is delivered".[4] According to this definition, simple tests, as well as sophisticated machines, are technologies. While our discussion will focus on new and expensive techniques, it is equally applicable to diagnostic tests in general.

The framework

In addition to their clinical uses, diagnostic technologies may contribute to a better understanding of human physiology and mechanisms of disease. Positron emission tomography scanning is likely to provide important information about tissue metabolism in health and disease, irrespective of its diagnostic use. While elucidation of disease mechanisms constitutes an important reason for developing and using diagnostic technologies, this goal is largely independent of clinical considerations and by itself would lead to much more limited dissemination. For these reasons we shall limit our discussion to the accuracy and benefit of technologies in the clinical context.

Depending on the point of view, there are a number of criteria for concluding that a diagnostic technology is ready for dissemination.[5,6] These criteria can be considered to form a hierarchy of progressively more rigorous evaluation, as follows:

- Technologic capability: The ability of the technology to perform to specifications in a laboratory setting has been demonstrated.
- Range of possible uses: The technology promises to provide important diagnostic information in a number of clinical situations.
- Diagnostic accuracy: The technology provides information that allows health care workers to make a more accurate assessment regarding the presence and severity of disease.
- Impact on health care providers: The technology allows health care workers to be more confident of their diagnoses and thereby decreases their anxiety and increases their comfort.
- Therapeutic impact: The therapeutic decisions made by health care providers are altered as a result of application of the technology.
- Patient outcome: Application of the technology results in benefit to the patient.

This schema has been modified from the one suggested by Fineberg and colleagues[5] primarily through the addition of the second and third criteria. Fineberg and colleagues did not attempt to provide guidelines whereby one could determine whether these criteria have been adequately met, nor did they discuss how far this hierarchy needs to go before a technology should be considered appropriate for dissemination. The latter will depend on a number of factors, including present practice, the cost and untoward effects of the technology and one's values. These issues will be considered in detail.

Technologic capability

Establishing the technologic capability of a new test is generally undertaken by physicists, biochemists, physiologists and manufacturers. Our discussion will concern the clinical evaluation of the new technology, which begins when the basic scientists and manufacturers have produced a product that meets laboratory specifications.

Range of possible uses

When the technology comes out of the laboratory into the clinical setting, it must first be applied to many patients with a large number of diverse conditions. The goal of this exercise is not to establish accuracy but rather to delineate the possible uses of the technology. Striking impressions often result from this exploratory phase: for example, the ease with which computed tomography (CT) scanning identifies hemorrhage, the difficulty in delineating lesions in the first weeks following thrombotic stroke, and the usefulness of magnetic resonance (MR) imaging in demonstrating demyelinating lesions.

The major criteria for patient selection during this phase should be a good idea of what the underlying condition is and a reasonable expectation that the new technology will provide important information. Attempts to limit the spectrum of conditions may prevent a full appreciation of the technology's uses. Those interpreting the test results should not be blind to who the patients are or what underlying condition they have. In fact, the more clinical information the better, for this phase of instrument development is a learning process in which unexpected correlations are discovered, interpretations refined and important hypotheses for subsequent testing generated. Problems can arise if certain elements of more advanced studies (such as establishing reliability, carefully defining the patient population or blinding the interpreters of the test results) are incorporated in what is still essentially an exploratory study. This may have two deleterious consequences: the full potential of the early phase of evaluation may not be realized, and a spurious impression of the adequacy of studies in establishing the usefulness of the test may be created.

Diagnostic accuracy

For a diagnostic technology to be clinically useful it must be able to accurately distinguish diseased and nondiseased and to quantitate the severity of an illness or condition. The pitfalls in trying to determine the accuracy of diagnostic tests and the best ways around them have been thoroughly studied and described.[1,2] We will briefly summarize them here.

