RESEARCH METHODS & REPORTING

Target practice: choosing target conditions for test accuracy studies that are relevant to clinical practice

Clinicians should seek information about how well a new test detects disease that will benefit from clinical intervention rather than simply the presence of any disease

S J Lord *research fellow*^{1,2}, L P Staub *PhD candidate*¹, P M M Bossuyt *professor of clinical epidemiology*³, L M Irwig *professor of epidemiology*²

¹National Health and Medical Research Council Clinical Trials Centre, University of Sydney, 92-94 Parramatta Road, Locked Bag 77, Camperdown NSW 2050, Australia; ²Screening and Test Evaluation Program, School of Public Health, University of Sydney; ³Department of Clinical Epidemiology and Biostatistics, Academic Medical Centre, University of Amsterdam, Amsterdam, Netherlands

Test accuracy varies depending on how the presence of disease is defined. For example, the sensitivity of computed tomographic (CT) colonography to detect colorectal neoplasia has been estimated at 96% for detecting invasive cancer, 86% for medium to large polyps, but as low as 45% when polyps of all sizes are included in the definition of disease.¹ To interpret these results, we need to decide whether all polyps, or what type of polyps, are important to detect.

In this paper, we explain how this principle applies when reading studies of test accuracy. When a disease threshold or set of criteria is available to define a clinically meaningful subset of disease, estimates of test accuracy for detecting the entire spectrum of disease will not apply to this subset. Therefore, clinicians need to look for estimates of test accuracy for detecting classifications of disease that are useful in clinical practice. The clinical value of a test cannot be interpreted from estimates of test accuracy if the disease definition is not clearly stated or if its clinical importance is ambiguous or unknown.

Defining disease for test accuracy studies

The diagnostic accuracy of a test measures how well it distinguishes between the presence and absence of disease. This is traditionally expressed as the sensitivity and specificity of the test. Test sensitivity is the proportion of patients with disease who are correctly identified. Test specificity is the proportion of patients without disease who are correctly identified.

To estimate sensitivity and specificity, accuracy studies require an explicit definition of disease as a dichotomous “present or absent” outcome and a reference standard that can be used to verify the true disease status (box). Typically, the definition of disease is based on the best available information about the pathological or molecular basis of disease. For example, trisomy 21 is defined by the presence of an additional chromosome 21. Here, the diagnostic threshold is reasonably clear: everyone

with this characteristic is classified as having the syndrome; those without are excluded. Accuracy studies of screening tests for trisomy 21, such as ultrasound assessment of fetal nuchal translucency, measure how well the tests can correctly classify individuals according to this binary definition of disease using chromosome studies as the reference standard.

However, clinical disease is often not binary: there is no clear fixed threshold to distinguish between the presence and absence of disease. Furthermore, cases vary in severity and clinical consequences and are rarely managed as a single condition. Since test accuracy varies depending on what spectrum of disease is examined, its accuracy in detecting all disease may not reflect accuracy in detecting clinically relevant cases at one end of this spectrum.

The figure² shows the common scenario where a disease has a broad spectrum of pathological presentations, ranging from mild to severe based on histology, extent, or location. For example, the presence of carotid artery stenosis can refer to all stenotic lesions, from mild plaques to fully occluded arteries. At a clinical level, disease is usually further classified based on evidence about differences in patient prognosis or treatment outcomes for subgroups of disease. These subgroups are used to guide decisions about management such as a choice between monitoring and treatment for clinically important cases.

The figure distinguishes between clinically important disease (disease associated with symptoms or a high risk of future clinical events) and disease that will benefit from treatment (disease for which the benefits of a specified treatment exceed the harms). Clinical studies are needed to define the threshold or boundaries for these subgroups. For some diseases these clinical definitions will overlap. For genital *Chlamydia trachomatis* infection, for example, all cases of disease are clinically important and recommended for treatment. For other diseases, different subgroups can be defined.

Example of interpretation of test sensitivity and specificity: computed tomographic colonography to detect colorectal neoplasia in a screening population*Target*

To identify polyps that need further investigation by colonoscopy and biopsy. This requires a threshold for disease to be specified—eg, one or more polyps ≥ 10 mm (research shows high risk of advanced adenoma and progression to cancer for this subgroup)

Accuracy of computed tomographic colonography¹

93% sensitivity—of 100 patients with polyps ≥ 10 mm, 93 will be correctly identified as requiring colonoscopy and biopsy and seven will be misclassified as having normal or clinically unimportant results and will miss out on appropriate investigation and management

97% specificity—Of 100 patients with no polyps ≥ 10 mm, 97 will be correctly classified as having normal or clinically unimportant results and three will be misclassified as having polyps and receive unnecessary further investigation

The target condition

When measuring test accuracy, the target condition is the classification of disease you wish to detect. To define the target condition we must think about the clinical decisions the test will be used to guide and determine the most appropriate threshold or criteria to dichotomise the presence or absence of disease for these decisions. The principle of defining the target condition to represent a clinically relevant classification of disease is not new² and is well recognised for some diseases, such as colorectal neoplasia and carotid artery stenosis.

