

When Is Measuring Sensitivity and Specificity Sufficient To Evaluate a Diagnostic Test, and When Do We Need Randomized Trials?

Sarah J. Lord, MBBS, MS; Les Irwig, MBBCh, PhD; and R. John Simes, MBBS, MS, MD

The clinical value of using a new diagnostic test depends on whether it improves patient outcomes beyond the outcomes achieved using an old diagnostic test. When can studies of diagnostic test accuracy provide sufficient information to infer clinical value, and when do clinicians need to wait for results from randomized trials? The authors argue that accuracy studies suffice if a new diagnostic test is safer or more specific than, but of similar sensitivity to, an old test. However, if a new test is more sensitive than an old test, it leads to the detection of extra cases of disease. Results from treatment trials that enrolled only patients detected by

the old test may not apply to these extra cases. Clinicians need to wait for results from randomized trials assessing treatment efficacy in cases detected by the new diagnostic test, unless they can be satisfied that the new test detects the same spectrum and subtype of disease as the old test or that treatment response is similar across the spectrum of disease.

Ann Intern Med. 2006;144:850-855.

For author affiliations, see end of text.

www.annals.org

Diagnostic tests are used to confirm, exclude, classify, or monitor disease to guide treatment. Their clinical value depends on whether the information they provide leads to improved patient outcomes; this can be assessed by randomized trials that compare patient outcomes from the new diagnostic test versus the old test strategy. However, randomized trials of test-and-treatment strategies are not routinely performed. They are not required for marketing approval, and they are not always feasible because they require large sample sizes. As a result, new diagnostic tests frequently enter clinical practice without evidence of improved patient outcomes.

Studies of diagnostic test accuracy can show how well diagnostic strategies that include a new test identify the presence or absence of disease compared with an old test strategy by comparing each with the results of a reference standard test. However, if clinicians use only information about test accuracy to decide whether to adopt a new diagnostic test, they sometimes may harm patients (for example, if subsequent treatments are unsafe or ineffective). It is therefore worthwhile to investigate how best to decide if clinicians can rely on evidence about test accuracy or if they need to wait for patient outcome results from randomized trials. Fryback and Thornbury (1) described a hierarchy of 6 levels of evidence for the assessment of a diagnostic test: 1) the technical quality of test information; 2) diagnostic accuracy; 3) change in the referring physician's diagnostic thinking; 4) change in the patient management plan; 5) change in patient outcomes; and 6) societal costs and benefits. This framework does not provide guidelines about if and when lower levels of evidence are adequate to assess a test and always requires randomized trials for conclusions about improved patient outcomes. Some researchers have suggested that accuracy studies alone may sometimes suffice (2–6). We provide a practical framework to help clinicians decide whether a new diagnostic test can be adopted on the basis of evidence of test accuracy alone or if they need to wait for results from randomized trials.

A FRAMEWORK FOR DECIDING IF EVIDENCE OF TEST ACCURACY WILL SUFFICE

Studies of diagnostic test accuracy can suffice if clinicians already have evidence from randomized trials showing that treatment of the cases detected by the diagnostic test improved patient outcomes (**Figure 1**). This approach may seem straightforward; however, it requires a clear understanding of the proposed use and benefits of the new test, as well as careful consideration of whether the cases detected are representative of the patients included in treatment trials.

The benefits of a new diagnostic test will vary according to how it is used. Investigators of a new diagnostic test need to explicitly state who will be tested, where the new test will fit in the existing diagnostic pathway, and what tests it will supplement or replace. This information will allow them to identify the expected benefits of adopting the new test and, therefore, the most relevant questions to ask to assess its value. Test attributes generally fall into 3 categories: 1) The test is safer or is less costly; 2) the test is more specific (excludes more cases of nondisease) and thus avoids unnecessary treatment; and 3) the test is more sensitive (detects more cases of disease) and thus promotes more appropriate treatment.

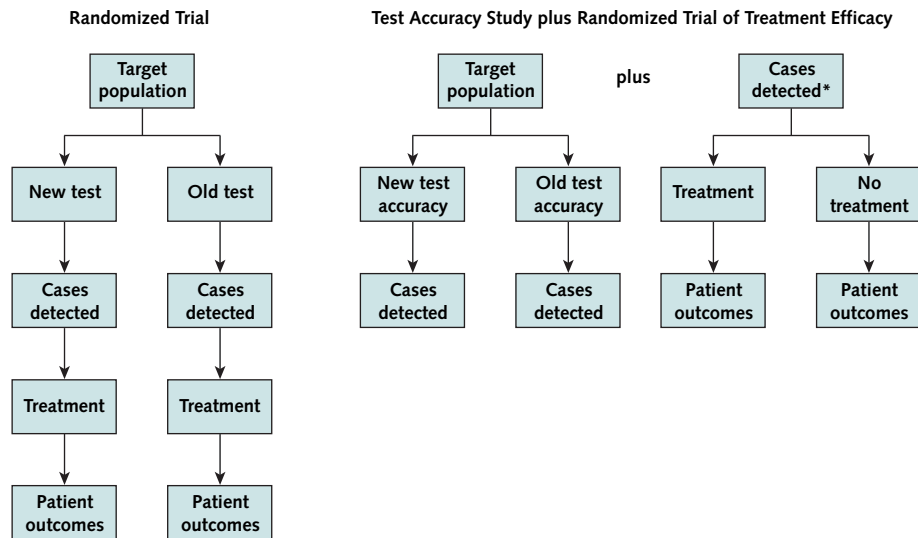
We propose a simple sequence of questions to guide decisions about whether evidence of test accuracy will suffice for each of these 3 categories (**Figure 2**). The first step in assessing a new diagnostic test is to classify it according to whether it is more sensitive than the old test. We describe the key concepts of this framework using simple examples to describe situations where the new test offers better safety or specificity with similar sensitivity, followed by

See also:

Web-Only

Conversion of figures and table into slides

Figure 1. Trial evidence versus linked evidence of test accuracy and treatment efficacy.



