This Chapter is Draft Version 0.4

This Chapter is undergoing editing and review, and thus does not constitute the final agreed formal directions for Cochrane Reviewers. However, the editors believe that it is close enough to completion for it to be helpful and not to be seriously misleading, although further sections, clarifications and some changes may still be added. It will be superseded by Version 1.0 in due course.

Please cite this version as

Bossuyt PM, Leeflang MM. Chapter 6: Developing Criteria for Including Studies. In: *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy* Version 0.4 [updated September 2008]. The Cochrane Collaboration, 2008.

This statement was added to this Chapter by Jon Deeks in September 2008

Copyright statement was added to this Chapter and typos were corrected by Tess Moore September 2008

The example was removed from this chapter by Jon Deeks in September 2008

This chapter was last edited by Patrick Bossuyt in April 2008.

CHAPTER 6

DEVELOPING CRITERIA FOR INCLUDING STUDIES

PM Bossuyt, MM Leeflang

Diagnostic accuracy studies are used to obtain how well a test, or a series of tests, is able to correctly identify diseased patients or, more generally, patients with the target condition, the condition of interest.

In general, diagnostic test accuracy, such as expressed in a test's sensitivity and specificity, is not a fixed test characteristic. Instead, accuracy statistics describe the performance of the test in a particular situation, defined by the population, the setting, the type of test and the level of prior testing. In a different population, such as in children versus adults, in a different setting, a rural center in the developing world versus an urban health center, or with a different strategy for pretesting, the sensitivity and specificity are likely to change.

The inclusion criteria should therefore follow on from the primary and secondary objectives of the review. They should be sufficiently specific without being too restrictive, and informative without being too general.

Reviewers need to distinguish between the inclusion of studies into the review and the inclusion of studies into specific meta-analyses within the systematic review. The inclusion criteria for the review can be broad while specific meta-analyses may be focused on a subgroup of studies that can be reasonably combined.

6.1 Types of studies

In a typical diagnostic test accuracy study, patients receive the index test, possibly one or more comparator tests, and the reference standard: the test or procedure used to classify patients as having the target condition or not.

Diagnostic accuracy studies are, in principle, cross-sectional studies. At the time of inclusion, all patients either have the target condition, or they do not, and there is clinical uncertainty about their status. The index test and its comparators are intended to reduce that uncertainty. The clinical reference standard is the best available method for establishing the presence or absence of the target condition. In the troponin study, for example, patients with chest pain are enrolled. Some of them have myocardial infarction, while the others have a different condition, such as oesophagitis.

In some diagnostic accuracy studies, verification of the index test results is based on information that will only be available after inclusion in the study. In most cases, such data are obtained during follow-up of the patients. Such studies have been labeled 'delayed cross-sectional' [REF Knottnerus].

Diagnostic test accuracy studies should be distinguished from prognostic or predictive accuracy studies. In prognostic accuracy studies, test information is used to identify patients that will have an event later on, such as disease recurrence, or sudden death. That event has not happened at the time the test was taken. In predictive accuracy studies, the test is used to identify patients that benefit from treatment, and those that do not.

There are two basic methods to recruit participants for a diagnostic accuracy study. In one set of studies, a single set of inclusion criteria is used. These studies are sometimes called 'cohort type accuracy studies', although that terminology is not entirely appropriate, because accuracy studies are essentially cross-sectional, not longitudinal. In the troponin example, this could mean that all patients admitted to the coronary care unit with suspected myocardial infraction are included. Such studies have also been referred to as 'single-gate' studies [Ref Rutjes].

In other studies, different sets of criteria are used for those with and those without the target condition. These studies are sometimes called 'case-control type accuracy studies', although that terminology is also not entirely appropriate, because accuracy studies are essentially cross-sectional, not longitudinal. In the troponin example, this could mean that two set of participants are recruited: patients with a verified myocardial infarction, and patients without myocardial infarction, but with a different condition. Such studies have also been referred to as 'two-gate' studies [Ref Rutjes].

Two-gate or 'case-control' type studies can be prone to bias, depending on the way in which the inclusion criteria have been defined. Early studies may compare the test results in 'cases' with sever disease with those in healthy 'controls'. If only sections of the spectrum of disease and spectrum of non-diseased are included, the estimated accuracy may not be applicable to the clinical question.

There is no a priori reason to exclude accuracy studies with a two-gate or 'case-control' design. Depending on the question, studies with a limited spectrum may have to be qualified as such in the quality appraisal phase of the review, and omitted from the meta-analysis.

If the review addresses a question that concerns an index test and one or more comparator tests, the comparison can be made in three different ways. The strongest design is based on direct or so-called head-to-head comparisons. In a direct comparison, the index test and the comparator are evaluated in the same study population. Such direct comparisons can be fully paired or not. In a fully paired direct comparison, all study participants receive all tests, as well as the reference standard. Fully paired comparisons are efficient, in terms of the resulting precision relative to the number of study participants. If the design is not fully paired, participants receive only a subset of the tests. Within the set of not fully paired designs, randomized direct comparisons offer the best opportunity to avoid selection bias. In such randomized direct comparisons of diagnostic accuracy, study participants are

randomly allocated to receive the index test or the comparator. All test results are then verified by the reference standard.

