Empirical methods for systematic reviews and evidence-based medicine
van Enst, W.A.

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.
Chapter 9

Summary and general discussion
The principal focus of this thesis is on methodological issues and challenges in conducting systematic reviews, the highest level of evidence to guide clinical decisions. This thesis provides further insight on how the validity of reviews can be improved. In this chapter we summarize and discuss the results described in this thesis, and the implications for clinical practice and for future research.

A high-quality systematic review consists of several steps to arrive at a valid answer to a research question. In this thesis, we have studied and evaluated various steps of the review process: identifying studies, assessing the quality of the primary studies and the quality of reporting of primary studies, assessing heterogeneity, and assessing publication bias. A considerable number of methodological issues were identified. Most of our research projects focused on methodological issues of the review process and meta-analysis of diagnostic test accuracy (DTA) studies, a relatively young field of evidence-based research.

Chapters 2 and 3 addressed the identification of studies. Chapter 2 focused on the use of prospective trial registers for the identification of interventional studies (randomized controlled trials; RCTs), in addition to searching the commonly used electronic databases. This evaluation showed that in a cohort of 210 Cochrane systematic reviews of interventions, about one third (38.1%) of the authors searched clinical trial registers. A search portal for multiple trial registers was used in 70% of these reviews. Thirty-five per cent of the searches resulted in identification of additional relevant trials, of which 14.3% of the ongoing of unpublished data actually were selected for inclusion in the review. In most of these cases (71.4%), the trial was still ongoing and therefore classified as ‘studies awaiting classification’. These findings indicate that the uptake of prospective trial registers for Cochrane reviews has started slowly and should improve in the years to come. In chapter 3 we studied how a search restriction to MEDLINE only affected the summary estimates of the meta-analyses of DTA studies. The relative diagnostic odds ratio (DOR) comparing studies that were uniquely identified in MEDLINE to all studies was 1.04 (95% CI 0.94 to 1.15), meaning that a restriction to studies indexed on MEDLINE studies would only slightly exaggerate the pooled estimate of the DOR, but this increase was not significant. The sensitivity would decrease 0.08% (95% CI -1% to 1%) and specificity 0.1% (95% CI -0.8% to 1%). For a substantial number of reviews (57%) all included studies were indexed on MEDLINE, which had no association with the comprehensiveness of the search strategies used in the reviews. We concluded that omitting
non-MEDLINE studies would not significantly affect the summary estimates of DTA meta-analyses. However, the impact for individual reviews is still unpredictable.

Chapter 4 described current practices of quality assessment in 65 DTA reviews. The quality of the included studies was formally assessed in 92% of the reviews, of which 64% used QUADAS (3% used QUADAS II). In 72% of these reviews, the results of the quality assessment were discussed, while only 9% linked the results of the quality assessment to the conclusion. Half of the reviews that linked the quality assessment to the conclusions had not performed a meta-analysis because of severe methodological heterogeneity and high risk of bias. Quality assessment was further mentioned in 43% of the abstracts of the reviews, while here only 5 reviews linked the outcome of the quality assessment to the conclusion in the abstract. We concluded that the reporting of systematic reviews of DTA should improve to provide readers of the review with a more valid perspective on the performance of the evaluated tests in clinical practice.

Chapter 5 presents an overview of reviews assessing the quality of reporting of DTA studies following the Standards for Reporting of Diagnostic Accuracy (STARD) initiative. Complete and transparent reporting of primary test accuracy studies is essential for review authors and end-users of the review to assess the validity of the design, conduct and analysis of those studies and to enable interpretation of the results. Sixteen reviews were included that had evaluated the quality of reporting, defined as the adherence to STARD, of 1,496 test accuracy studies. We concluded that the quality of reporting was suboptimal. Out of the 25 STARD-items, the mean number of STARD items that scored positive varied from 9.1 to 14.3 with a median of 12.8 items. The number of items that was reported has slightly improved since the introduction of STARD in 2003 (1.41 items; 95% CI: 0.65 to 2.18). It was worrisome that half of the reviews had median proportions of adherence under 50% (scored on less than 13 STARD items). Analysis on item-level indicated that seven items scored particularly low: item 10 (persons executing the tests), item 11 (blinding of readers), item 13 (methods for calculating test reproducibility), item 16 (number of eligible patients not undergoing either test), item 20 (adverse events), item 22 (handling of missing results), and item 24 (estimates of test reproducibility). This is alarming because several of these items can be related to biased results. Overall, although a small improvement of reporting quality was measured in the years after the introduction of STARD, there is still considerable room for improvement. Adherence to STARD should be further
promoted and recommended among researchers, editors and peer reviewers from the stage of designing the study and onwards.