*Cases detected by the new and old test may not show similar response to treatment.

consideration of situations where the test is more sensitive. We also consider other, more complex scenarios. In each of these examples, we have some existing evidence of treatment efficacy for cases detected by an old test, and therefore, the rationale for testing has already been established.

When a New Test Has Similar Sensitivity to an Old Test

If a new diagnostic and old diagnostic test have similar sensitivity, it is generally reasonable to assume that they will detect the same true cases of disease. However, a new test may offer other positive attributes, such as better safety or more specificity than an old test. If studies of test accuracy show that the new test offers other positive attributes without a loss of sensitivity, it is logical to assume that cases detected by either test will show the same response to treatment. Therefore, new trials assessing treatment efficacy in the cases detected by the new test are not needed.

When a new test and an old test show similar sensitivity and specificity, the value of the new test corresponds to the benefits of avoiding the adverse events or costs associated with the old test. For example, Doppler ultrasonography has been adopted as a less invasive replacement for venography for diagnosis of deep venous thrombosis (7). Given that trials have shown that treatment of this condition is effective (8), studies of test accuracy showing that Doppler ultrasonography and venography have similar sensitivity and specificity suffice to determine the value of the new test (Table, example 1).

If a new test is more specific than an old test, it avoids the harms of unnecessary further investigation or treatment. In this scenario, evidence showing that the new test has similar sensitivity but better specificity than the old test will suffice to determine the new test's value. For example, new immunochemical fecal occult blood testing kits have

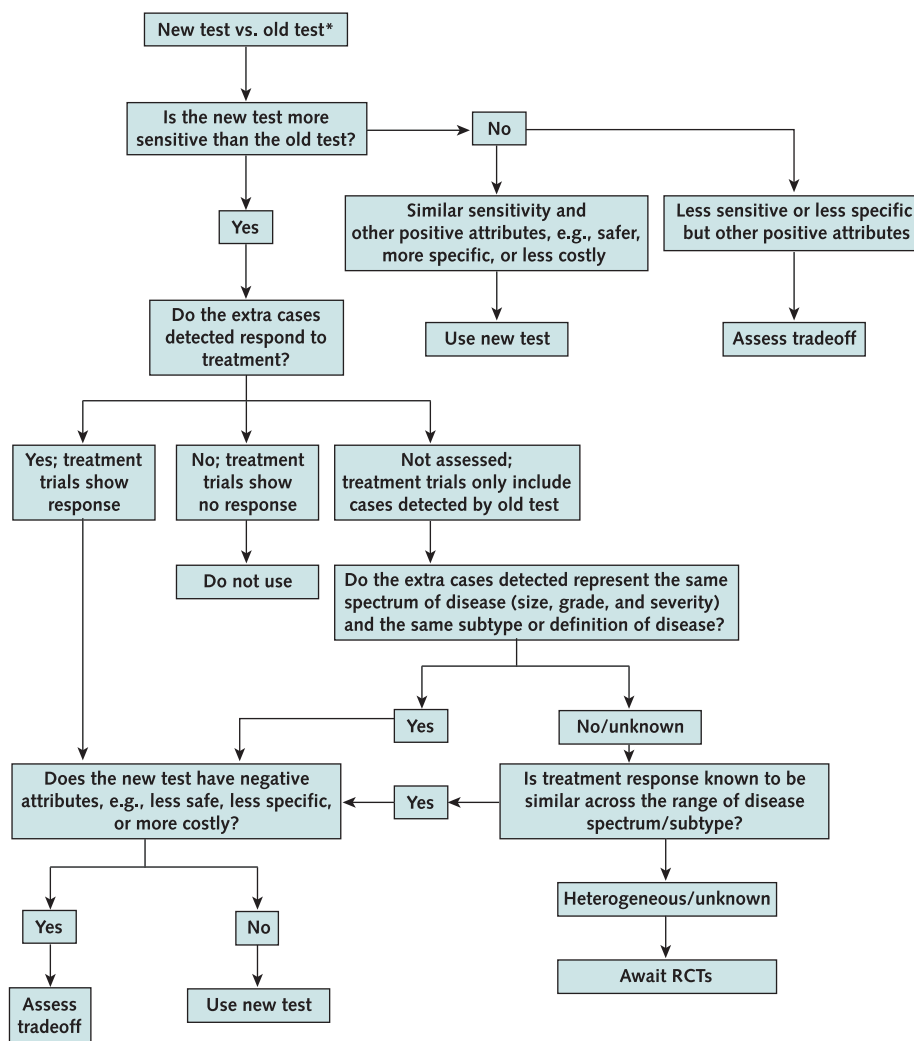
been proposed to offer increased specificity, without a loss of sensitivity, for detecting human blood, reducing the high false-positive rate of the older guaiac-based test kits (9). Given that trials have already shown the benefit of detecting colorectal cancer with guaiac test kits, studies of test accuracy will suffice to determine whether the immunochemical tests have increased specificity without any substantial loss in sensitivity (Table, example 2).