Establishing accuracy implies independent comparison with a "gold standard". By independent, we mean that those interpreting the test results must be unaware of the results of the gold-standard procedure. The gold standard is often a more invasive or dangerous procedure, such as coronary artery angiography, which is a gold standard for electrocardiographic stress testing. If a definitive procedure is not available a substitute gold standard, such as long-term follow-up, may be adequate. In determining accuracy the precision should also be ascertained; both intra- and interobserver reliability should be established.

There are many situations in which no gold standard exists and adequate substitutes are not available (e.g., bronchial provocation tests for asthma, walking tests for functional exercise capacity in patients with chronic heart and lung disease, and strain-gauge plethysmography for diagnosis of the postphlebitic syndrome). In these situations one must rely on construct validity. To demonstrate construct validity one examines the relation between a new test and existing measures and looks at whether the new technology relates to other variables in the way one would expect if it is really measuring what it is supposed to measure.

When a gold standard is available and a new test is designed to detect the presence or absence of disease, sensitivity (the proportion of patients with disease correctly identified as such) and specificity (the proportion of patients without disease correctly identified) should be calculated. A receiver-operating characteristic (ROC) curve, which relates false-positive results to true-positive results at multiple cut-off points, can be constructed to help determine the cut-off point that gives the optimum combination of sensitivity and specificity. A more powerful method of establishing a test's usefulness is to examine the associated likelihood ratios, which allow estimates of the probability that disease is present at any level of a diagnostic test result.[7] If two tests are being compared their results should be interpreted independently against the same gold standard. In this situation ROC curves are used to determine which test is "better",[8] but comparison of likelihood ratios at various levels of test results is a more powerful method.

For diagnostic technologies in which level of function or severity of disease are the variables of interest, simple correlation coefficients (if the unit of measurement differs between the test and the gold standard) or an intraclass correlation coefficient that takes account of both random and systematic error[9] can be used to quantitate the relation between the test and the gold standard. An alternative approach is to divide the outcomes into clinically relevant levels and calculate chance-corrected agreement with a statistic called kappa,[10] or a modified version, called weighted kappa, which allows one to consider not only the fact that the technologies disagreed with the gold standard but also the extent of the discrepancy.[11]

In assessing accuracy we need to know whether the diagnostic technology is capable of identifying patients with mild as well as severe disease and of distinguishing them not from healthy individuals but from those with conditions easily confused with the disease of interest. Accuracy should be determined by examining representative patients with a suspected condition, applying the diagnostic technology under investigation and proceeding with independent application of the gold standard. Sensitivity will be greater in those with severe disease, and specificity will be greater in normal controls. Exercise radionuclide ventriculography as a diagnostic test for coronary artery disease provides a dramatic example.[12] When initially studied radionuclide ventriculography had a specificity of 94%. In subsequent studies the specificity fell to 49%; this was most probably due to the fact that the disease-free population was far healthier during the earlier studies. Patients who have a substantial pretest likelihood of coronary artery disease (i.e., the ones for whom we need the diagnosis confirmed) show a high incidence of nonspecific abnormalities in exercise radionuclide angiography.[12]

Impact on health care providers

Accurate diagnostic tests may influence neither therapy nor patient outcome, and yet many still receive wholehearted support from the medical community. CT scanning has been widely disseminated, and its use for many conditions has received endorsement from an impressive array of official agencies and consensus conferences without rigorous scientific evidence that patients benefit from its application. This may be because physicians and policy-makers are convinced of its benefit even in the absence of adequate data, because they see the demonstration of accuracy as sufficient reason for dissemination irrespective of benefit, or because the CT scan's ability to reduce the physician's anxiety and increase his or her confidence may favour rapid diffusion and official endorsement. Although one may think of more sinister possibilities (such as the increased power and status, and the financial advantages, acquired by the medical profession as it adopts more and more mysterious and apparently powerful gadgetry, or the unregulated promotional efforts of the companies responsible for developing the new technology) we suspect that the third explanation may be the most important. There is no doubt that it is immensely reassuring to know that a patient who has had a sudden change in neurologic status is not suffering from a condition (such as acute or chronic subdural hematoma) that requires immediate surgical intervention. Similar reassurance accompanies the knowledge that severe headaches of recent onset do not represent a brain tumour or that the comatose patient with a documented malignant disease in whom one has decided to do nothing more does, in fact, have tumour deposits throughout the brain. Such reassurance is even more powerful in these days of rampant medical litigation, when a mistake may have disastrous consequences for not only the patient but also the physician.

