

When Should a New Test Become the Current Reference Standard?

Paul Glasziou, MB, BS, PhD; Les Irwig, MB, BCh, PhD; and Jonathan J. Deeks, PhD

The evaluation of claims that a new diagnostic test is better than the current gold standard test is hindered by the lack of a perfect reference judge. However, this problem may be sidestepped by focusing on the clinical consequences of the decision rather than on estimation of accuracy. Consequences can be assessed by use of a "fair umpire" test that is not perfect yet can discriminate between disease and nondisease cases and is not biased in favor of 1 test.

This article discusses 3 principles to aid judgments about the value of new tests. First, the consequences are best examined in

cases with disagreement between the current and new tests. Second, resolving these disagreements requires a fair, but not necessarily perfect, umpire test. Finally, umpire tests include consequences, such as prognosis and response to treatment, as well as causal exposures and other test results.

Ann Intern Med. 2008;149:816-821.

For author affiliations, see end of text.

www.annals.org

Time is often a better diagnostician than the best anatomical pathologist.

—Clifton Meador

New diagnostic tests do not simply alter our diagnostic processes; they may also change whom we classify as having a disease. Gold standard tests are often imperfect, which leads to modification or change of the standard—sometimes by deliberate decision and sometimes by stealth. A definitional shift occurs when a new test detects additional cases of apparent disease but creates uncertainty about whether these additional cases should be classified and treated in the same way. For example, magnetic resonance imaging for suspected multiple sclerosis has widened the diagnosis in practice, but the correlation of lesions found on imaging with eventual clinical course is poor (1); hence, some patients will be falsely labeled and may receive unnecessary treatment. This problem is not unique to multiple sclerosis; new enzyme tests, such as troponins (2); DNA methods in microbiology, such as polymerase chain reaction tests (3); new cardiac hormone tests, such as B-type natriuretic peptide for heart failure (4); and new imaging methods, such as magnetic resonance imaging and positron emission tomography in neurologic and musculoskeletal diseases (5), are changing the spectrum of patients considered to have specific diseases. These changes have led to discussion and dispute about how the additional cases detected should be managed. As our diagnostic armamentarium continues to expand and improve, this dilemma will increasingly challenge us.

One trigger for changing the reference test is known flaws in the current reference test. For example, the insensitivity of culture for viruses and difficult-to-culture bacteria created interest in polymerase chain reaction methods, and problems with diagnosing diastolic heart failure created interest in alternatives to echocardiography. A second trigger is when a new technology, such as magnetic resonance imaging or positron emission tomography, suggests new diagnostic possibilities or stages of disease not previously perceived. This need or perception does not necessarily translate into greater diagnostic accuracy that leads to greater clinical benefits. Of course, not all new diagnostic technologies will lead to reconsideration of the reference test. Some may be only improvements on a triage test, or may be less invasive or more convenient replacements of other tests (6). But tests that do change our definition of disease require more than a consideration of accuracy; they also require assessment of the clinical consequences of the change.

What principles should guide replacement of a current clinical reference test? Investigators may claim that the new test is better than existing reference standard tests on theoretical grounds, or is more reliable, or has better sensitivity. If the current reference standard is flawed, then we have no obvious means of deciding whether the proposed new reference standard is really better. Most previous work on this topic has considered the problem of estimating the accuracy of the new test (7) by using such methods as combined tests (using several tests as the reference standard) or discrepant analysis (8–10) (use of a third test to resolve disagreements between the 2 tests). However, estimation of accuracy is not essential in the decision to adopt a new reference standard (11). Any principles or criteria for replacing the current reference standard need to adequately assess the consequences of the switch, both nosologically and clinically.

We aim to set out the criteria and principles for accepting a new test as a better reference standard or component of a reference standard. The 3 key principles that assist with deciding on the value of the new test are as follows.

See also:

Print

Key Summary Points 817
Related article..... 777

Web-Only

Conversion of graphics into slides

Key Summary Points

Whether a new test (such as polymerase chain reaction) should replace a current reference test (such as culture) can be determined by a “fair umpire” test.