Our assessment is more complex when we consider a new test that involves a tradeoff between positive and negative attributes. We deal with this situation later.

When the New Test Is More Sensitive than the Old Test

If a new test is more sensitive than an old test but has similar specificity, its value is directly related to the treatment response in the extra cases detected. If treatment response has already been assessed by treatment trials enrolling patients detected by the new test, decisions to use the test will be based on whether these trials showed that treatment improves patient outcomes. In this instance, evidence of test accuracy linked with evidence of treatment efficacy replaces the need for new randomized trials (Figure 1).

There will also be a good match between tested and treated populations if the results for patients identified by a new test are analyzed in a treatment trial as potential predictors of treatment response. For example, trials of adjuvant tamoxifen among women with early breast cancer have shown that the estrogen receptor status of the tumor determines its response to treatment (10). These trials demonstrate the clinical value of testing for estrogen receptor status, as well as the effectiveness of tamoxifen therapy. However, more commonly, treatment trials have only en-

Figure 2. Assessing new tests using evidence of test accuracy, given that treatment is effective for cases detected by the old test.

RCT = randomized, controlled trial. * New test = diagnostic strategies that include the new test; old test = standard diagnostic strategies that do not include the new test.

rolled cases detected by an old test. In these situations, clinicians need to consider whether the results apply to cases detected by the new test.

Assessing Whether the Extra Cases Detected by a New, More Sensitive Test Respond to Treatment

Clinicians first need to ask whether the extra cases detected by a new diagnostic test represent the same spectrum or subtypes of disease as those included in treatment trials. If they do, then randomized trials may not be required (Figure 2). If they do not, or if clinicians cannot be certain that they do, then clinicians also need to ask whether treatment response is known to be similar across the range of disease spectrum or subtype. These considerations are important regardless of the magnitude of the difference in sensitivity between tests.

At one extreme, a good match between tested and

treated populations is possible if the reference standard used to determine test sensitivity and specificity is also the starting point for trials conducted in similar populations. For example, computed tomography colonography is more sensitive for the detection of large colorectal polyps when scanning is performed with the patient in prone and supine positions versus supine positioning alone, without reduced specificity (11). Treatment trials showing improved survival following the early detection and treatment of colorectal polyps have been based on cases detected by colonoscopy, the reference standard for the computed tomography colonography accuracy studies. All of these tests are used to identify the same disease characteristic (adenomatous colorectal polyps) and spectrum of disease (classified by polyp size), so it is reasonable to conclude that computed tomography colonography with dual positioning will improve pa-

tient outcomes compared with supine positioning alone (Table, example 3).

At the other extreme, a good match between tested and treated populations will not be possible if a new test measures a different biological characteristic to define disease and leads to a different selection of cases for treatment. For example, a new nuclear magnetic resonance spectroscopy technique to measure lipoprotein particle size and concentration has been proposed as a more accurate test than plasma lipoprotein cholesterol levels to identify patients who will benefit from lipid-lowering therapy (12). However, the effectiveness of treatment for the extra patients detected by the new rather than the old test has not been assessed (Table, example 4).

Often we need to consider situations between these 2 extremes, including situations where the evidence is incomplete. Accuracy studies and treatment trials are usually done by different investigators at different times and in different settings. Variations in the clinical setting and spectrum of disease may lead to genuine differences (heterogeneity) in test accuracy and treatment effect (13, 14),

limiting clinicians' ability to link evidence between studies and draw conclusions about the value of a new test. Consider magnetic resonance angiography versus conventional arteriography for detecting lower-limb arterial disease. Although magnetic resonance angiography is a more sensitive test for patients with claudication, the effectiveness of surgical and percutaneous interventions has been well established only in patients with critical limb ischemia (15). Whether this evidence can be linked to infer that magnetic resonance angiography will improve patient outcomes depends on whether clinicians can assume that test accuracy and treatment response are similar for patients with different disease severity.

A new test may also produce a shift in the spectrum of disease detected. For example, breast magnetic resonance imaging in addition to mammography is more sensitive than mammography alone for the early detection of invasive breast cancer in young women at high risk. However, the extra cases detected may represent a different spectrum of disease that does not show improved treatment response. Therefore, the value of breast magnetic resonance imaging