When test results will be used to rule in or rule out disease from further investigation or management, we need to consider whether all cases of disease are clinically important and, if not, what is the threshold for identifying cases that are clinically important. This is essentially a prognostic question and is best answered by studies that report patient outcomes for subgroups of disease using different disease thresholds. Clinically important carotid artery stenosis, for example, is commonly defined as $\geq 50\%$ narrowing of the artery diameter, based on evidence from prognostic studies about the risk of stroke for patients with this severity of disease compared with those with mild stenosis or no disease.³

When test results will be used to select treatment, the target condition is best defined by treatment trials. For example, trial evidence suggests clear benefits from surgery exceeding harms only for those with 70-99% stenosis.⁴ Thus the target condition for test accuracy can be based on these trial defined boundaries.

Although the target condition and reference standard are sometimes regarded as interchangeable terms, there is an important distinction between them: the target condition represents the classification of disease required to investigate a particular clinical problem (such as important carotid artery stenosis), whereas the reference standard represents the best available method to detect this condition (such as angiography). If more than one target condition is important, test accuracy can be measured separately for each.

The choice of target condition is related to but separate from other factors that limit the application of test research into practice, including the importance of measuring test accuracy in patient groups similar to those who will have the test in practice and reporting separate accuracy estimates for different patient subgroups.^{5 6}

Why target condition should be clinically defined

When the target condition matches the best available evidence based classification of disease for diagnosis or treatment

decisions, the clinical consequences of a true and false positive and negative test result can be clearly appreciated: test sensitivity represents the proportion of patients with the condition who will receive appropriate management; and test specificity represents the proportion of patients without the condition who will avoid further unnecessary tests or treatment. In contrast, if the definition of the target condition is not clearly defined, or its clinical importance is ambiguous, estimates of sensitivity and specificity may underestimate or overestimate the clinical value of the test. Consider human papillomavirus (HPV) testing to screen for precancerous cervical abnormalities. Here the target is high grade cervical intraepithelial neoplasia rather than all HPV infections, many of which do not warrant further investigation. HPV test sensitivity for detecting high grade intraepithelial neoplasia can be directly interpreted in clinical terms; test sensitivity for detecting all HPV infection cannot.⁷

The target condition should always be based on the best available evidence about the threshold or criteria for the intervention that the test will be used to guide. For CT colonography screening for colorectal cancer and pre-malignant polyps, the critical decision is whether to refer for colonoscopy. The risk of advanced adenoma is low for tiny polyps and increases with polyp size.⁸ This evidence has been used to support a referral threshold for polyps ≥ 10 mm or, at most, ≥ 6 mm. Thus the most clinically relevant target for measuring test accuracy is having at least one polyp over the threshold size regardless of (unknown) histology.¹

Unfortunately, poor reporting of the definition of the target condition and its relevance for clinical decisions is widespread in accuracy studies, even for conditions where clinically meaningful subgroups are well defined. In their systematic review of CT colonography for detecting colorectal polyps, Halligan et al found that only half of the accuracy studies (12/24) reported sufficient data to construct a two by two table to calculate per patient sensitivity and specificity according to polyp size.¹ In Wardlaw et al's review of imaging tests for carotid artery stenosis, 41 studies met current standards for reporting accuracy results. Of these, 30 (73%) did not provide data using standard criteria for defining clinically important (50-99%) stenosis, 12 (29%) did not provide data for 70-99% stenosis, and 7 (17%) provided data for neither of these subgroups.⁹

These reviews also show the potential magnitude of bias if test accuracy for detecting a broad disease spectrum is used as a proxy for a more narrowly defined clinically important subgroup (CT colonography test sensitivity: 93% for detecting polyps ≥ 10 mm; 86% for polyps ≥ 6 mm, and as low as 45% for all polyps); or if test accuracy for detecting one segment of the disease spectrum is applied to another segment (Doppler

ultrasound sensitivity 89% for detecting 70-99% carotid artery stenosis; 36% for 50-69% stenosis⁹).