Assuming that diagnostic technologies do have an impact on health care workers, what weight should we give this effect in our decisions about resource allocation? For example, if an expensive technology reassures health care providers but does not influence patient outcome should it be adopted? Certainly a test's effects on health care workers should influence our decisions about its use, but careful thought needs to be given to the appropriate measurement of the effect as well as to its importance.

Reasonable discussion of this so far neglected area cannot begin until data about the extent to which diagnostic technologies do provide reassurance to health care workers are collected. Diamond and Forrester[13] have highlighted the distinction between estimates of the probability that a disease is present and confidence in that estimate. That is, a given estimate of the probability that a disease is present (e.g., it is believed that there is a 25% chance a patient has coronary artery disease) may be associated with a great deal of confidence (e.g., it is quite certain that the probability is close to 25%) or very little confidence (e.g., the estimate of 25% is just a guess; the real probability could be much higher or be close to zero). Confidence in probability estimates is likely to be related to physician expertise and experience and is likely to increase as data about patients accumulate. This issue should be explored further in studies of diagnostic technologies.

Therapeutic impact

A test result may have diagnostic impact and still not affect therapy: a health care worker may be unaware of the significance of a test result or unfamiliar with available treatment, the change in probability of disease may be insufficient to alter therapy, the patient may refuse treatment, there may be no therapy available, or the patient may already be receiving the best possible therapy. To change morbidity or mortality or improve the quality of life a diagnostic test must provide information that changes therapy. If the test results lead to institution of an intervention whose effectiveness is known, patient benefit follows. If unproven therapy is instituted a change in health status as a result of the diagnostic technology remains a possibility.

How would one go about showing that therapy has changed as a result of a new diagnostic technology? The best way would be a randomized controlled trial in which patients would be randomly assigned to one of two diagnostic plans, only one of which would include the technology under investigation. The new technology might be added (e.g., exercise stress testing might be instituted before a patient who has been treated for myocardial infarction leaves the hospital), or the two arms of the experiment might contain alternative technologies (e.g., CT scanning and MR imaging).

It has been argued that clinical trials are likely to be too cumbersome or impractical for regular evaluation of diagnostic technologies.[14] Problems include the need for a large number of patients, the need for preliminary use of the technology in practice for clinicians to develop expertise in interpretation of the results, and rapid developments in technology, which may make the results of a trial obsolete by the time they appear.[14,15] Given the difficulties of randomized controlled trials are there alternative ways of assessing a test's therapeutic impact?

One strategy is simply to review patient records and evaluate whether the diagnostic test altered patient management.[16] Retrospective review has numerous problems, including the difficulty of determining what would have been done if the test had not been available. A more effective method would be to ask physicians about their plans for further diagnosis and therapy before the test is performed, then give them the results and see if their plans change.

Using before–after studies based on clinicians' reports of their plans for therapy is enticing. First, the expense and logistic difficulties of a randomized controlled trial are avoided. Second, no patient is denied a potentially beneficial technology.

There are, however, major problems with this study design:[17,18] changes in therapy that are believed to be beneficial may, in fact, be harmful; inaccurate diagnostic tests can have deleterious therapeutic impact; clinicians differ systematically in their assessment of whether a given test result contributed to management;[19,20] it may be difficult to consistently be aware of clinicians' plans before the test results are available; clinicians' reports of what they would do before the test result is available may differ from what they actually would have done were the technology not available; while all patients receive the potential benefits of the test, they also are all exposed to its known and unknown hazards; and the design is in most cases limited to "add-on" technologies as opposed to those that replace existing tests.