Table. New Diagnostic Test Assessment Framework and Examples*

Test and Indication	Proposed Benefits of New Diagnostic Test	Does the Evidence of Effective Treatment Apply to the Cases Detected by the New Test?	Can Studies of Test Accuracy Suffice?
Example 1: Doppler ultrasonography vs. venography for detection of deep venous thrombosis	Safer, less costly	Yes; no change to the definition or spectrum of disease.	Yes; the value of the new test may be inferred from an assessment of its relative safety and cost.
Example 2: new vs. standard FOBT for early detection of colorectal cancer	More specific	Yes; no change to the definition or spectrum of disease.	Yes; the value of the new test may be inferred from an assessment of its relative safety and cost and the benefits of avoiding a false-positive result.
Example 3: supine and prone positioning for CT colonography vs. supine positioning alone for detection of adenomatous colorectal polyps	More sensitive	Yes; no change to the definition or spectrum of disease because cases detected by both tests are subsequently confirmed by colonoscopy.	Yes; the value of the new test may be inferred from an assessment of its relative sensitivity, given no substantial loss in specificity and existing trial evidence that treatment improves patient outcomes.
Example 4: NMR spectroscopy vs. plasma lipoprotein levels for the detection of hyperlipidemia	More sensitive	Unknown; it is unknown whether NMR detects patients who would benefit more or less from lipid-lowering agents than those whose disorder was detected using the existing test.	No; trial evidence is required to determine the value of the new test.
Example 5: MRI vs. mammography for earlier detection of invasive breast cancer in women at high risk for the disease	More sensitive	No; it is uncertain if any benefits of treatment at the earlier stage of disease outweigh the harm of overdiagnosis of cancer that would never have presented clinically.	No; trial evidence is required to determine the impact of testing on patient outcomes or, at least, the interval breast cancer rate.
Example 6: PET, MRI, and EEG vs. MRI and EEG to detect an epileptogenic focus in patients with medically refractory epilepsy who are being considered for surgery	More sensitive	Uncertain; it is uncertain whether patients with functional lesions detected by PET for whom existing standard imaging yields negative or inconclusive results will show the same treatment response to surgery as patients with structural lesions that can be detected with standard imaging alone.	Judgment is needed about whether clinicians require randomized trials to assess treatment response in the extra cases detected by PET or whether they can rely on existing trials conducted in patients with disease detected by MRI and EEG and observational evidence about treatment response in cases detected by PET.

* CT = computed tomography; EEG = electroencephalography; FOBT = fecal occult blood test; MRI = magnetic resonance imaging; NMR = nuclear magnetic resonance; PET = positron emission tomography.

is uncertain without trials comparing patient outcomes after early versus standard detection (16) (Table, example 5).

When the possibility of a spectrum effect and its impact on treatment efficacy are considered, it is useful to determine whether a new test performs consistently in different populations and whether treatment of the detected condition is effective across different patient subgroups. Previous studies that identify subgroups in which the test is less accurate or the treatment less effective may be available and may assist in making judgments about generalizability between the cases detected by a new test and patients included in treatment trials. In many cases, generalizability will depend on expert opinion. Sometimes, differences in the spectrum of disease detected may be obvious because the new test detects disease of a different size, stage, grade, or severity. In other cases, this judgment will be less clear.

Consider the use of positron emission tomography as an incremental test to detect an operable epileptogenic focus in patients with medically refractory epilepsy who have no structural lesion on magnetic resonance imaging. There is trial evidence that seizure control improves after surgery for patients with structural lesions (17). However, the extra cases detected by positron emission tomography may represent a different spectrum of disease. Thus, clinicians' decisions whether evidence of test accuracy will suffice will depend on whether it is plausible that the response to treatment is similar in patients with structural lesions (the treated population in this example) and those with functional lesions detected by positron emission tomography (positive results on the new test in the tested population) (Table, example 6). The willingness of clinicians to accept assumptions about generalizability depends on the estimated costs of a new test and the severity of the consequences if the assumptions are proved false. If the costs or consequences are substantial, clinicians should wait for direct evidence from trials that include the extra cases detected by the new test.

OTHER CONSIDERATIONS

The examples discussed here do not cover all possible scenarios. For example, it is possible for a new test to have overall sensitivity and specificity that are similar to those of an old test but to detect different cases of disease if it identifies patients at a different disease spectrum. However, we think our approach provides a reasonable and efficient framework to deal with most situations. It helps clinicians to quickly identify more complex situations where a new test offers a tradeoff between positive and negative attributes (for example, if a new test is less invasive but also less specific than the existing test). Here, the benefits of better safety need to be assessed against the harms arising from additional false-positive results. Such tradeoffs can be assessed by using a decision analytic model to assess the benefits and harms of new and old tests, including the rates and the consequences of true-positive, false-positive, true-

negative, and false-negative results, as well as test complications. Decision analysis also allows clinicians to compare the effects of testing in populations with a different prevalence of disease. If a new test offers better sensitivity but has other negative attributes, decision analysis may not be appropriate because it relies on assumptions that evidence of test accuracy can be linked to evidence of treatment efficacy. If this linkage is uncertain, clinicians will need to call for new randomized trials. In these situations, trials investigating the effect of treatment in patients who have positive results on the new test and negative results on the old test may be more efficient and more clinically relevant than trials conducted on all patients in whom the new test yields positive results (18).

CONCLUSION

Whenever clinicians use a new diagnostic test because it is more sensitive than an old test, they need to be clear about the assumptions linking this evidence to improved patient outcomes, such as evidence that the new test detects the same spectrum of disease as the old test or has similar treatment efficacy across the spectrum of disease. The selection of effective tests is just as essential as the selection of effective treatments; thus, whenever clinicians' assumptions about a new diagnostic test are in doubt, they should wait for evidence from randomized trials.

From The University of Sydney, Sydney, Australia.

Grant Support: The authors have received funding from the Australian Medical Services Advisory Committee for the development of guidelines for the assessment of diagnostic technologies and National Health and Medical Research Council Program Grants (253602, 402764)

Potential Financial Conflicts of Interest: None disclosed.