For conditions where the threshold for clinically important disease has not yet been defined, interpreting test accuracy is even more challenging. For example, studies have shown that CT pulmonary angiography is more sensitive than ventilation-perfusion (V/Q) scanning for detecting pulmonary emboli. These studies commonly include all emboli, regardless of location and size, in the definition of disease. However, the clinical importance of small subsegmental pulmonary emboli is uncertain.¹⁰ If some of the extra cases detected by angiography represent this end of the disease spectrum, the improved sensitivity may not translate into improved patient outcomes. Ignoring this uncertainty when interpreting test accuracy results will overestimate the clinical value of angiography.^{10 11} The problems of new diagnostic technologies that are capable of detecting early or milder disease that is of uncertain clinical importance have been well documented in other areas, including cardiovascular and infectious diseases,^{12 13} spinal disorders,¹⁴ and various cancers.^{15 16}

Clinical implications

Current guidelines for reporting and appraising studies of test accuracy do not include guidance about choosing the most relevant target condition.^{17 18} To make evidence based judgments about the clinical value of a test, clinicians need to be clear about what they are seeking to detect. When reading test accuracy studies, they should look for information about what target condition was chosen and why. Studies should include an explicit statement about the clinical consequences of correctly detecting or excluding this target and interpret the results in these terms. If the target condition is not clearly defined or does not match your reason for testing, seek more relevant evidence.

Reports of a new test having higher sensitivity for detecting potentially serious disease are compelling, but interpretation is not straightforward if the new test shifts the threshold for detecting disease. To recognise these situations, readers should consider whether the extra information provided by a more sensitive test will lead to a broader definition of disease than existing tests. If so, the clinical value of the new test will depend on whether the prognosis or response to treatment of the extra cases detected is likely to be similar to that of cases detected by existing tests. When the value is uncertain, authors should be explicit about the lack of evidence. For example, examining the clinical value of adding magnetic resonance imaging (MRI) to CT to rule out spinal injuries in obtunded patients, Schoenfeld et al reported that there are "little objective data correlating many of the identifiable MRI soft tissue abnormalities with the clinical assessment of instability."¹⁴

Research implications

Collaboration between clinicians and researchers is essential to define the clinical role of the test and choose the most appropriate target condition. This dialogue should focus on identifying the major clinical decisions the test will be used to guide and locating the best available evidence to determine the optimal disease threshold or criteria for making this decision.

Clinically defined target conditions will sometimes be substantially different from the traditional pathological definitions of disease. For example, a recent study of imaging strategies for acute abdominal pain used a target condition of patients requiring treatment within 24 hours rather than a traditional pathological diagnosis of appendicitis or other

disease, resulting in more clinically relevant estimates of accuracy.¹⁹

The target condition is likely to change as new tests and treatments are introduced that shift the threshold for clinically important disease and the threshold at which treatment benefits exceed harms. Therefore, we encourage researchers to report test results using different disease thresholds in a multidimensional table. This will allow readers to select accuracy estimates that are most relevant to how they will use the test. Ideally, study data would also be stored in a publicly accessible format that could be re-examined when new indications for tests and treatments arise.

Attempts to define the target condition will often point to the need for clinical studies that better define optimal thresholds for diagnosis and intervention. Consider quantitative polymerase chain reaction testing for cytomegalovirus in patients who have received transplants; these tests are highly sensitive for detecting copies of viral DNA in peripheral blood, but the optimal criteria for starting (expensive) treatment for this devastating disease has not yet been established.¹³

We recognise that even when trial evidence is available to develop standard disease criteria for treatment decisions, in practice the treatment threshold is not fixed. Clinical decisions about treatment thresholds may vary for different patient groups and according to clinician experience, clinician and patient preferences, and resources. Even so, we believe the development of standard evidence based disease classifications is essential to ensure accuracy estimates are meaningful to practice and to promote consistency between studies that will allow better comparisons between tests and over time.

We thank Eveline Staub, Michael Solomon, and Reginald S A Lord for their comments on drafts of this manuscript. We also thank Angela Webster and Benjamin Jonker for providing examples.

Funding: This work was supported through an Australian National Health and Medical Research Council Project Grant (No 571044) and Program Grant (No 402764).

Contributors: PB suggested the target condition as the topic for this paper, and SL and LS drafted the paper. LI led discussions to help develop and structure the ideas presented. The paper builds on earlier work from LI and PB about improving the applicability of test accuracy estimates. All authors made substantial contributions to improve the paper and approved the final version. SL is guarantor.