A fair umpire test may have errors even larger than the tests under evaluation. What makes it a fair umpire is that its errors are independent of those of both tests.

Possible umpires include causal exposures, concurrent testing, prognosis, or response to treatment.

To decide whether the new test or current reference standard test is more accurate requires the fair umpire to be applied only to cases in which these tests’ results differ.

1. The consequences of the new reference test can be understood through the disagreements between the old and new reference tests.

2. Resolving the disagreements between old and new test requires a fair, but not necessarily perfect, “umpire” test.

3. Possible umpire tests include causal exposures, concurrent testing, prognosis, or the response to treatment.

We examine each principle in turn and then discuss how they may work together.

DISAGREEMENTS BETWEEN OLD AND NEW TESTS

A new reference standard test may lead to a diagnostic spectrum shift by either broadening or narrowing a diagnosis. Most commonly, an apparently more sensitive test broadens the diagnosis by detecting earlier or less consequential cases. Troponins in chest pain and magnetic resonance imaging in breast disease or suspected multiple sclerosis are typical examples of this. Less commonly, the test will narrow the diagnosis by excluding cases that were previously diagnosed, either because the disease was inconsequential or the reference test yielded an incorrect diagnosis. For example, nerve conduction studies narrowed the range of patients considered to have carpal tunnel syndrome. Narrowing also occurs when specific diagnoses are removed from a broad nonspecific category, such as when antibody testing led to reclassification of some patients previously labeled as having the irritable bowel syndrome (12).

To understand the consequences of switching reference tests, we may use the following principle to focus investigation:

Principle 1: The consequences of the new reference test can be understood through the disagreements between the old and new reference tests.

The shaded cells in **Table 1** show these consequences through the hypothetical comparison of a new and old

reference standard test. The consequence of switching from the old to the new reference test is that the spectrum of disease in the patients being treated shifts from cell C (old test positive; new test negative) to cell B (old test negative; new test positive).

The possible disagreements in **Table 1** may be divided into 2 simpler cases (**Table 2**). First, the new test may detect extra possible cases. This apparent additional sensitivity of the new test may also involve a shift to the detection of earlier or less severe cases, such as additional cases of myocardial ischemia detected by troponin (2) or cases of celiac disease detected by endomysial antibody (13). However, these extra cases may also be false positives or cases of less severe illness, so careful assessment is required. Second, the new test may detect fewer possible cases. This apparently better specificity of the new test may also involve some shift from earlier or less severe and less consequential cases. For example, concerns about white-coat hypertension have led to increasing use of ambulatory blood pressure monitoring to reduce false-positive diagnoses of hypertension. Similar concerns have been expressed about overdiagnosis of many conditions, particularly from screening techniques (such as lung and breast cancer screening) (14). However, these cases may also be false negatives, and careful assessment is again required. For example, because we would wish to see that persons classified as nonhypertensive by ambulatory blood pressure monitoring have a similar prognosis to those with no hypertension, the clinical consequences for the discordant group need careful assessment.

It is possible to have cases in both cells B and C of **Table 1**. The crucial question is not simply the numbers in the discordant cells (B and C), but the nature of those cases. Are the additional cases detected by the new test actual (consequential) cases of the disease or merely false-positive results? And are the cases not detected by the new test actual (consequential) cases of the disease (that is, is the test yielding false-negative results?) If the numbers were equal but the additional cases detected by the new test (cell B) were more serious cases that had a greater gain from treatment than the missed cases (cell C), then the new test would provide a net benefit. Conversely, if the additional

Table 1. Possible Consequences of Switching from Old to New Reference Tests

New Reference Test	Old Reference Test	
	Positive	Negative
Positive	A. Agreement; no change in management	B. Apparent cases detected only by the new test
Negative	C. Apparent cases detected only by the old test	D. Agreement; no change in management

Table 2. Performance of New Test if (I) More Sensitive or (II) More Specific

New Reference Test	Old Reference Test	
	Positive	Negative
I. New test possibly more sensitive		
Positive	A. Agreement	B. Apparent new cases
Negative	C. Nil	D. Agreement
II. New test possibly more specific		
Positive	A. Agreement	B. Nil
Negative	C. Apparent non-cases	D. Agreement

cases were less consequential and benefited less from treatment, then the shift would be undesirable.