Requests for Single Reprints: Sarah J. Lord, MBBS, MS, National Health and Medical Research Council Clinical Trials Centre, University of Sydney, 88 Mallett Street, Camperdown NSW, 2050 Australia; e-mail, sally.lord@ctc.usyd.edu.au.

Current author addresses are available at www.annals.org.

References

1. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. *Med Decis Making*. 1991;11:88-94. [PMID: 1907710]
2. Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. *CMAJ*. 1986;134:587-94. [PMID: 3512062]
3. Barratt A, Irwig L, Glasziou P, Cumming RG, Raffle A, Hicks N, et al. Users' guides to the medical literature: XVII. How to use guidelines and recommendations about screening. Evidence-Based Medicine Working Group. *JAMA*. 1999;281:2029-34. [PMID: 10359392]
4. Busse R, Orvain J, Velasco M, Perleth M, Drummond M, Gurtner F, et al. Best practice in undertaking and reporting health technology assessments. Working group 4 report. *Int J Technol Assess Health Care*. 2002;18:361-422. [PMID: 12053427]
5. Harris RP, Helfand M, Woolf SH, Lohr KN, Mulrow CD, Teutsch SM, et al. Current methods of the US Preventive Services Task Force: a review of the

process. *Am J Prev Med.* 2001;20:21-35. [PMID: 11306229]

6. Biesheuvel CJ, Grobbee DE, Moons KG. Distraction from randomization in diagnostic research. *Ann Epidemiol.* 2006. [PMID: 16386925]

7. Dauzat M, Laroche JP, Deklunder G, Ayoub J, Quéré I, Lopez FM, et al. Diagnosis of acute lower limb deep venous thrombosis with ultrasound: trends and controversies. *J Clin Ultrasound.* 1997;25:343-58. [PMID: 9282799]

8. Gould MK, Dembitzer AD, Doyle RL, Hastie TJ, Garber AM. Low-molecular-weight heparins compared with unfractionated heparin for treatment of acute deep venous thrombosis. A meta-analysis of randomized, controlled trials. *Ann Intern Med.* 1999;130:800-9. [PMID: 10366369]

9. Ouyang DL, Chen JJ, Getzenberg RH, Schoen RE. Noninvasive testing for colorectal cancer: a review. *Am J Gastroenterol.* 2005;100:1393-403. [PMID: 15929776]

10. Tamoxifen for early breast cancer: an overview of the randomised trials. Early Breast Cancer Trialists' Collaborative Group. *Lancet.* 1998;351:1451-67. [PMID: 9605801]

11. Fletcher JG, Johnson CD, Welch TJ, MacCarty RL, Ahlquist DA, Reed JE, et al. Optimization of CT colonography technique: prospective trial in 180 patients. *Radiology.* 2000;216:704-11. [PMID: 10966698]

12. Otvos JD, Jeyarajah EJ, Cromwell WC. Measurement issues related to lipoprotein.

13. Bailey KR. Generalizing the results of randomized clinical trials. *Control Clin Trials.* 1994;15:15-23. [PMID: 8149769]

14. Irwig L, Bossuyt P, Glasziou P, Gatsonis C, Lijmer J. Designing studies to ensure that estimates of test accuracy are transferable. *BMJ.* 2002;324:669-71. [PMID: 11895830]

15. Koelemay MJ, Lijmer JG, Stoker J, Legemate DA, Bossuyt PM. Magnetic resonance angiography for the evaluation of lower extremity arterial disease: a meta-analysis. *JAMA.* 2001;285:1338-45. [PMID: 11255390]

16. Irwig L, Houssami N, Armstrong B, Glasziou P. Evaluating new screening tests for breast cancer [Editorial]. *BMJ.* 2006;332:678-9. [PMID: 16565097]

17. Wiebe S, Blume WT, Girvin JP, Eliasziw M. A randomized, controlled trial of surgery for temporal-lobe epilepsy. *N Engl J Med.* 2001;345:311-8. [PMID: 11484687]

18. Bossuyt PM, Lijmer JG, Mol BW. Randomised comparisons of medical tests: sometimes invalid, not always efficient. *Lancet.* 2000;356:1844-7. [PMID: 11117930]

SERVICES FOR PDA USERS

See the "PDA Services" heading at the top of the *Annals* home page at www.annals.org to learn more about the following services available to PDA users.

Annals Reader for Palm PDA

Editors' Notes, abstracts, and full texts. Images and references are not available.

Reminders from PDA to the desktop computer about articles. The reminders include a link to the full text of the article at www.annals.org.

Storage of multiple issues of *Annals*. You can control how many issues to store on your PDA or have the reader automatically delete back issues.

Issues may be stored on external memory. Back issues are available starting with July 2003.

Storage of favorite articles (separate from storage of back issues).

Storage of collections of previously published articles from *Annals*.

Wireless access to the current issue for an Internet-enabled PDA

Current Author Addresses: Drs. Lord and Simes: National Health and Medical Research Council Clinical Trials Centre, University of Sydney, 88 Mallett Street, Camperdown NSW, 2050 Australia.

Dr. Irwig: School of Public Health, A27, Edward Ford Building, University of Sydney, NSW 2006 Australia.

