Systematic reviews of diagnostic test accuracy
Leeflang, M.M.G.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Summary and discussion
8.1 Summary

Good evidence of the accuracy of diagnostic tests is required to make rational decisions about the provision, selection and application of tests, and to guide the interpretation of test results. Systematic reviews and meta-analyses of test accuracy studies may be the preferred source of such evidence, but building reviews and summarizing study results can be methodologically challenging. The objective of the research reported in this thesis was to provide empirical evidence to guide the development of systematic reviews of diagnostic test accuracy. We specifically addressed the search process, the incorporation of study quality, and the analysis of the data.

Chapter 1 explained the challenges that systematic reviews of diagnostic test accuracy pose and provided an overview of the most recent developments in the methodology for conducting systematic reviews and meta-analyses of diagnostic test accuracy studies. The methods discussed in this Chapter are a reflection of the review methods that will be advocated by The Cochrane Collaboration. The Cochrane Collaboration is the largest international organisation preparing, maintaining and promoting systematic reviews and from October 2008 their Cochrane Database of Systematic Reviews will also include systematic reviews of diagnostic test accuracy.

Diagnostic test accuracy reviews aim to identify and evaluate all available evidence about a specific index test or a comparison of tests. If the yield of the initial search of the literature is too large, a diagnostic search filter can be helpful to reduce this number. The aim of the study reported in Chapter 2 was to assess the fraction of relevant studies that did not pass methodological search filters for diagnostic test accuracy studies. We also determined to what extent the diagnostic search filters decrease the number of studies that need to be screened to find one relevant article. The use of search filters for diagnostic studies led to an inevitably loss of relevant articles, varying from an average of 2% of the total number of relevant primary articles used in this study to 42%. The major reasons for this loss of articles are the poor indexing of diagnostic studies in MEDLINE and the wide range of possible designs for diagnostic accuracy studies. Search filters are also not guaranteed to reduce the number of studies, so their impact on search efficiency will be small. We feel therefore that the use of diagnostic search filters in the development of a systematic review should be discouraged.

The objective of the research in Chapter 3 was to examine to what extent different strategies of defining and incorporating quality of included studies affect the results of meta-analyses of diagnostic test accuracy. We re-analyzed the data from 30 systematic reviews by applying three strategies that varied both in the definition of quality and in statistical approach: (1) restricting the analysis to high-quality subsets; (2) multivariable adjustment for a predefined set of quality items; and (3) multivariable adjustment for significant quality items. We found no evidence for
Summary and discussion

our hypothesis that adjustment for quality in the meta-analysis will lead to less optimistic summary diagnostic accuracy estimates with less variability in results among better-quality studies. The effect of each strategy varied much from one review to another, but also within a single review.

Chapter 4 addressed a possible source of bias when evaluating a test that produces a continuous result: the post-hoc determination of an optimal cut-off value. Optimal cut-off values for continuous test results are often derived in a data-driven way. As this may lead to overoptimistic measures of diagnostic accuracy, we determined the magnitude of bias in sensitivity and specificity associated with data-driven selection of cut-off values in simulated data sets. Three alternative approaches (assuming a specific distribution, leave-one-out, smoothed ROC) were examined for their ability to reduce this bias. The magnitude of bias caused by data-driven optimization of cut-off values was inversely related to sample size. The distribution of the test results had little impact on the amount of bias if sample size was held constant. More robust methods of optimizing cut-off values were less prone to bias, but the performance deteriorated if the underlying assumptions were not met.

Chapter 5 highlighted a possible source of heterogeneity between studies: differences in the prevalence of the target condition. Although it is sometimes claimed that sensitivity and specificity do not depend on disease prevalence, we provide a number of real life examples in which accuracy varied with prevalence. Changes in prevalence and accompanying changes in sensitivity and specificity may be caused by clinical or artefactual variability between studies. Clinical variability refers to differences in the clinical situation. For example, a patient population with a higher disease prevalence may include more severely diseased patients, in which the test performs better. Artefactual variability refers to effects on prevalence and accuracy associated with study design, for example the verification of index test results by a reference standard. Sensitivity and specificity are not fixed test characteristics, but test properties that describe the behaviour of the test in a particular situation. As the setting, filter, or patient group changes, prevalence and accuracy may change. For this reason, variation in disease prevalence and test accuracy between studies can act as a flag for clinicians to detect important differences in study population or study design, affecting accuracy.

In Chapter 6 we systematically reviewed the accuracy of fibronectin tests for the prediction of pre-eclampsia, one of the most important causes of maternal and fetal mortality and morbidity worldwide. Only five studies reported sufficient data to calculate accuracy estimates, such as sensitivity and specificity. At a sensitivity of at least 50%, specificities ranged between 72 and 96% for cellular fibronectin. For total fibronectin, these numbers were 42 to 94%. Due to the small number of studies and the clinical heterogeneity between studies, we refrained from doing a meta-analysis.
Chapter 8

Chapter 7 contained a systematic review of the diagnostic accuracy of galactomannan detection in serum for the diagnosis of invasive aspergillosis (IA) in immunocompromized patients. Twenty-nine studies were included in the meta-analyses. We translated the results of the meta-analyses results to a clinical example. If we use the test at cut-off value 0.5 in a population of 100 patients with a disease prevalence of 8%, that will mean that 2 patients who have IA, will be missed (sensitivity 79%, 21% false negatives) and that 17 patients will be treated unnecessarily (specificity of 82%, 18% false positives). If we use the test at cut-off value 1.5 in the same population, 3 IA patients will be missed (sensitivity 62%, 38% false negatives) and 5 patients will be treated unnecessarily (specificity of 95%, 5% false positives). To improve our understanding of the consequences of false positive and false negative test results in patients, we need more information about (1) the timing of a positive galactomannan test in the course of disease; (2) the timing of positive results in additional tests (for example, clinical signs or CT); and (3) whether earlier treatment improves survival in these patients.