Deciding on the balance of consequences of the new test is not simple. A randomized trial may be required to demonstrate the value (or nonvalue) of treatment in patients in cells B and C. For example, the introduction of troponins has altered our ability to detect ischemic myocardial damage but led to uncertainty about appropriate treatments for this new spectrum of patients. This uncertainty led to several randomized trials in non-ST elevation myocardial infarction, which, when pooled, suggest that early revascularization is beneficial (15), at least for some subgroups (16). Other examples or scenarios may be simpler, such as the identification of additional surgically correctable epileptic foci by positron emission tomography (5), and well-documented before–after case series may be sufficient (17).

Randomized trials will only need to be considered for patients in the discordant cells (cells B and C in **Table 1**). However, such trials are sometimes unnecessary (18) or not feasible. We should begin with an examination of any known flaws in the current reference test (19) to see whether these might explain the differences. This is a useful start but alone is insufficient, and the next 2 principles provide guidance on what other methods may help us decide whether the new test is better.

THE “FAIR UMPIRE” PRINCIPLE

For the cases in the discordant cells in **Table 1**, we need to decide which test is better. This requires a third test to serve as a referee, or umpire, for disagreements. At first, this would seem impossible, as any other test is likely to be inferior to the competing tests. However, our second principle suggests that perfection is not needed.

Principle 2: Resolving the disagreements between old and new tests requires a fair, but not necessarily perfect, umpire test.

Ideally, the umpire test would discriminate perfectly between persons who truly have the disease and those who do not. In reality, the umpire test will be imperfect. How-

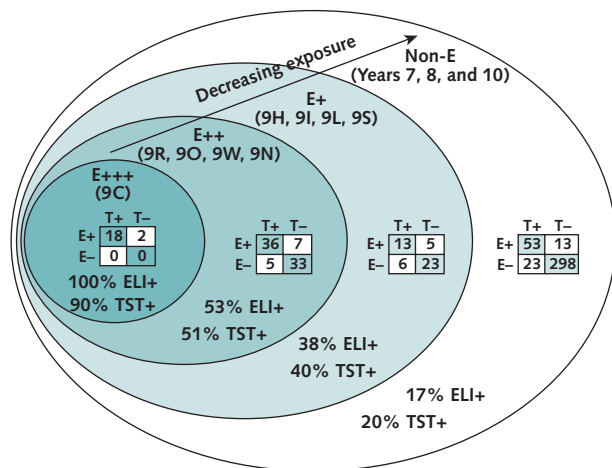
ever, if the umpire is fair—that is, it does not systematically favor either test—then it may be able to accurately distinguish the better reference standard even if it is less accurate than either the old or new test. The umpire test can be thought of as comprising 2 referees: a perfect discriminator and a coin toss, each of which judges half the cases. Provided that a sufficient number of patients are tested, the umpire test will be able to tell which test is better. The umpire test must also meet 2 crucial criteria: It must have some ability to discriminate between disease and non-disease cases, and it must be unbiased (its errors must be conditionally independent of the new and old tests). The conditional independence required here is that regardless of the true disease state of a patient, the results of either reference test being judged do not influence the results of the umpire test. For example, several choices are available to serve as an umpire test for clinical criteria versus magnetic resonance imaging for multiple sclerosis. Computed tomography or alternative magnetic resonance imaging methods, such as magnetic resonance spectroscopy or diffusion-weighted imaging, are likely to make related misclassifications and hence not be conditionally independent, whereas future prognosis is more likely to be conditionally independent. This conditional independence will usually be difficult to demonstrate and may need to be argued from the biological principles underlying the testing process.