COMMENTS AND RESPONSES

A Philosophical Approach to Diagnostic Test Evaluation

TO THE EDITOR: In their recent article, Lord and colleagues (1) propose a framework to decide when measuring diagnostic test accuracy is sufficient to evaluate a diagnostic test and when a randomized trial is needed. Their work, in accordance with most previous papers on the subject, considers accuracy and patient outcome studies as separate phases in the evaluation of a diagnostic test: Accuracy studies show how well a diagnostic test identifies the true presence or absence of disease, and patient outcome studies show whether a patient classification based on the new test better predicts patient outcome. We feel that this phased evaluation of diagnostic tests lacks appreciation of the fact that disease, by definition, is a prognostic statement and that, consequently, accuracy studies cannot be interpreted independent of patient outcome.

The prognostic character of disease is intuitively attractive. All diseases, without exception, in some way influence a patient's ability to function according to his or her aims or goals. Having red hair clearly is not a disease, nor is the presence of variations on normal anatomy, such as an accessory pancreatic duct, or genetic polymorphisms, that have no bearing on the individual's functioning. This is what Christopher Boorse referred to when he defined a disease entity as the state of an individual that interferes with or prevents the normal function of an organ or a system of organs (2). The normal function of an organ can be seen in relation to value-free goals, such as the individual's survival and reproduction (also called an essentialist approach), or to value-laden goals, such as welfare and quality of life (also called a holistic approach) (3). Of note, however, the realization of these goals takes place in the future. Consequently, disease cannot be defined as a current state of being but must include a reference to a future state.

Accuracy studies correlate a diagnostic test or test regimen with the current presence or absence of disease as measured with a gold standard or reference test. However, accepting that disease cannot be defined without reference to a future state logically implies that patient outcome is the ultimate gold standard test. In other words, the ideal diagnostic test perfectly predicts patient outcome and the ideal accuracy study is, therefore, a patient outcome study. A situation in which the evaluation of each new test requires lengthy follow-up, however, is not desirable. Accuracy studies may bypass the need for follow-up by using a reference test as an approximation of patient outcome. This approximation is appropriate only if the test's association with patient outcome is known with acceptable certainty and can be quantitatively described.

In conclusion, we believe that accuracy and patient outcome studies have the same objective, that is, to determine how well a diagnostic test classifies patients according to patient outcome (instead of disease presence). Accuracy studies are a way to bypass the need for lengthy follow-up by using a less-than-perfect reference test that approximates patient outcome. This approximation is appropriate only if its association with patient outcome can be quantitatively described. In all other circumstances, patient outcome studies must be performed.

Gabe S. Sonke, MD, PhD
Netherlands Cancer Institute
Amsterdam, the Netherlands

André L.M. Verbeek, MD, PhD
Lambertus A.L.M. Kiemeny, PhD
Radboud University Nijmegen Medical Centre
Nijmegen, the Netherlands

Potential Financial Conflicts of Interest: None disclosed.

References

1. Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? *Ann Intern Med.* 2006;144:850-5. [PMID: 16754927]
2. Boorse C. Health as a theoretical concept. *Philos Sci.* 1977;44:552-73.
3. Gammelgaard A. Evolutionary biology and the concept of disease. *Med Health Care Philos.* 2000;3:109-16. [PMID: 11079338]

IN RESPONSE: We agree with Dr. Sonke and colleagues that one purpose of test accuracy studies is to determine whether the test provides accurate information about long-term clinical outcomes, that is, what is the prognostic value of the test information? We also agree that an ideal study to achieve this purpose is one in which the reference standard is a good proxy for patient outcome. In the absence of effective treatment, this information would be all that is required. However, if treatment is available, we also need to ask whether the test identifies patients whose outcomes would be improved by using this treatment, that is, what is the therapeutic value of testing? This is the issue we addressed in our article. Assessing the therapeutic value of a test does not just involve asking whether it provides accurate prognostic information. We also need to consider whether it identifies patients whose prognosis will improve with treatment (1).

In some cases, accuracy studies suffice because we already have evidence about effective treatments for the cases detected. In other situations, we require new randomized, controlled trials to assess the effect of test and treatment on patient outcomes.

To illustrate the difference between these 2 questions, let's consider a new test, such as a more detailed ultrasonographic study for detecting acute deep venous thrombosis. As Dr. Sonke and colleagues suggest, an accuracy study would suffice for conclusions about the prognostic value of the test if the relationship between the spectrum of disease detected by an abnormal venogram (the reference standard) and subsequent patient morbidity and mortality has been adequately established. If not, we would need a patient outcome study to assess the new test. However, to assess the therapeutic value of the test, we still need to consider whether detection improves patient outcomes. If evidence suggests that anticoagulant therapy is effective treatment for the spectrum of disease detected by the new test, a comparison of the accuracy and safety of the new test versus those of a standard test will suffice for conclusions about the new test's therapeutic value. Consider the use of D-dimer testing to identify patients at high risk for recurrence of deep venous thrombosis who may benefit from a longer course of anticoagulant therapy: Patient outcome studies have shown the test is prognostic for recurrence of deep venous thrombosis, but does this evidence suffice for conclusions about its therapeutic value? Neither simple prognostic studies nor existing treatment trials allow conclusions about whether

using the test and subsequent treatment in this population would improve patient outcomes, and therefore, a new randomized, controlled trial is required. Such a trial has been done and provides evidence about both the prognostic value of testing and the effectiveness of treatment in patients with positive test results (2). Hence, sometimes new trials are required to determine the therapeutic value of testing, whereas in other circumstances, linking of accuracy studies to existing treatment trials may suffice.

