D iagnostic 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
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 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-
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-

---

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

- **New test vs. old test**
  - **Yes**
    - **Similar sensitivity and other positive attributes, e.g., safer, more specific, or less costly**
    - **Do the extra cases detected respond to treatment?**
      - **Yes; treatment trials show response**
        - **Use new test**
      - **No; treatment trials show no response**
        - **Assess tradeoff**
  - **No**
    - **Less sensitive or less specific but other positive attributes**
    - **Do the extra cases detected represent the same spectrum of disease (size, grade, and severity) and the same subtype or definition of disease?**
      - **Yes**
        - **Is treatment response known to be similar across the range of disease spectrum/subtype?**
          - **Yes**
            - **Use new test**
          - **No/unknown**
            - **Await RCTs**
      - **No/unknown**
        - **Assess tradeoff**

**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.*
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**

<table>
<thead>
<tr>
<th>Test and Indication</th>
<th>Proposed Benefits of New Diagnostic Test</th>
<th>Does the Evidence of Effective Treatment Apply to the Cases Detected by the New Test?</th>
<th>Can Studies of Test Accuracy Suffice?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 1: Doppler ultrasonography vs. venography for detection of deep venous thrombosis</td>
<td>Safer, less costly</td>
<td>Yes; no change to the definition or spectrum of disease.</td>
<td>Yes; the value of the new test may be inferred from an assessment of its relative safety and cost.</td>
</tr>
<tr>
<td>Example 2: new vs. standard FOBT for early detection of colorectal cancer</td>
<td>More specific</td>
<td>Yes; no change to the definition or spectrum of disease.</td>
<td>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.</td>
</tr>
<tr>
<td>Example 3: supine and prone positioning for CT colonography vs. supine positioning alone for detection of adenomatous colorectal polyps</td>
<td>More sensitive</td>
<td>Yes; no change to the definition or spectrum of disease because cases detected by both tests are subsequently confirmed by colonoscopy.</td>
<td>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.</td>
</tr>
<tr>
<td>Example 4: NMR spectroscopy vs. plasma lipoprotein levels for the detection of hyperlipidemia</td>
<td>More sensitive</td>
<td>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.</td>
<td>No; trial evidence is required to determine the value of the new test.</td>
</tr>
<tr>
<td>Example 5: MRI vs. mammography for earlier detection of invasive breast cancer in women at high risk for the disease</td>
<td>More sensitive</td>
<td>No; it is uncertain if any benefits of treatment at the earlier stage of disease outweigh the harm of overdetection of cancer that would never have presented clinically.</td>
<td>No; trial evidence is required to determine the impact of testing on patient outcomes or, at least, the interval breast cancer rate.</td>
</tr>
<tr>
<td>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</td>
<td>More sensitive</td>
<td>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.</td>
<td>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.</td>
</tr>
</tbody>
</table>

* CT = computed tomography; EEG = electroencephalography; FOBT = fecal occult blood test; MRI = magnetic resonance imaging; NMR = nuclear medicine 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**

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