It may be argued that these problems do not significantly mar the validity of before–after study designs of therapeutic impact that rely on physician judgement. We believe it is more likely that such studies will overestimate patient benefit. The only way to know for sure is to do what has been done for uncontrolled trials or for those using historical controls: compare their results to those of randomized controlled trials that ask the same question. Preliminary evidence comes from two before-after studies that found that CT scanning can decrease the frequency of abdominal surgery.[16,21] In the only randomized controlled trial of CT scanning patients presenting with undiagnosed abdominal masses were randomly assigned to receive CT or conventional imaging.[22] The proportion of patients who received laparotomy was actually higher in the group that underwent CT scanning (39%) than in the control group (32%). Sample size limitations make the results of this study far from definitive, but the results suggest that further comparisons between before-after studies of therapeutic impact and randomized controlled trials that examine patient benefit be conducted before the former are accepted as valid in the assessment of diagnostic technologies.

However, there are circumstances in which one can be more confident of the validity of before-after therapeutic impact studies. If change in therapy immediately follows receipt of new diagnostic information, or if the test is clearly responsible for an important change in treatment plan, the therapeutic impact of the technology is established. If changes in management have been shown to be effective in well designed randomized controlled trials or obviate the need for an invasive procedure no further studies will be required.

More often, though, therapeutic impact studies that rely on clinical judgement will have a role as exploratory studies. If no therapeutic impact is found it is extremely unlikely that the technology is of benefit. On the other hand, if the initial study results suggest therapeutic impact more rigorous investigations must be undertaken.

Patient outcome

Does one really need to go beyond determining therapeutic impact or even diagnostic impact before concluding that a technology warrants dissemination? There have been many instances in which a diagnostic technology provided information but failed to change clinically relevant outcomes. In one case, application of a diagnostic test (measurement of serum cholesterol levels) when followed by specific therapy (clofibrate administration), actually increased the rate of death.[23] Emergency endoscopy in patients bleeding from the upper gastrointestinal tract provides increased diagnostic information without altering rates of surgery, length of hospital stay or rate of death.[24] Chest radiography is an accurate tool in ascertaining the presence of carcinoma of the lung, and radiographic screening for lung cancer has therapeutic impact, in that more patients undergo surgery and at an earlier stage of disease, but evidence to date suggests that the outcome does not change.[25,26] Nonstress tests that monitor fetal heart rate for abnormalities clearly add information and have been widely disseminated, but they do not change perinatal morbidity or mortality.[27,28] These examples illustrate the wisdom of demonstrating improvement in patient outcome before a diagnostic technology becomes widely disseminated.

It has been suggested that randomized controlled trials are extremely difficult and may not be feasible for many diagnostic technologies. A key feature of such trials of therapeutic technology — blinding — is difficult in trials of diagnostic technologies in which the physician may have to be aware of the diagnostic procedure. This limitation introduces the possibility of bias in the application of other tests or the institution of a treatment regimen. Feinstein[29] has recently argued that cohort studies of patients who have been given a test according to clinical judgement or availability may provide valid results if important confounders are considered. One problem with this approach is that it is unlikely that all important confounders can be identified and adequately measured. For example, a group of Australian neurologists examined the effect of CT scanning on mortality in stroke patients by comparing the outcome of patients seen in 1978 who underwent CT scannning with that of patients seen in 1974 before CT scanning was introduced.[30] To ensure that it was the scanning that was making the difference, they chose patients matched for all the prognostic variables they thought relevant. The 1978 group had a lower mortality, apparently providing dramatic evidence of the impact of CT scanning on outcome. However, the investigators then assessed mortality in another matched group: stroke patients seen in 1978 when the CT scanner wasn't working. This group had the lowest mortality, comparable to that of the other patients studied in 1978 and lower than that of the group studied in 1974. The conclusions are that patients in the more recent group were not as sick (in ways that the investigators could not measure except by looking at rates of death) as the historical control group and that concurrent randomized controls are necessary to establish the benefit of diagnostic technologies.