Sarah J. Lord, MBBS, MS

Les Irwig, MBCh, PhD

R. John Simes, MBBS, MS, MD

The University of Sydney

Camperdown, New South Wales 2050, Australia

Potential Financial Conflicts of Interest: None disclosed.

References

1. Lijmer JG, Bossuyt PM. Diagnostic testing and prognosis: the randomised controlled trial in diagnostic research. In: Knottnerus JA, ed. *The Evidence Base of Clinical Diagnosis*. London: BMJ Books; 2002:61-80.
2. Palareti G, Cosmi B, Legnani C, Tosetto A, Brusi C, Iorio A, et al. D-dimer testing to determine the duration of anticoagulation therapy. *N Engl J Med*. 2006;355:1780-9. [PMID: 17065639]

Life Expectancy and Colorectal Cancer Screening

TO THE EDITOR: Gross and colleagues (1) provide important information demonstrating the expected reduction in life expectancy that accompanies advancing age and higher burden of comorbid conditions among patients with newly diagnosed colorectal cancer. While their observations may be particularly informative when considering treatment options after diagnosis of colorectal cancer, the direct relevance of their observations to decision making about screening is somewhat more complex. If the benefit of endoscopic screening for colorectal cancer derives from an earlier diagnosis of cancer, then life expectancy at the time of diagnosis is particularly relevant. However, to the extent that the endoscopy offers benefit through the identification and removal of colorectal polyps among patients without colorectal cancer, decisions about screening might better reflect consideration of life expectancy among individuals without colorectal cancer. While the issues of advancing age and comorbid conditions are still quite relevant (2, 3), knowing the extent to which the estimates of Gross and colleagues are sensitive to the presence of colorectal cancer would be of interest.

Robert C. Burack, MD, MPH

Wayne State University

Detroit, MI 48202

Potential Financial Conflicts of Interest: None disclosed.

References

1. Gross CP, McAvay GJ, Krumholz HM, Paltiel AD, Bhasin D, Tinetti ME. The effect of age and chronic illness on life expectancy after a diagnosis of colorectal cancer: implications for screening. *Ann Intern Med*. 2006;145:646-53. [PMID: 17088577]
2. Lin OS, Kozarek RA, Schembre DB, Ayub K, Gluck M, Drennan F, et al. Screening colonoscopy in very elderly patients: prevalence of neoplasia and estimated impact on life expectancy. *JAMA*. 2006;295:2357-65. [PMID: 16720821]

3. Walter LC, Covinsky KE. Cancer screening in elderly patients: a framework for individualized decision making. *JAMA*. 2001;285:2750-6. [PMID: 11386931]

IN RESPONSE: Colorectal cancer screening can benefit patients through 2 mechanisms. First, a screening colonoscopy will increase the likelihood of early-stage cancer diagnosis among patients with prevalent disease (1). Because of the slow-growing nature of colorectal cancer, clinical benefits of this stage shift take years to accrue (1, 2). This is because the colorectal cancer survival benefit is due to the detection and treatment of early-stage cancer that otherwise would not have resulted in death for at least 5 years. Hence, further understanding of life expectancy after early-stage cancer diagnosis should inform decision making about screening by identifying patients who are unlikely to live long enough to receive the benefit. Second, as Dr. Burack emphasizes, colorectal cancer screening can also benefit patients by the removal of precancerous lesions and the decrease in the incidence of colorectal cancer. However, the interval between the onset of an adenoma and the clinical diagnosis of cancer has been estimated to be as long as 20 years (3).

Trial data support these estimates. Investigators of the Minnesota Colon Cancer Control Study randomly assigned more than 46 000 patients to a fecal occult blood test (FOBT) group or to a control group (1). Patients randomly assigned to the annual FOBT group had a significantly lower cumulative colorectal cancer mortality rate than that of the control group (0.59% vs. 0.88%) after 13 years of follow-up. They were also more likely to receive a diagnosis of early-stage disease (1). Yet differences in incidence took longer to accrue. Whereas incidence rates at 13 years did not significantly differ, the incidence rate at 18 years was significantly lower in the annual FOBT group than in the control group (3.2% vs. 3.9%; $P < 0.001$) (4). Given that the clinical benefits of the stage shift may be observed in as little as 5 years, whereas those of adenoma removal may take 15 years or longer, we feel that the shorter-term interval should guide decision making. We agree that further work is needed to explore factors associated with life expectancy in the cancer-free population. While this analysis was beyond the scope of our study, the relative survival of early-stage colon cancer approaches 90% (5). Therefore, although substantial differences in life expectancy estimates between patients with early-stage colorectal cancer and those without cancer would be expected, empirical analyses should validate this assumption.

Cary P. Gross, MD

Gail J. McAvay, PhD

Mary E. Tinetti, MD

Yale University

New Haven, CT 06520

Potential Financial Conflicts of Interest: None disclosed.