EXAMPLE: A NEW TEST IN A TUBERCULOSIS OUTBREAK

The tuberculin skin test (the reference test for latent tuberculosis infection) is known to be compromised, particularly in terms of specificity, because *Mycobacterium bovis* bacille Calmette–Guérin vaccination and environmental mycobacterial exposure can lead to false-positive results. The test may also fail to detect disease in patients with compromised immune response. Interferon- γ enzyme-linked immunospot (ELISpot) assays for early secretory antigenic target-6 and culture filtrate protein-10 gene products, which are specific for *M. tuberculosis* infection and not *M. bovis*, have the potential to replace the tuberculin skin test as a reference test.

Ewer and colleagues (20) describe a school-based cohort study of latent tuberculosis infection in a disease outbreak situation. After identification of the index case of tuberculosis, tuberculin skin tests and ELISpot tests were performed in 535 children in the school to diagnose latent tuberculosis infection. The tests generally agreed but were discordant in some cases (**Figure 1**). Tuberculosis exposure might be an imperfect but fair umpire for these tests. Likelihood of latent tuberculosis infection depends on proximity and duration of exposure to an infectious case; investigators therefore measured proximity of exposure to the disease by classifying children according to the amount of time they spent in the same room as the index case during his infectious period. This depended on where they were in the same school year (9C), whether they shared lessons

Figure 1. Results of 2 tests for tuberculosis, stratified by exposure to the index case.



The tuberculin skin test (*TST*) is the old test, and the enzyme-linked immunospot (*ELI*) assay is the new test. E+++ = class of index case; E++ = classes of students who regularly shared classes with the index case; E+ = students in 4 classes of the same year who did not regularly share classes with the index case; Non-E = students in different years. (Reproduced from Ewer and colleagues [20], with permission of *The Lancet*.)

(9R, 9O, 9W, and 9N) or only weekly activities (9H, 9I, 9L, and 9S), and whether they were in a different school year (Figure 1). The investigators measured duration of exposure for the 148 students in year 9 by identifying hours of lessons shared with the index case through class lists and the school timetable.

The better test here would be the one that more strongly correlated with exposure. The ELISpot showed the stronger gradient (Figure 1) with proximity of exposure, with 100% positive in the highest exposure category and 17% in the lowest exposure category, compared with 90% and 20%, respectively, for the tuberculin skin test. Formal statistical analysis of these trends (exploiting the paired nature of the data) showed that the difference could not be explained by chance ($P = 0.03$) (20). Analysis of the data on duration of exposure showed a larger and more significant difference in the same direction ($P = 0.007$).

This example illustrates the use of exposure as an imperfect but unbiased reference standard to prove the superior value of a new technology, particularly in terms of specificity. Further work (21) has since validated the test against the subsequent clinical outcome of development of tuberculosis. Follow-up of a cohort of 908 child tuberculosis contacts, of whom 15 developed active tuberculosis, determined that the incidence rate was 21 per 1000 person-years for children with positive ELISpot results and 17 for children with positive tuberculin skin test results. In addition, because of ELISpot's higher specificity, targeting preventive treatment to children with positive ELISpot results would have required

treatment of fewer children than relying on tuberculin skin test results while preventing a similar number of tuberculosis cases. Confirming whether the prognostic power of ELISpot is significantly higher than the tuberculin skin test will require larger studies that identify more cases of clinical disease.

THE TESTS OF TIME AND TREATMENT

Unlike the competing old and new reference tests, which are needed for immediate diagnosis, the umpire test need not be confined to the same narrow time window. In particular, both longer follow-up and exploration of previous exposures are feasible for evaluating tests but usually not for making a diagnosis. So, our final principle is:

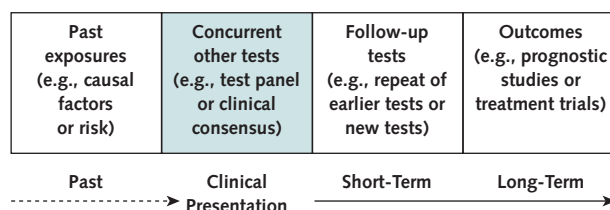
Principle 3: Possible umpire tests include causal exposures, concurrent testing, prognosis, or response to treatment.