While there is no doubt that randomized controlled trials of diagnostic technologies are difficult, are certainly not applicable to all situations and are limited by the difficulty associated with blinding, they are nevertheless possible. Trials of diagnostic technologies conducted to date include studies of nonstress tests in pregnant women,[27,28] CT scanning in the assessment of abdominal masses,[22] endoscopy in patients with acute gastrointestinal bleeding,[24] chest radiography in men at risk of carcinoma of the lung,[25] endoscopic cholangiography versus transhepatic cholangiography in patients with jaundice,[31] and multiphasic screening at the time of admission to hospital.[32] Clearly, there are many methodologic challenges still to be met, but the same was true of randomized controlled trials of therapeutic technologies two decades ago.

Methodologic standards for studies of patient outcome

The methodologic standards for trials of therapeutic technologies are equally applicable to diagnostic tests. They include the necessity for true randomization, pre- or posthoc stratification for potential confounders, consideration of possible cointervention and contamination, adequate sample size, and measures (such as blinding) to minimize potential bias.[33] However, several points need to be made about the special challenges posed by randomized controlled trials of diagnostic technologies.

Care must be taken to identify the appropriate role of the new technology. If it is added on to existing technologies all patients must receive the full conventional examination and then be randomly assigned to receive or not receive the new procedure. If the new test is designed to replace existing methods patients must receive an identical examination prior to randomization to receive or not receive the conventional or experimental technology.

Diagnostic possibilities (and diagnostic confidence) and therapeutic plans should be elicited from health care providers before and after application of the technology, for this will help clarify the mechanism of any effects that are found. This is analogous to measuring the biologic effects, in addition to the clinically relevant outcomes, of a new treatment in an attempt to clarify mechanisms of action.

Patient selection for studies of diagnostic technology must be appropriate. For example, there will be patients in whom the pretest likelihood of disease is so low that even a positive result on the new test will not lead to institution of therapy (or so high that therapy would be administered despite a negative result). The inclusion of patients in whom the test can have an impact on neither therapy nor outcome in a randomized controlled trial of a new technology will decrease the power of the study. The analogy here is restricting entry to a therapeutic trial to "high-risk, high-response" patients.

Mechanisms for estimating the accuracy of the technology should be built into the trial. After all, one wouldn't expect benefit from an inaccurate test. This is also an argument for proceeding directly from the preliminary stage of establishing the range of possible uses to randomized controlled trials in which patient benefit is the primary measure of outcome. If this approach is taken the hierarchy of diagnostic accuracy, impact on health care providers, therapeutic impact and patient outcome can all be examined in a single study.

All clinically relevant aspects of patient outcome should be measured. These may include reductions in rate of death, length of hospital stay and number of complications from more invasive tests, as well as improvement in quality of life. By quality of life we mean both a person's ability to undertake activities that he or she finds rewarding and enjoyable, and the way he or she feels. A diagnostic technology may change quality of life even when other, more easily measurable, variables show no change. Although the assessment of quality of life is an intimidating task, guidelines for its measurement are becoming more available.[34,35]

An example of the importance of measuring the impact of an intervention on quality of life is provided by Sox and associates,[36] who found that among patients presenting with noncardiac chest pain, those randomly assigned to receive routine measurements of creatine phosphokinase and electrocardiography showed less short-term disability than did patients who did not undergo these investigations. Just as the physician may find that negative results of CT scanning decrease his or her anxiety about the possibility of a brain tumour in patients with severe headaches of recent onset, so may the patients. This reassurance can be extremely important for the worried patient.

The therapeutic value of diagnostic tests, as this reassurance value might be labelled, is worthy of considerably more investigation than it has received to date, but we would like to include one caveat for those who might try. The reassurance value of the test for the patient may be confounded with its reassurance value for the physician, and unless the latter is measured the extent of this confounding may be impossible to assess. For example, patients with a severe headache who undergo CT scanning may have a stronger conviction that they don't have a brain tumour than patients who are spared the test only because the physician expressed some hesitation about the matter. The appropriate conclusion from such a study would be not to recommend CT scanning for patients with headache but to educate physicians so they realize that if the results of a careful neurologic examination are negative a brain tumour can be virtually ruled out.[37,38] Physicians might then be able to provide the reassurance that would quell the patient's anxieties.