Assessment of heterogeneity is more complex for test accuracy results, mostly due to the bivariate nature of the data. **Chapter 6** focused on the assessment of heterogeneity in 65 systematic reviews of diagnostic test accuracy. In 12 of these the authors decided not to pool the results, for which severe heterogeneity was mentioned as main reason. In 53/65 reviews methods were used to address heterogeneity. A stratified analysis was performed in 47.2%, meta-regression in 35.8% and sensitivity analysis in 22.6%. Many sources of heterogeneity were explored compared to the number of primary studies in a meta-analysis (median ratio 1:5). Based on these findings we made suggestions on what to consider and report on when exploring sources of heterogeneity in diagnostic test accuracy reviews.

**Chapter 7** and **8** both concerned selective publication. It has been suggested that smaller studies are more likely to be published when they show significant positive results. Larger studies may be more likely to be submitted, accepted and published regardless of the estimated effect. This mechanism, which is termed small study effect, can hamper the validity of a systematic review. In **Chapter 7** the methods that are currently used by DTA review authors to detect publication bias in their meta-analyses are described. In a cohort of 114 reviews, 41.2% of the authors assessed publication bias with graphical and/or statistical methods that are aimed to identify small study effects. Most of the used methods are developed to investigate the relationship between treatment effect and study size. Funnel plot evaluation was done in 31 reviews using a wide variety of diagnostic parameters. Statistical tests that assess small study effects were used in 41 reviews. The test described by Deeks (1), which is specifically designed for meta-analyses of diagnostic test accuracy, was only used in 29.2%. The most frequently used test was the Egger test (43.9%) (2). This test, however, has inflated type-1 errors in DTA meta-analyses. In an additional evaluation we used data from the included studies to compare the concordance between the various tests. Agreement between tests (defined as being concordant with respect to significant or non-significant results) ranged between 66% (Deeks vs. Egger) and 87% (Begg vs. Egger) (3), even though each test is aimed to measure the same concept (small study effects). We suggest that reviewers use the Deeks test and be careful with the interpretation of the results as the mechanisms driving publication bias of test accuracy studies are not known. In **Chapter 8** we investigated the presence of small study effects or time lag effect in meta-analyses of diagnostic test accuracy.
Instead of identifying the anticipated small study effects, we found the opposite: studies with small sample sizes had lower accuracy than larger studies. A possible explanation for this finding was a statistical artefact: odds ratios are overestimated in small samples due to the inherent properties of logistic regression models. However, this could not fully explain this unexpected result. However, we did identify a time lag effect. Comparing the 25% most recently published studies with the 25% first published studies, a degree in accuracy over time was found, but it was not as strong as for intervention reviews. We concluded that some of the typical mechanisms associated with selective publication of RCTs are less prominent in test accuracy research.

**Discussion and future recommendations**

Identification of all available studies is the basis of every systematic review. Missing studies may hamper the validity of the review and, therefore, extensive searches are recommended (4-6). Establishment of prospective trial registers has been considered of great importance by many parties (7). Prospective registers, however, seem to be underused not only by Cochrane review authors (8) but also by editors (9). Prospective registers are a valuable new source of information to identify ongoing studies and non-published trial results, among unpublished studies (10) and its potential is gradually growing. The quality of registration and transparency of clinical trial results still improves over time (11), driven by initiatives like the AllTrials campaign – a petition for registration and reporting of all trials and their results (12) – and the requirement of the American Food and Drug Administration service (FDA) to upload all results in ClinicalTrials.gov, within one year after completion (13). The FDA requirement of providing a summary of the primary and secondary outcomes should be supported by and part of all registries in the near future. Additionally, there is a movement going on among major stakeholders towards registering studies involving human subjects beyond clinical trials, like diagnostic test accuracy and prognostic studies. We expect that these great improvements will raise the awareness about trial registries’ usefulness for the scientific community.

However, we have to deal with some barriers first before we can fully profit from the potential benefits of trial registration. Improvement of the user friendliness and advanced search options of the WHO Search Portal (14), a single point of access to all (inter)national trial registries, can be an important next step to increase its usage. At the moment, the search engines of most registers have very limited options, leading to unsuccessful usage or no usage of trial registers (14). Additional guidance for review authors on how
to incorporate unpublished trials or unpublished data in their review should be developed. The Cochrane Handbooks could be a good place for these instructions. To identify the barriers of individual users, an online platform can be initiated for researchers to share their obstacles and ideas for improvement. In addition, editors and peer-reviewers should be encouraged to crosscheck the registered items with published study reports to identify selective reporting of outcomes. This process can be facilitated by automatic downloads of item entries from trial registers (15).

Required registration of diagnostic test accuracy studies will definitely facilitate empirical research on the mechanisms and possible explanations that drives publication bias of test accuracy studies. It is of great importance to understand what type of test accuracy studies take longer to be published or do not get published at all. Our study on small sample size and time lag effects (Chapter 8), two well-known effects in therapeutic meta-analyses, seem to be less prominent in test accuracy research. However, there are some hints that publishing overly optimistic test performances may lead to high cost and harming patients (16).