Accepting a new reference standard is essentially equivalent to accepting a new definition of the disease. Hence, we need to explore the consequences of that shift in definition. Two important issues with cases added or dropped by the new reference standard are their taxonomic similarity and their similarity in clinical behavior, or the prognosis and the response to treatment. The taxonomic similarity is based on features present at the time of presentation, including possible etiologic factors. Although the similarity of presenting pattern is important, the overall purpose of diagnosis is to guide management, that is, prognostic advice and treatment. Hence, clinical behavior is the more important guiding principle.

As suggested by Figure 2, we may refine these issues of pattern and behavior further by considering the development of a disease over time: past exposure, current features, or future measurements or outcomes. The measures suggested in Figure 2 may come from a single study, but more often will come from a portfolio of linked studies. In these stages and studies, the focus should be on the extra cases (the shaded cells B and C in Table 2) and on whether the extra cases are similar in 4 ways:

1. Causation: Are the etiologic agents or risk factors similar?
2. Clinical pattern at presentation: Are the symptoms, signs, and test results similar?

Figure 2. Possible comparison points for additional cases.



3. Natural history or prognosis: Are the development and consequences of the disease over time similar?

4. Response to treatment: Do cases respond in the same way to standard treatments?

Although response to treatment is usually the most critical criterion, it is insufficient because we want the diagnosis to be useful in guiding treatment; in this case, we also want clinical similarity. For example, a grand mal seizure and acute anxiety may both respond to benzodiazepines, but we would not consider these disorders to be the same diagnosis because their causes, clinical patterns, and prognosis are dissimilar. Likewise, anxiety and thyrotoxicosis may look similar, but they are separate diagnoses because they differ in cause, prognosis, and response to treatment. Although evidence on all 4 criteria may not be possible, these examples suggest that we should generally require evidence on at least 2 of the criteria to appropriately triangulate.

COMBINING THE CRITERIA

The precise choice from the above criteria depends in part on the reason for choosing a new reference test. For instance, if the main aim is prognosis, then the prognostic implication of changing the reference test is essential. Consider suspected amyotrophic lateral sclerosis, for which no treatment, clear cause, or definitive means of diagnosis are known. Because the label has significant consequences, any new test for the condition should improve our ability to predict who will progress in the expected pattern of amyotrophic lateral sclerosis. Thus, the central 2 criteria would be the second and third criteria of pattern and prognosis.

In all cases, the evaluation will be an imperfect measure of the true disease state. For example, we find for a disease with a low fatality rate that the additional cases (Table 1, cell B) have a fatality rate of 4% compared with 1% for the missed cases (Table 1, cell C). This higher fatality rate suggests the new test is probably a better choice; however, outcomes do not differ for most cases in either cell. For example, B-type natriuretic peptide has been shown to be a strong prognostic marker in heart failure. Echocardiography detects similar cases of heart failure, but B-type natriuretic peptide seems to be a better predictor of prognosis than echocardiography, suggesting that it is a better reference test for heart failure. Unfortunately, the comparison of outcomes is confounded by treatment—patients with poor left ventricular ejection fraction but normal B-type natriuretic peptide levels have been more likely to be treated than those with normal left ventricular ejection fraction but high B-type natriuretic peptide levels, which distorts the comparison of the discordant cells.

DISCUSSION

The 3 principles above will help guide our evaluation of the consequences of a new versus an old reference test. Although we have emphasized the clinical consequences of

a change in the reference test, other differences between the new and old test will clearly enter decision making. Such differences will include the perspectives of policymakers (such as costs), clinicians' views about direct harms (such as invasiveness or radiation), and patients' views about these and other issues (such as discomfort and harms).

For a given target condition, we may choose different reference tests depending on whether we need them for research or policy purposes, such as population-based disease surveillance, or clinical purposes, such as screening, prognosis, or treatment decisions of individual patients. For example, the research definition of the chronic fatigue syndrome includes a 6-month minimum duration, but a clinician seeing someone with 5 months of fatigue might rightly wish to start graded exercise, cognitive therapy, or other effective treatments.

