Sources of bias in diagnostic accuracy studies and the diagnostic process

Toshi A. Furukawa, Gordon H. Guyatt

∞ See related article page 469

Making accurate diagnoses requires knowing the performance characteristics of the tests we use. How can we know whether to trust studies that purport to establish those performance characteristics?

The fundamental design of studies assessing the accuracy of diagnostic tests involves a comparison of the test under investigation with a "gold" or reference standard. In many cases, the gold standard is impractical in the clinical setting because it is too invasive (e.g., pulmonary angiography), is too costly or time-consuming (e.g., long-term follow-up for diagnosis of seizures or syncope) or is possible only after the patient is dead (neuropathological examination for Alzheimer's disease). The test under investigation can be any method of obtaining information from the patient: clinical history taking, physical examination, laboratory investigation, imaging, questionnaire or pathological examination. Investigators assess the accuracy of the test by the extent of its agreement with the gold standard. One can best express this agreement in the form of likelihood ratios, although a number of less satisfactory alternatives (sensitivity and specificity, diagnostic odds ratios, or areas under ROC [receiver-operating-characteristic] curves) are also available.

There are many variations possible to this prototypical diagnostic accuracy study, which may or may not introduce biases (i.e., systematic over- or underestimation of the true accuracy of a diagnostic test). Critical examination of diagnostic accuracy studies started around 1980, and we now have some 90 quality-assessment tools,[1] the most notable of which include the STARD (Standards for Reporting of Diagnostic Accuracy)[2] and the QUADAS (Quality Assessment of Diagnostic Accuracy Studies).[3] The STARD is a checklist of 25 items and flow diagram to improve the reporting of original studies, whereas the QUADAS is a list of 14 checkpoints for assessing the quality of original studies to include them in systematic reviews. The authors of both checklists explicitly acknowledge that our understanding of the determinants of the quality of diagnostic accuracy studies is growing, and they intend to update their products accordingly.

In this issue (page 469), Rutjes and colleagues report on their investigation of 15 potential sources of bias in diagnostic accuracy studies by empirically examining their association with estimates of diagnostic accuracy in each study.[4] Their work complements previous empirical works by groups led by Lijmer[5] and Whiting.[6] Both Rutjes and colleagues[4] and Lijmer and associates[5] sampled primary studies from systematic reviews, a sampling frame that is likely to be systematically different from the total pool of diagnostic accuracy studies (and may include stronger or weaker original investigations). Whiting and colleagues,[6] on the other hand, systematically searched for both primary and secondary studies that examined influences of design features on estimates of diagnostic accuracy; the limitation of their work is that they provided narrative, rather than quantitative, summaries of their results.

The 3 studies have identified a number of design features that consistently appear to introduce bias (see online Appendix 1 at www.cmaj.ca/cgi/content/full/174/4/481/DC1). These include poor selection of patients with and without the target diagnosis (spectrum bias); failure to use the same gold standard on all patients, irrespective of their test results (verification bias); and awareness of the results of the gold standard or test by those interpreting the results of the other (lack of blinding). Clinicians should pay special attention to these aspects of diagnostic accuracy studies before deciding to apply the results to their own practice, regardless of whether they are reviewing a single primary study or a systematic review of many such studies.[1]

We should not ignore potential sources of bias even when supporting empirical data are lacking.

The limitations of the 3 reviews suggest that we should not ignore potential sources of bias even when supporting empirical data are lacking. For instance, empirical support for bias associated with a poor choice of gold standard is limited (see online Appendix 1 at www.cmaj.ca/cgi/content/full/174/4/481/DC1). Nevertheless, common sense tells us that, if the gold standard lacks reliability or validity, assessments of test properties will be pointless. Malignant melanoma provides a recent example of the unreliability of the accepted gold standard: a nationwide sample of practising pathologists in the United Kingdom agreed only modestly in the diagnosis of malignant melanoma without the use of standardized diagnostic criteria (kappa = 0.45). Satisfactory agreement was achieved when standardized diagnostic criteria were used and some disease categories were modified.[7]

DOI:10.1503/cmaj.060014

To establish that studies have provided unbiased estimates of test properties, and that a particular test is indeed accurate, is but one step in the diagnostic process. In order to better serve our patients, we need to generate appropriate differential diagnoses with a sense of the relative likelihood of competing alternatives. Ideally, we will think in quantitative terms, generating estimates of pretest probabilities. In addition, we must understand the sequential application of each diagnostic test. In practice, the diagnostic process proceeds in a stepwise fashion, starting with a particular set of signs and symptoms. Clinical prediction rules or guides, which simultaneously consider a range of diagnostic information, provide one solution to incorporating the results of multiple tests to arrive at an accurate diagnosis.[8]

Toshi Furukawa is with the Department of Psychiatry and Cognitive-Behavioral Medicine, Nagoya City University Graduate School of Medical Sciences, Nagoya, Japan, and Gordon Guyatt is with the Departments of Internal Medicine and of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ont.

Competing interests: None declared.

REFERENCES

1. Whiting P, Rutjes AW, Dinnes J, et al. A systematic review finds that diagnostic reviews fail to incorporate quality despite available tools. *J Clin Epidemiol* 2005;58:1-12.
2. Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clin Chem* 2003;49:7-18.
3. Whiting P, Rutjes AW, Reitsma JB, et al. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol* 2003;3:25.
4. Rutjes AWS, Reitsma JB, Di Nisio M, et al. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ* 2006;174(4):469-76.
5. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;282:1061-6.
6. Whiting P, Rutjes AW, Reitsma JB, et al. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Ann Intern Med* 2004;140:189-202.
7. A nationwide survey of observer variation in the diagnosis of thin cutaneous malignant melanoma including the MIN terminology. CRC Melanoma Pathology Panel. *J Clin Pathol* 1997;50:202-5.
8. McGinn T, Guyatt G, Wyer P, et al. Clinical prediction rules. In: Guyatt G, Rennie D, editors. *Users' guides to the medical literature: a manual for evidence-based clinical practice*. Chicago: AMA Press; 2002.

Correspondence to: *Dr. Toshi A. Furukawa, Professor, Department of Psychiatry and Cognitive-Behavioral Medicine, Nagoya City University Graduate School of Medical Sciences, Mizuho-cho, Mizuho-ku, Nagoya, 467-8601 Japan; furukawa@med.nagoya-cu.ac.jp*