With indirect comparisons of study accuracy, estimates of the accuracy of the respective tests are obtained in different study groups. The accuracy of the index test is estimated in one set of studies, while the accuracy of the comparator test is estimated in a different set of not or only partially overlapping studies. As diagnostic test accuracy is not a fixed test property, such indirect comparisons can be prone to selection bias.

Preferentially, comparative accuracy reviews should be based on fully paired or randomized designs. A first exploration of the literature should reveal whether there are sufficient studies to limit inclusion to studies with such a design.

6.2 Participants

The description of the patients has to include a number of items. The first is the set of patient characteristics that described the clinical problem. Which presentations would clinicians recognise as suggesting the clinical problem? The inclusion criteria should include all relevant clinical characteristics with which such patients would present. Note that we use the plural ‘presentations’ to allow for the review covering a range of presentations of interest. In particular, avoid restricting to a particular clinical subgroup or age or gender when first defining presentations. Such restriction can be considered later as outlined below.

Include the setting, if applicable. The accuracy of tests has been known to vary between primary care and secondary or tertiary care, and between screening and diagnostic uses.

Then outline what prior tests patients will have had and how that redefines the range of presentations of interest. For example, a new test may only be of interest in people in whom a prior test has been negative.

Unfortunately, many diagnostic test accuracy studies do not specify the nature of participants and so using extensively defined participant-based criteria will result in exclusion of potentially informative studies. On the other hand information about patient selection is essential as this factor may influence the diagnostic accuracy of the tests.

6.3 Index tests

The research question guiding the review research will have a definition of the index test(s) to be included in the review. Inevitably there will be usually some variability in the index test being evaluated. Components of variability can include the test positivity threshold, or details of the tests being evaluated, such as the manufacturer, type of image processing, or different types of blood markers. A general principle here is that, because the users of the Cochrane Library will be world-wide, reviewers should not narrow the review using index test criteria too much unless there are compelling clinical or policy reasons to do so. In a later stage, differences between index tests or between index test varieties can be explored, for example by subgroup analyses.

6.4 Comparator tests

A comparator test will be included in the review whenever reviewers are interested in the diagnostic accuracy of the index test relative to the diagnostic accuracy of other tests. The accuracy of the index test of the comparator will be assessed by comparing the respective results with the results of the clinical reference standard. Be aware that the reference standard cannot be the comparator in the review!

The index test can be a new test, relative to the comparator tests, which are currently used in practice, or all tests may be already in use in clinical practice. A comparator test may come up when the research question is about replacing one test by a newly developed test. For example, when we want to review whether MRI has at least an equal diagnostic accuracy compared to the CT, the index test will be MRI and the (currently used) CT will be the comparator. In this example, the diagnostic accuracy of both the MRI and the CT will be assessed by comparing both against the same reference standard, for example clinical follow-up.

6.5 Target condition

Tests are used to reduce uncertainty about the presence of the target condition in patients. While disease may describe a state that is often tightly defined, based on microbiological, pathological or histological findings, a target condition is a more clinically relevant term that describes patients with particular clinical history, examination and test outputs. For these patients the benefits of a specific course of management – further diagnostic testing, monitoring, or the initiation, modification or termination of treatment – can be evaluated in research.

The target condition can refer to a particular disease, a disease stage, or to any other identifiable condition that may prompt clinical actions, such as further diagnostic testing, or the initiation, modification or termination of treatment.

Several diagnostic tests can be used to identify more than one condition. For example, chest x-rays are used in the diagnosis of infection, malignancy and inflammatory diseases. Reviewers will need to define their target condition of interest and the reference test used to define the target condition (see 6.6).

6.6 Reference standards

The clinical reference standard is the test, series of tests, or set of procedures that is used to determine the presence or absence of the target condition in patients. Ideally the reference standard is the best available, clinically accepted, error-free procedure to do so.

For many conditions, a series of procedures can be used to establish the presence or absence of the target condition, such as multiple modes of imaging, additional lab tests, or clinical follow-up. In most applications, these reference standards are not interchangeable. They may not have the same degree of error, and may not identify the same segment of the disease spectrum.

Ideally, reviewers should define upfront the reference standard that is going to be used in the review. To avoid ambiguity and bias, a single reference standard should be used in all studies in the review.

References

Knotnerus A, van Weel C, Muris JWM. Evaluation of diagnostic procedures. BMJ 2002;324:477–80.

Rutjes AWS, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PMM. Case–Control and Two-Gate Designs in Diagnostic Accuracy Studies. Clinical Chemistry 2005;51:1335–1341.