References

1. Mandel JS, Bond JH, Church TR, Snover DC, Bradley GM, Schuman LM, et al. Reducing mortality from colorectal cancer by screening for fecal occult blood. Minnesota Colon Cancer Control Study. *N Engl J Med*. 1993;328:1365-71. [PMID: 8474513]
2. Kronborg O, Fenger C, Olsen J, Jørgensen OD, Søndergaard O. Randomised study of screening for colorectal cancer with faecal-occult-blood test. *Lancet*. 1996;348:1467-71. [PMID: 8942774]
3. Loeve F, Boer R, van Oortmarssen GJ, van Ballegooijen M, Habbema JD. The

MISCAN-COLON simulation model for the evaluation of colorectal cancer screening. *Comput Biomed Res.* 1999;32:13-33. [PMID: 10066353]

4. Mandel JS, Church TR, Bond JH, Ederer F, Geisser MS, Mongin SJ, et al. The effect of fecal occult-blood screening on the incidence of colorectal cancer. *N Engl J Med.* 2000;343:1603-7. [PMID: 11096167]

5. Ries LAG, Harkins D, Krapcho M, Mariotto A, Miller BA, Feuer EJ, et al., eds. *SEER Cancer Statistics Review, 1975-2003*. Bethesda, MD: National Cancer Institute; 2006. Accessed at http://seer.cancer.gov/csr/1975_2003/ on 3 April 2007.

CLINICAL OBSERVATION

Institutional Review Board Review of Multicenter Studies

Background: Consensus suggests that institutional review boards (IRBs) are operating beyond their capacities (1). For example, multicenter studies require assessment of the same protocol at each participating site's IRB, which can lead to variation in the review of individual protocols (2). Proposed reforms of the IRB system seek centralized (3) and regional (4) review of multicenter studies to improve the quality of review, reduce repetition, and increase efficiency. Although the precise number of multicenter reviews nationally is uncertain, it is growing (1). But data are lacking to estimate the potential benefits and cost savings of centralized review.

Methods: We surveyed 121 IRBs at academic medical centers to determine the number of new and continuing reviews at the local site in 2004, the fraction of reviews that were of multicenter studies, and the average number of other centers involved.

Findings: We received responses from 73 IRBs (60% response rate), which were roughly evenly distributed by annual protocol review volume (Table). Only 14 IRBs (19%) knew or could estimate the number of institutions involved in their multicenter studies. Only 32 IRBs (44%) recorded whether a submitted protocol was part of a multicenter study. The average number of new multicenter reviews among the IRBs was 62.

Discussion: Most IRBs seem to lack mechanisms to collect precise information about multicenter studies. Yet, IRBs should know of other centers involved in multicenter studies—if only to monitor amendments and adverse events occurring at other sites. Centralized review improves tracking and management of multicenter studies. It could also yield cost savings from the benefits of economies of scale within larger IRBs (5). From the ratio of the number of new multicenter reviews to the total number of new reviews for each class of

IRB, we estimate that centralization of multicenter reviews could yield cost savings between roughly 10% and 35% of the total costs (Table). However, infrequent and inconsistent tracking of multicenter information by IRBs means that numerical estimates of financial benefits are imprecise. For many institutions facing escalating costs of administering an IRB, such cost reductions could be substantial.

Conclusion: These findings suggest that few IRBs document the multicenter status of studies they review. Estimates from data also suggest that institutional, managerial, financial, and quality benefits from centralization of IRB review could be substantial.

Gregory K. Sobolski, BA, BS

Leonardo Flores, BA

Ezekiel J. Emanuel, MD, PhD

National Institutes of Health

Bethesda, MD 20892

Disclaimer: The views contained herein are those of the authors and do not necessarily represent the views of the National Institutes of Health, U.S. Public Health Service, or U.S. Department of Health and Human Services.

Acknowledgment: The authors thank Amy Reid for careful assistance with collection of the data.

Potential Financial Conflicts of Interest: None disclosed.

References

- Burman WJ, Reves RR, Cohn DL, Schooley RT. Breaking the camel's back: multicenter clinical trials and local institutional review boards. *Ann Intern Med.* 2001;134:152-7. [PMID: 11177319]
- McWilliams R, Hoover-Fong J, Hamosh A, Beck S, Beaty T, Cutting G. Problematic variation in local institutional review of a multicenter genetic epidemiology study. *JAMA.* 2003;290:360-6. [PMID: 12865377]
- Christian MC, Goldberg JL, Killen J, Abrams JS, McCabe MS, Mauer JK, et al. A central institutional review board for multi-institutional trials. *N Engl J Med.* 2002;346:1405-8. [PMID: 11986418]
- Wood A, Grady C, Emanuel EJ. Regional ethics organizations for protection of human research participants. *Nat Med.* 2004;10:1283-8. [PMID: 15580245]
- Sugarman J, Getz K, Speckman JL, Byrne MM, Gerson J, Emanuel EJ, et al. The cost of institutional review boards in academic medical centers [Letter]. *N Engl J Med.* 2005;352:1825-7. [PMID: 15858200]

Table. Characteristics of Respondent Institutional Review Boards*

Characteristic	All IRBs (n = 73)	Low-Volume IRBs (<350 protocols) (n = 23)	Medium-Volume IRBs (350–699 protocols) (n = 29)	High-Volume IRBs (>699 protocols) (n = 21)
Average new protocols (tracking data), n (n)	599 (73)	191 (23)	502 (29)	1179 (21)
Average new multicenter protocols (tracking data), n (n)	62 (32)	66 (16)	43 (12)	115 (4)
Cost per IRB action, \$†	1021	1455	1104	839
Ratio of new multicenter protocols to new protocols, %	–	35	9	10

* IRB = institutional review board.

† Data from Sugarman et al. (5).