For clinical purposes, reference standards would be either the test usually used to confirm a diagnosis, such as histology or culture, or the final arbiter test used in unclear cases, such as pulmonary angiography for pulmonary embolism. Such selective use will make the switch of reference standard complex, because the new test may not simply replace the old test but may also be used for a wider variety of clinical situations. For research purposes, a diagnostic reference standard is used for standardization across research settings and times and may not be suitable for clinical purposes. For example, clear definitions of chronic fatigue, posttraumatic stress disorder, or asthma are needed for comparison of prevalence across time, but these definitions may exclude some patients with the potential to benefit from treatment and are not clinically appropriate. Hence, application of the 3 principles may differ depending on the purpose. A further complication is that both research and clinical reference standards may be a composite (22) of several tests; for example, most psychiatric illness is defined by multiple features and myocardial infarction is usually defined by a combination of clinical, echocardiographic, and cardiac enzyme criteria.

Previous work on the problem of new tests pointed out that “discrepant cell” analysis—in which a third test is used to resolve disagreements to then estimate the sensitivity and specificity—will usually lead to biased and optimistic estimates (8). This bias occurs even if the umpire test (the third test) is unbiased because discrepant cell analysis ignores the errors when both the new and old tests are (incorrectly) in agreement. This bias creates problems for estimation of absolute accuracy of a test but is less of an issue for relative accuracy; unfortunately, it has also led to a general reluctance to examine the discordant cells. Our 3 principles are not aimed at absolute estimation of accuracy, but rather focus on the discordant cells to assess the consequences of adopting the 2 possible competing reference tests, recognizing that both may be imperfect. This focus is similar to Lord and colleagues' suggestion (23) that trials of new diagnostic tests might usefully focus on the additional positive cases detected by a new test. In addition, we sug-

gest that the case for the new test needs to be made on a portfolio of studies. These studies will usually begin with etiologic exposure and concurrent testing but should aim for longer-term follow-up of the patients studied (prognostic cohort studies) and sometimes initiate trials in patients for whom the tests disagree. A single element will rarely be sufficient to persuade us of the superiority of the new test.

Advances in diagnostic technology will continue, and some will lead to changes in the definition of disease. We suggest that the main problem to consider, guided by 3 clarifying principles, is the consequences of any shift in which patients will be treated.

From the University of Oxford, Oxford, United Kingdom; University of Sydney, Sydney, New South Wales, Australia; and University of Birmingham, Edgbaston, Birmingham, United Kingdom.

Acknowledgment: The authors thank Gordon Guyatt, Ajit Lavani, Sally Lord, Jenny Doust, and Chris Hyde for their helpful comments on drafts.

Grant Support: In part by funding from a UK National Institute for Health Research program grant and from the Australian National Health and Medical Research Council Program grant 402764 to the Screening and Test Evaluation Program.

Potential Financial Conflicts of Interest: None disclosed.

Requests for Single Reprints: Paul Glasziou, MB, BS, PhD, Centre for Evidence-Based Medicine, Department of Primary Health Care, University of Oxford, Oxford OX3 7LF, United Kingdom; e-mail, paul.glasziou@dphpc.ox.ac.uk.

