REVIEW 10.1111/1469-0691.12474

Systematic reviews and meta-analyses of diagnostic test accuracy

M. M. G. Leeflang

Clinical Epidemiology and Biostatistics and Bioinformatics, Academic Medical Centre, University of Amsterdam, Amsterdam, the Netherlands

Abstract

Systematic reviews of diagnostic test accuracy summarize the accuracy, e.g. the sensitivity and specificity, of diagnostic tests in a systematic and transparent way. The aim of such a review is to investigate whether a test is sufficiently specific or sensitive to fit its role in practice, to compare the accuracy of two or more diagnostic tests, or to investigate where existing variation in results comes from. The search strategy should be broad and preferably fully reported, to enable readers to assess the completeness of it. Included studies usually have a cross-sectional design in which the tests of interest, ideally both the index test and its comparator, are evaluated against the reference standard. They should be a reflection of the situation that the review question refers to. The quality of included studies is assessed with the Quality Assessment of Diagnostic Accuracy Studies-2 checklist, containing items such as a consecutive and all-inclusive patient selection process, blinding of index test and reference standard assessment, a valid reference standard, and complete verification of all included participants. Studies recruiting cases separately from (healthy) controls are regarded as bearing a high risk of bias. For meta-analysis, the bivariate model or the hierarchical summary receiver operating characteristic model is used. These models take into account potential threshold effects and the correlation between sensitivity and specificity. They also allow addition of covariates for investigatation of potential sources of heterogeneity. Finally, the results from the meta-analyses should be explained and interpreted for the reader, to be well understood.

Keywords: diagnosis, diagnostic test accuracy, evidence-based medicine, meta-analyses, sensitivity and specificity, systematic reviews
Article published online: 26 November 2013
Clin Microbiol Infect 2014; **20:** 105–113

Corresponding author: M. M. G. Leeflang, Clinical Epidemiology and Biostatistics and Bioinformatics, Academic Medical Centre, University of Amsterdam, Room J1b-209, PO Box 22700, 1100 DE Amsterdam, the Netherlands
E-mail: m.m.leeflang@amc.uva.nl

Introduction

Practising evidence-based medicine starts with a clinical question [1]. For example, a general physician might want to know whether testing for papilloma virus can replace cytology for the diagnosis of cervical cancer, as it is cheaper and easier to perform, or a haematologist might wonder whether a molecular test is needed on top of clinical judgement before a patient is treated for invasive fungal diseases. For questions such as these, the sensitivity and specificity of a diagnostic test may be helpful.

Systematic reviews are at the heart of evidence-based medicine. These literature overviews are performed in a systematic and transparent way, and they are explicit about where their study base comes from and how included references were selected. The quality of included studies is assessed and, if appropriate, the results are quantitatively summarized in a meta-analysis. These explicit methods limit bias, and improve the reliability of conclusions [2]. Systematic reviews also enable us to establish whether findings are consistent and can be generalized over different situations.

Healthcare professionals looking for evidence about diagnostic tests may turn to systematic reviews of diagnostic test accuracy. These reviews summarize the sensitivity and specificity of a test, and sometimes other measures as well, such as predictive values, likelihood ratios, ORs, or summary receiver operating characteristic (ROC) curves [3]. Sensitivity is defined as the probability of a person with the disease of interest having a positive test result, and specificity is defined as the probability of a person without the disease having a negative test result. These refer to the clinical situation in

which a test is being used, and are different from analytical sensitivity (referring to the ability of the test to measure low concentrations of a substance) and analytical specificity (usually referring to cross-reactivity). They may also be different from more technical definitions of sensitivity and specificity, such as the ability to distinguish between cases and (often healthy) controls. These analytical and technical measures are important in the earlier phases of test development, whereas clinical sensitivity and specificity are used to indicate the performance of a test in clinical practice [4].

The number of diagnostic test accuracy reviews has rapidly increased, especially over the last 5 years. A quick MEDLINE search revealed that the number of systematic reviews or meta-analyses with diagnosis, diagnostic, test, testing, tests or accuracy in the title grew from 748 at the end of 2008 to 2068 in November 2013. However, readers find it difficult to grasp the concept of a diagnostic test accuracy review, and this may limit their use in practice [5].

This overview describes the steps involved in a diagnostic test accuracy systematic review, while focusing on the link with the clinical question. We hope to explain for readers what they can expect from a diagnostic accuracy review, and how the results of these reviews can be used in clinical and laboratory practice. A selection of 20 diagnostic test accuracy reviews in infectious diseases will serve as an illustration [6–25] (see Appendix). These reviews come from a set of reviews published between September 2011 and January 2012 that we used to survey which meta-analytic methods authors use [26].

Review question

The first and most important step in a systematic review is question formulation. The review question guides the rest of the review: it dictates the relevant study design and study characteristics, the potential biases to be expected, the appropriate meta-analysis technique, and the interpretation of results. The review question includes some basic elements: the patients or population who will undergo the test in practice, the test(s) of interest and comparator test(s), and the target condition or disease of interest, as defined by the reference standard. When papilloma virus testing is compared with cytology, the patient population consists of women who will be tested for cervical cancer. The test of interest is called the index test, here being virus testing. Its comparator in this case is cytology. The disease of interest is cervical cancer; the term target condition refers to a more specific definition, e.g. a specific stage of cancer, or treatable cancer [27].

A key element in diagnostic accuracy is the reference standard. This is the test used to define the target condition, and the underlying assumption is that it reflects the truth. For cervical cancer, a valid reference standard is histopathology. By design, the reference standard is assumed to be flawless. The reference standard sets the reference, and sensitivity and specificity are expressed as the proportion of reference standard positives with a positive index test result, and the proportion of reference standard negatives with a negative index test result, respectively. It is therefore impossible to show that an index test is better than the reference standard, even if this would be the case in reality.

To place the review question in a context and to enable better interpretation of the results, the place of the test(s) in the diagnostic pathway should be described [28]. It matters whether the test is used as a first-line test to decide who should be referred for further testing, or whether the test will be used to start treatment on the basis of the test result. If a test is used as a first-line test, then the composition of the sample and the consequences of a false-positive or false-negative test result will be different from those in a more specialized situation. A first-line test, also called a triage test, may be useful even when the sensitivity or specificity is not high, depending on the steps that will be taken after testing. If the test is used to determine who should be treated and who should not be treated, it will be important to not miss any diseased patients (requiring high sensitivity), and it may be also be important to prevent the treatment of non-diseased persons (requiring high specificity), especially when the treatment is invasive or burdensome. A systematic review on molecular assays for neonatal sepsis aimed to investigate whether the sensitivity of these assays would be higher than 98% and the specificity higher than 95%, based on the balance between missing almost no neonate with sepsis and overtreatment of neonates without sepsis [19]. Authors may find it difficult to firmly state a minimally accepted sensitivity and specificity beforehand. Alternatively, one could hypothesize that the sensitivity and specificity in the current study should at least be as high as previously reported, or that the sensitivity and specificity of the index test should at least be as high as those of the comparator test(s).

An important secondary objective of a diagnostic test accuracy review is to investigate potential sources of heterogeneity. How do the sensitivity and specificity of a test differ between adults and children, or between primary care and secondary care, or between different subtypes of the test? For example, the objective of a systematic review on antigen tests for tuberculosis was to estimate the diagnostic accuracy of antigen detection tests using different clinical specimens in adults and children with and without human immunodeficiency virus infection [9].