Competing interests: All authors have completed the ICMJE unified disclosure form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare support from the Australian National Health and Medical Research Council Project Grant (No 571044) and Program Grant (No 402764) for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years, and no other relationships or activities that could appear to have influenced the submitted work.

Provenance and peer review: Not commissioned; externally peer reviewed.

- Halligan S, Altman DG, Taylor SA, Mallett S, Deeks JJ, Bartram CI, et al. CT colonography in the detection of colorectal polyps and cancer: systematic review, meta-analysis, and proposed minimum data set for study level reporting. *Radiology* 2005;237:893-904.
- 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.
- Autret A, Saudeau D, Bertrand PH, Pourcelot L, Marchal C, De Boisvilliers S. Stroke risk in patients with carotid stenosis. *Lancet* 1987;388-90.
- Rothwell PM, Eliasziw M, Gutnikov SA, Fox AJ, Taylor DW, Mayberg MR, et al. Analysis of pooled data from the randomised controlled trials of endarterectomy for symptomatic carotid stenosis. *Lancet* 2003;361:107-16.
- Mulherin SA, Miller WC. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. *Ann Intern Med* 2002;137:598-602.

Summary points

Studies of test accuracy traditionally measure how well the test distinguishes between the presence and absence of disease

Such studies may underestimate or overestimate a test's clinical value if a narrower spectrum of disease is relevant for diagnosis and management decisions

Accuracy studies should use clinically relevant disease as the target condition

Definition of the target condition should be based on evidence from prognostic studies and treatment trials

- 6 Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Ann Intern Med* 2004;140:189-202.
- 7 Snijders PJ, van den Brule AJ, Meijer CJ. The clinical relevance of human papillomavirus testing: relationship between analytical and clinical sensitivity. *J Pathol* 2003;201:1-6.
- 8 Hassan C, Pickhardt PJ, Kim DH, Di GE, Zullo A, Laghi A, et al. Systematic review: distribution of advanced neoplasia according to polyp size at screening colonoscopy.. *Alimen Pharmacol Ther* 2010;31:210-7.
- 9 Wardlaw JM, Chappell FM, Best JJ, Wartolowska K, Berry E, NHS Research and Development Health Technology Assessment Carotid Stenosis Imaging Group. Non-invasive imaging compared with intra-arterial angiography in the diagnosis of symptomatic carotid stenosis: a meta-analysis. *Lancet* 2006;367:1503-12.
- 10 Anderson DR, Kahn SR, Rodger MA, Kovacs MJ, Morris T, Hirsch A, et al. Computed tomographic pulmonary angiography vs ventilation-perfusion lung scanning in patients with suspected pulmonary embolism: a randomized controlled trial. *JAMA* 2007;298:2743-53.
- 11 Glassroth J. Imaging of pulmonary embolism: too much of a good thing? *JAMA* 2007;298:2788-9.
- 12 Morrow DA. Clinical application of sensitive troponin assays. *N Engl J Med* 2009;361:913-5.
- 13 Szczepura A, Westmoreland D, Vinogradova Y, Fox J, Clark M. Evaluation of molecular techniques in prediction and diagnosis of cytomegalovirus disease in immunocompromised patients. *Health Technol Assess* 2006;10:1-176.
- 14 Schoenfeld AJ, Bono CM, McGuire KJ, Warholc N, Harris MB. Computed tomography alone versus computed tomography and magnetic resonance imaging in the identification of occult injuries to the cervical spine: a meta-analysis. *J Trauma* 113;68:109-13.
- 15 Esserman L, Thompson I. Solving the overdiagnosis dilemma. *J Natl Cancer Inst* 2010;102:582-3.
- 16 Welch HG, Black WC. Overdiagnosis in cancer. *J Natl Cancer Inst* 2010;102:605-13.
- 17 Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *BMJ* 2003;326:41-4.
- 18 Whiting P, Rutjes AW, Reitsma JB, Bossuyt P, Kleijnen J. 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.
- 19 Lameris W, van RA, van Es HW, van Heesewijk JP, van RB, Bouma WH, et al. Imaging strategies for detection of urgent conditions in patients with acute abdominal pain: diagnostic accuracy study. *BMJ* 2009;338:b2431.

Accepted: 24 March 2011

Cite this as: *BMJ* 2011;343:d4684

© BMJ Publishing Group Ltd 2011

Figure

Defining the target condition to represent a clinically relevant spectrum of disease