Blinding of patients in studies of diagnostic technologies (such as by mock CT scanning) may be ethically questionable. Although in some situations blinding of physicians may be possible (such as when a physician receives a verbal report of a test without being aware of which test or combination of tests led to the result), it will often be difficult or impossible. However, it will usually be possible and highly desirable that those who are assessing the outcome (such as interviewers adminstering questionnaires on quality of life) be blinded.

We have discussed the difficulties of assessing the accuracy of tests for which a gold standard does not exist and the need to resort to construct validity: the determination of whether a technology relates to other measures in the manner that one would expect if it is really measuring what it is supposed to measure. Another, possibly more satisfactory, approach is to consider construct validity only in passing and to proceed straight to determining whether the application of the technology results in patient benefit. For example, one could randomly assign patients suspected of having asthma to receive or not receive a bronchial provocation test. If it is found that asthma is better controlled in patients in the experimental group, then there is very strong evidence that the bronchial provocation test is a good measure of the severity of asthma.

Given that new technologies are often expensive, rigorous economic evaluation should be built into randomized controlled trials of diagnostic tests. Methodologic standards of economic evaluation in clinical trials are available.[39-41] The inclusion of economic evaluation underscores the importance of assessing the impact of a diagnostic technology on patient outcome. If the clinical evaluation yields rigorous evidence concerning benefit the incremental cost of using the test can be related to the incremental improvement in patient outcome. The relation between incremental cost and benefit can in turn be compared with the costs and benefits associated with other health care interventions.[17,39-41] If the cost per life-year or quality-adjusted life-year gained is lower than that associated with many alternative uses of health care resources, it is likely that the new technology represents an efficient use of health care resources.

When are randomized controlled trials of patient outcome unnecessary?

There are a number of situations in which the most stringent tests of benefit are not appropriate, as follows:

- Patient benefit from the test is so dramatic that even the results of observational studies leave no room for doubt. The use of electrocardiography for dysrhythmias associated with treatment of known efficacy is one example. The use of CT head scanning in the context of head trauma is sufficiently dramatic in decreasing the need for exploratory surgery that it can probably also be included in this category.[42,43]
- The new technology produces the same or fewer untoward effects *and* is equally or less expensive than existing alternatives and has been shown to be more accurate.
- If controlled trials demonstrate that application of a diagnostic technology leads to the institution of a therapy that previous randomized controlled trials have proven effective or to the termination of harmful therapy (as might happen when a patient without a disease is mistakenly given a toxic treatment), benefit can be considered established. For example, impedance plethysmography and leg scanning have been shown to be comparable to venography in diagnosing deep vein thrombosis.[44] Because heparin is known to do more good than harm in treating deep vein thrombosis, that impedance plethysmography leads to appropriate administration of heparin is a sufficient demonstration of its usefulness. Given the difficulty in performing studies that examine differences in outcome, it may be worth while to concentrate diagnostic technology assessment in areas where treatment is known to do more good than harm on the basis of existing results of randomized controlled trials or other definitive evidence.

Summary

The clinical assessment of diagnostic technologies should begin with an exploratory stage in which potential application of the new test is studied. Ideally, the accuracy, impact on health care providers, therapeutic impact, patient outcome, and pecuniary costs and benefits of the technology should then be systematically assessed. These steps need not be sequential but under the right circumstances may be established in a single trial. There are a large number of situations in which shortcuts are appropriate; for example, if a new test is shown to be both more accurate and less expensive than existing alternatives, its usefulness is established. Nevertheless, demonstration of accuracy is ordinarily not sufficient for dissemination of a new technology. While before–after studies with physicians' assessments of therapeutic impact are less costly than randomized controlled trials, the results may overestimate the benefit of the new diagnostic technology. In many situations methodologically rigorous randomized controlled trials that test whether a diagnostic technology not only improves accuracy and changes therapy but also improves outcome will be required. Attention to this framework for assessing diagnostic tests will avoid premature dissemination of expensive new technologies; ignoring the framework will result in inefficient use of increasingly limited health care resources.