8.2 General discussion

Systematic reviews of diagnostic test accuracy studies are more complicated than systematic reviews of randomized trials. This starts already with question formulation, where the actual or anticipated role of the test in clinical practice and specifications of the patient spectrum are important items to include. In the work reported in this thesis, we specifically addressed the next three steps of a systematic review of diagnostic test accuracy studies: the search process, the incorporation of study quality, and the analysis of the data.

Identification of diagnostic test accuracy studies is complicated by the poor indexing of diagnostic studies in bibliographic databases and the wide range of possible designs for diagnostic accuracy studies. When The Cochrane Collaboration started with its Database of Systematic Reviews and with the developments of systematic reviews of interventions, the same identification problems were encountered for intervention studies. Since then, much effort has gone into the implementation of a clear, unequivocal indexing term for these studies (every randomized controlled trial is now labelled with publication type “randomized controlled trial”) and the development of a register of randomized controlled trials and clinical trials (CENTRAL). One could question whether the same efforts should go into the indexing and registering of diagnostic test accuracy studies.

There is so much variation in diagnostic test accuracy studies that labelling may turn out to be complicated. For example, data on diagnostic test accuracy can be hidden in studies that did not have test accuracy estimation as their primary objective. At this moment, there is no evidence that missing one or two studies will lead to other conclusions. Furthermore, as we saw in Chapter 1, the studies that were not
Summary and discussion

retrieved by searches were most often older studies, in which outdated diagnostic devices may have been used anyway. In many instances, searching with terms for index test(s) and target condition will suffice.

It may be more efficient to put efforts into promoting informative reporting in individual studies and better implementation of the StARD statement. Complete and transparent reporting of diagnostic test accuracy studies may improve their visibility and thus the retrieval of these studies. Better reporting of study design features is also needed for a better understanding of the relation between methodological quality and biased results. Although there is evidence that individual quality items produce biased results, the intertwined effects of quality items cannot be predicted. Furthermore, the importance of different quality items will vary from one research project to another. Research in these directions is also hampered by poor reporting of study characteristics.

Better reporting of study characteristics may also improve explorations of sources of heterogeneity between studies other than methodological quality. Examples are differences in patient spectrum, in setting or in referral pattern of patients, and in the test under evaluation. Heterogeneity is more a rule than an exception in diagnostic test accuracy reviews, which make random effects models the recommended method for meta-analysis of such data. Heterogeneity can be investigated by including study characteristics as covariates in models for meta-analysis, but drawing conclusions based on these investigations only makes sense if enough information is provided by the included studies.

Troublesome identification of studies and poor reporting of methodological quality and study characteristics complicate the interpretation of the results of meta-analyses of diagnostic test accuracy. Summary estimates of sensitivity and specificity alone provide insufficient information. First, they are no fixed properties of a test. Accuracy measures may change from setting to setting and from population to population. However, to what extent they differ can be difficult to assess, because these changes may be confounded by other, often poorly reported study characteristics and flaws in the methodological design. Second, a test is never used on its own. Diagnostic tests are used to reduce uncertainty about a patient’s health status. The results of previous tests may influence the extent to which other tests reduce remaining uncertainty. As long as individual studies do not take combinations of tests into account, reviewing comparative diagnostic questions (add-on, triage, replacement) has to be limited to indirect comparisons between tests, investigated in different patient populations and against different reference standards. Again, the comparative accuracy of these tests may be confounded by other study characteristics. Third, to be able to judge the clinical usefulness of a test or test combination, information should be provided about the consequences of false positive and false negative test results. In case of a false positive test result, the following questions will be important. Will these patients be referred for (invasive) further testing, will they receive a relatively cheap and harmless drug therapy, or will they be referred
Chapter 8

for surgery? How many patients on an annual basis will have a false positive test result? In case of false negative test results, we should ask whether this is a severe condition, or not, and will patients be sent home and never seen again, or will they be followed up? Again the question arises how many patients will be involved. When there is an inconsistency or inconclusiveness in the answers to these questions, the studies that are included in the review may provide information about what is usually done in clinical practice, and how the supposed role of the test(s) under evaluation can be applied in those situations.

In conclusion, the development of methods for identification of studies, for the assessment of methodological quality, for meta-analysis, and for the investigation of statistical and clinical heterogeneity will profit from better reporting of design and characteristics of individual studies. Although quality of reporting is not the same as the methodological quality of a study, better reporting of study characteristics will improve the interpretation of study results and thus the overall quality of a study.

The development and conduct of systematic reviews of diagnostic test accuracy is complicated and the methodology is still in progress. Translating the results presented in those reviews into policy and practice may be even more challenging. It requires knowledge of diagnostic research methodology but also of the clinical context in which the tests are used. When reporting of original research improves, we would like to urge authors of diagnostic test accuracy reviews to guide their readers in understanding the implications of their results. Only then can they rest assured that these results will find their way into clinical practice and improve patient care.