Identification of diagnostic test accuracy studies is quite complex (5;17). Searches for test accuracy studies should not be performed using a search filter (17;18), unlike for identifying RCTs because diagnostic studies will easily be missed as a result of poor indexing. Therefore, it is often necessary to screen thousands of hits to identify all relevant papers in the systematic review process. Since the number of publications about test accuracy is rapidly increasing, the expected workload will also proportionally increase. In order to advocate efficient searching we assessed the effect on the summary estimates of DTA meta-analyses (Chapter 3) and we found no significant effect. Our results can help reviewers with the decision to restrict their search to MEDLINE when a comprehensive search is not possible due to limited time and resources. Unfortunately, because these results are based on a small number of reviews we can not draw firm conclusions. The impact of a limited search on an individual review is still difficult to predict. Confirmation of our results in other meta-epidemiological studies, preferably stratified for specialty, are warranted. To date, the number of Cochrane DTA reviews might be sufficient to enable a replication of our study in a sample of high quality DTA reviews, with high-quality initial searches. If our results are confirmed a strong recommendation about limiting the searches to MEDLINE may be given.
Systematic reviews are the cornerstone of evidence-based medicine and are used to guide clinical practice (19). The Grading of Recommendations Assessment, Development and Evaluation (GRADE) tool (20) is increasingly used in Cochrane reviews to summarize the main results and to assign levels of evidence for each critical and important outcome (21). GRADE is also adopted by many international guideline developers and policy makers, like the National Institute for Health and Care Excellence (NICE) and the World Health Organization (WHO) (22). GRADE helps guideline groups with assigning levels of evidence for each outcome and provides a framework for translating the evidence into recommendations (including the strength thereof) in a systematic and transparent manner. For the purpose of assigning levels of evidence in systematic reviews, GRADE assesses five domains which are summarised in a so-called Summary of Findings (SoF) Table: study limitations (risk of bias), inconsistency of results (heterogeneity), indirectness of the evidence (applicability), imprecision, and publication bias. The results of this thesis indicate that grading test accuracy evidence is challenging for most of these items. As a result of the poor quality of reporting of primary diagnostic test accuracy studies (Chapter 5), essential information needed for the assessment of these five domains is often missing (23). Editors should be motivated to adopt STARD and adherence to STARD should be enforced. Researchers should also play a role in this and report their study methods and results in such a way, that end-users can easily judge the risk of bias to allow a judgement of their confidence in the results.

In addition, the relationship between QUADAS items and bias in DTA studies is not clear yet. Review authors seem to struggle with how to incorporate risk of bias assessments in the interpretation of the results (Chapter 4) (24). Ideally, more studies are needed to assess this relationship, but such evaluations are only possible when primary studies are well reported. Only then sound guidance to assess this GRADE domain can be developed.

Similar reasoning applies to the GRADE domain ‘Inconsistency’, with which review authors seem to struggle as well (Chapter 6). Heterogeneity is the rule rather than the exception in DTA meta-analyses. Therefore, random effects models are always recommended (25), but exploration of heterogeneity should also be performed. Assessment of heterogeneity in systematic reviews of diagnostic test accuracy is challenging, if not impossible, mostly due to the bivariate nature of the data. Further empirical studies are needed to enable developing guidance, or even another interpretation for this GRADE domain.

Scoring the GRADE domain ‘Publication bias’ for DTA studies is even
more challenging. This thesis indicated that several well-known mechanisms resulting from like small study or time lag effects are not (as strongly) prevalent in DTA studies (Chapter 8) as identified for randomized trials. According to GRADE-guidance, funnel plot asymmetry may lead to downgrading (26). This item is vulnerable for drawing incorrect conclusions, because this method does not seem to be suitable to detect publication bias in meta-analyses of DTA outcomes (Chapter 7) (27). As stated above, empirical research on possible mechanisms of selective publication in DTA studies is needed because it could be possible that this phenomenon works very differently in the DTA domain.

As a result of poor quality of reporting of primary DTA studies, poor linkage of study quality and the occurrence of bias, unclear guidance on how to assess and deal with heterogeneity in DTA meta-analyses, and lack of clarity of the mechanisms of selective publication in the DTA domain, most results of DTA reviews will be labelled as ‘Low’ or ‘Very low’ after GRADE assessment, which will decrease the confidence in the results of those reviews. This is an undesired situation, because it may obstruct relevant changes in medical care. Some GRADE domains for DTA studies, therefore, may need further fundamental empirical evidence, and maybe some of the domains should be reconsidered until we have better understanding of mechanisms specific for test accuracy studies.
REFERENCES

1. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. J Clin Epidemiol 2005 Sep;58(9):882-93.
12. All Trials Registered | All Results Reported. All Trials Campaign 2014 May 6Available from: URL: http://www.alltrials.net/


22. GRADE working group. Organizations that have endorsed or that are using GRADE. http://www.gradeworkinggroup.org/ 2014 February 6