References

1. Department of Clinical Epidemiology and Biostatistics, McMaster University Health Sciences Centre: How to read clinical journals: II. To learn about a diagnostic test. *Can Med Assoc J* 1981; 124: 703–710

2. Ransohoff DF, Feinstein AR: Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med* 1978; 299: 926–929

3. Griner PF, Mayewski RJ, Mushlin AI et al: Selection and interpretation of diagnostic tests and procedures. Principles and applications. *Ann Intern Med* 1981; 94 (4 pt 2): 557–592

4. Banta HD, Behney CJ: Policy formulation and technology assessment. *Milbank Mem Fund Q* 1981; 59: 445-479

5. Fineberg HV, Bauman R, Sosman M: Computerized cranial tomography. Effect on diagnostic and therapeutic plans. *JAMA* 1977; 238: 224-227

6. Banta HD, Behney CJ, Willems JS: *Toward Rational Technology in Medicine: Considerations for Health Policy*, Springer-Verlag, New York, 1981

7. Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ont.: Interpretation of diagnostic data: 5. How to do it with simple maths. *Can Med Assoc J* 1983; 129: 947-954

8. Swets JA, Pickett RM, Whitehead SF et al: Assessment of diagnostic technologies: advanced measurement methods are illustrated in a study of computed tomography of the brain. *Science* 1979; 205: 753-759

9. Kramer MS, Feinstein AR: Clinical biostatistics. LIV. The biostatistics of concordance. *Clin Pharmacol Ther* 1981; 29: 111-123

10. Fleiss J: *Statistical Methods for Rates and Proportions*, Wiley, New York, 1973: 146-153

11. Cohen J: Weighted kappa. *Psychol Bull* 1968; 70: 213-230

12. Rozanski A, Diamond GA, Berman D et al: The declining specificity of exercise radionuclide ventriculography. *N Engl J Med* 1983; 309: 518-522

13. Diamond GA, Forrester JS: Metadiagnosis: an epistemologic model of clinical judgment. *Am J Med* 1983; 75: 129-137

14. Feinstein AR: An additional basic science for clinical medicine: II. The limitations of randomized trials. *Ann Intern Med* 1983; 99: 544-550

15. Alperovitch A: Controlled assessment of diagnostic techniques: methodological problems. *Effect Health Care* 1983; 1: 187-190

16. Robbins AH, Pugatch RD, Gerzof SG et al: Observations on the medical efficacy of computed tomography of the chest and abdomen. *Am J Roentgenol* 1978; 131: 15-19

17. Feeny D, Guyatt G, Tugwell P (eds): *Health Care Technology: Effectiveness, Efficiency and Public Policy* (in press)

18. Guyatt GH, Tugwell PX, Feeny DH et al: The role of before-after studies of therapeutic impact in the evaluation of diagnostic technologies. *J Chronic Dis* (in press)

19. Goldman L, Feinstein AR, Batsford WP et al: Ordering patterns and clinical impact of cardiovascular nuclear medicine procedures. *Circulation* 1980; 62: 680-687

20. Goldman L, Cohn PF, Mudge GH Jr et al: Clinical utility and management impact of M-mode echocardiography. *Am J Med* 1983; 75: 49-56

21. Wittenberg J, Fineberg HV, Ferrucci JT et al: Clinical efficacy of computed body tomography. *AJR* 1980; 134: 111-120

22. Dixon AK, Fry IK, Kingham JG et al: Computed tomography in patients with an abdominal mass: effective and efficient? A controlled trial. *Lancet* 1981; 1: 1199-1201

23. WHO cooperative trial on primary prevention of ischaemic heart disease using clofibrate to lower serum cholesterol: mortality follow-up. Report of the Committee of Principal Investigators. *Lancet* 1980; 2: 379-385

24. Dronfield MW, Langman MJS, Atkinson M et al: Outcome of endoscopy and barium radiography for acute upper gastrointestinal bleeding: controlled trial in 1037 patients. *Br Med J [Clin Res]* 1982; 284: 545-548

25. Brett GZ: The value of lung cancer detection by six-monthly chest radiographs. *Thorax* 1968; 23: 414-420

26. Ibid: Earlier diagnosis and survival in lung cancer. *Br Med J* 1969; 4: 260-262

27. Brown VA, Sawers RS, Parsons RJ et al: The value of antenatal cardiotocography in the management of high-risk pregnancy: a randomized controlled trial. *Br J Obstet Gynaecol* 1982; 89: 716-722

28. Flynn AM, Kelly J, Mansfield H et al: A randomized controlled trial of non-stress antepartum cardiotocography. Idem: 427-433

29. Feinstein AR: An additional basic science for clinical medicine: III. The challenges of comparison and measurement. *Ann Intern Med* 1983; 99: 705-712

30. Christie D: Before-and-after comparisons: a cautionary role. *Br Med J* 1979; 2: 1629-1630

31. Elias E, Hamlyn AN, Jain S et al: A randomized trial of percutaneous transhepatic cholangiography with the Chiba needle versus endoscopic retrograde cholangiography for bile duct visualization in jaundice. *Gastroenterology* 1976; 71: 439-443

32. Durbridge TC, Edwards F, Edwards RG et al: An evaluation of multiphasic screening on admission to hospital. Precis of a report to the National Health and Medical Research Council. *Med J Aust* 1976; 1: 703-705

33. Department of Clinical Epidemiology and Biostatistics, McMaster University Health Sciences Centre: How to read clinical journals: V. To distinguish useful from useless or even harmful therapy. *Can Med Assoc J* 1981; 124: 1156-1162

34. Kirshner B, Guyatt GH: A methodological framework for assessing health indices. *J Chronic Dis* 1985; 38: 27-36

35. Guyatt GH, Bombardier C, Tugwell PX: Measuring disease-specific quality of life in clinical trials. *Can Med Assoc J* (in press)

36. Sox HC Jr, Margulies I, Sox CH: Psychlogically mediated effects of diagnostic tests. *Ann Intern Med* 1981; 95: 680-685

37. Carrera GF, Gerson DE, Schnur J et al: Computed tomography of the brain in patients with headache or temporal lobe epilepsy: findings and cost-effectiveness. *J Comput Assist Tomogr* 1977; 1: 200-203

38. Larson EB, Omenn GS, Lewis H: Diagnostic evaluation of headache. Impact of computerized tomography and cost-effectiveness. *JAMA* 1980; 243: 359-362

39. Department of Clinical Epidemiology and Biostatistics, McMaster University Health Sciences Centre: How to read clinical journals: VII. To understand an economic evaluation (part A). *Can Med Assoc J* 1984; 130: 1428-1432, 1434

40. Idem: How to read clinical journals: VII. To understand an economic evaluation (part B). Ibid: 1542-1549

41. Drummond MF, Stoddart GL: Economic analysis and clinical trials. *Controlled Clin Trials* 1984; 5: 115-128

42. Zimmerman RA, Bilaniuk LT, Gennarelli T et al: Cranial computed tomography in diagnosis and management of acute head trauma. *Am J Roentgenol* 1978; 131: 27-34

43. Ambrose J, Gooding MR, Uttley D: E.M.I. scan in the management of head injuries. *Lancet* 1976; 1: 847-848

44. Hull R, Hirsh J, Sackett DL et al: Combined use of leg scanning and impedance plethysmography in suspected venous thrombosis. An alternative to venography. *N Engl J Med* 1977; 296: 1497-1500