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

References

- Whiting P, Harbord R, Main C, Deeks JJ, Filippini G, Egger M, et al. Accuracy of magnetic resonance imaging for the diagnosis of multiple sclerosis: systematic review. *BMJ*. 2006;332:875-84. [PMID: 16565096]
- Katz IA, Irwig L, Vinen JD, March L, Wyndham LE, Luu T, et al. Biochemical markers of acute myocardial infarction: strategies for improving their clinical usefulness. *Ann Clin Biochem*. 1998;35 (Pt 3):393-9. [PMID: 9635105]
- Ramers C, Billman G, Hartin M, Ho S, Sawyer MH. Impact of a diagnostic cerebrospinal fluid enterovirus polymerase chain reaction test on patient management. *JAMA*. 2000;283:2680-5. [PMID: 10819951]
- Doust JA, Pietrzak E, Dobson A, Glasziou P. How well does B-type natriuretic peptide predict death and cardiac events in patients with heart failure: systematic review. *BMJ*. 2005;330:625. [PMID: 15774989]
- Maehara T. Neuroimaging of epilepsy. *Neuropathology*. 2007;27:585-93. [PMID: 18021381]
- Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. *BMJ*. 2006;332:1089-92. [PMID: 16675820]
- Rutjes AW, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PM. Evaluation of diagnostic tests when there is no gold standard. A review of methods. *Health Technol Assess*. 2007;11:iii, ix-51. [PMID: 18021577]
- Hadgu A. The discrepancy in discrepant analysis. *Lancet*. 1996;348:592-3. [PMID: 8774575]
- Hadgu A. Discrepant analysis: a biased and an unscientific method for estimating test sensitivity and specificity. *J Clin Epidemiol*. 1999;52:1231-7. [PMID: 10580787]
- McAdam AJ. Discrepant analysis: how can we test a test? *J Clin Microbiol*. 2000;38:2027-9. [PMID: 10834948]
- Irwig LM, Groeneveld HT, Pretorius JP, Hnizdo E. Relative observer accuracy for dichotomized variables. *J Chronic Dis*. 1985;38:899-906. [PMID: 4055978]
- Sanders DS, Carter MJ, Hurlstone DP, Pearce A, Ward AM, McAlindon ME, et al. Association of adult coeliac disease with irritable bowel syndrome: a case-control study in patients fulfilling ROME II criteria referred to secondary care. *Lancet*. 2001;358:1504-8. [PMID: 11705563]
- Catassi C, Fabiani E. The spectrum of coeliac disease in children. *Baillieres Clin Gastroenterol*. 1997;11:485-507. [PMID: 9448912]
- Irwig L, Houssami N, Armstrong B, Glasziou P. Evaluating new screening tests for breast cancer [Editorial]. *BMJ*. 2006;332:678-9. [PMID: 16565097]
- Hoening MR, Doust JA, Aroney CN, Scott IA. Early invasive versus conservative strategies for unstable angina & non-ST-elevation myocardial infarction in the stent era. *Cochrane Database Syst Rev*. 2006;3:CD004815. [PMID: 16856061]
- O'Donoghue M, Boden WE, Braunwald E, Cannon CP, Clayton TC, de Winter RJ, et al. Early invasive vs conservative treatment strategies in women and men with unstable angina and non-ST-segment elevation myocardial infarction: a meta-analysis. *JAMA*. 2008;300:71-80. [PMID: 18594042]
- Glasziou P, Chalmers I, Rawlins M, McCulloch P. When are randomised trials unnecessary? Picking signal from noise. *BMJ*. 2007;334:349-51. [PMID: 17303884]
- Bossuyt PM, Lijmer JG, Mol BW. Randomised comparisons of medical tests: sometimes invalid, not always efficient. *Lancet*. 2000;356:1844-7. [PMID: 11117930]
- Lijmer JG, Mol BW, Heisterkamp S, Bossel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA*. 1999;282:1061-6. [PMID: 10493205]
- Ewer K, Deeks J, Alvarez L, Bryant G, Waller S, Andersen P, et al. Comparison of T-cell-based assay with tuberculin skin test for diagnosis of Mycobacterium tuberculosis infection in a school tuberculosis outbreak. *Lancet*. 2003;361:1168-73. [PMID: 12686038]
- Bakir M, Millington KA, Soysal A, Deeks JJ, Efee S, Aslan Y, et al. Prognostic value of a T-cell-based, interferon- γ biomarker in children with tuberculosis contact. *Ann Intern Med*. 2008;149:777-786.
- Alonzo TA, Pepe MS. Using a combination of reference tests to assess the accuracy of a new diagnostic test. *Stat Med*. 1999;18:2987-3003. [PMID: 10544302]
- 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]

Current Author Addresses: Dr. Glasziou: Centre for Evidence-Based Medicine, Department of Primary Health Care, University of Oxford, Oxford OX3 7LF, United Kingdom.

Dr. Irwig: Screening and Test Evaluation Program, School of Public Health, and University of Sydney, Sydney, New South Wales 2006, Australia.

Dr. Deeks: Unit of Public Health, Epidemiology, and Biostatistics; University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom.