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
General introduction
INTRODUCTION

Systematic reviews
Evidence-Based Medicine is the integration of best research evidence with clinical expertise and patient values (1). Systematic reviews have become the cornerstone of evidence-based medicine, which is reflected in the position systematic reviews have in the pyramid of evidence-based medicine (Figure 1). Systematic reviews are exhaustive summaries of all studies relevant to answer a specific research question. The advantage of systematic reviews over single primary studies is that they give a structured and transparent overview of all available evidence and its quality, following strict methods. If possible, a meta-analysis can be performed in which the results of all relevant studies are combined to provide an overall estimate of the effect (Figure 2). The strength of the conclusions is based on the overall rating of confidence in the estimated effects, which is only relevant in settings when recommendations for clinical practice are made. This overall confidence depends on methodological limitations of the primary studies, inconsistency of the results (heterogeneity), indirectness of the evidence (applicability), imprecision of the effect estimates and possible reporting biases (2;3).

The most prominent are reviews that evaluate the effectiveness of therapeutic interventions (4). Systematic reviews of healthcare interventions predominantly rely on the evidence from randomized controlled trials (RCTs). However, observational evidence about the harms and effects of interventions might also be required. Besides healthcare interventions, reviews may address other evidence-based medicine (EBM) areas such as etiology, prognosis or diagnosis, although these types of systematic reviews are still less prevalent in the medical literature. Such reviews summarize the results of study types other

Figure 1. Evidence-based medicine pyramid. Study designs are hierarchically ordered based on their relevance that the design presents unbiased results to guide patients’ care.
than RCTs (e.g. cohort studies, case-control studies or cross-sectional studies) and are, therefore, more complex.

At the moment, systematic reviews are regarded as the highest level of evidence and are used to guide clinical practice and decision making (5). Funding agencies and biomedical journals rely on systematic reviews to ensure justification of further research (6,7). A major journal, The Lancet, is now asking authors to report the results of new research within the context of existing systematic review evidence. Systematic reviews have become the most important source of information for making decisions in health care.

Figure 2. A hypothetical meta-analysis that combines the study results of different studies. The results of six individual studies are combined into one summary measure of effect: a risk ratio of 0.51.

### History of systematic reviews and The Cochrane Collaboration

Although systematic reviews are very common nowadays, it was only 30 years ago that they were first developed in the field of medicine (8). Before the introduction of systematic reviews, clinicians and policy makers had to rely on single studies that are more prone to random error and demand substantial time of the clinician to keep up to date. Archie Cochrane (Figure 3), a clinician and epidemiologist who lived between 1909 and 1988, pointed out its shortcomings when he wrote the book Effectiveness and Efficiency: Random Reflections on Health Services published in 1972 (9). He kept challenging the medical profession and wrote “It is surely a great criticism of our profession that we have not organised a critical summary, by specialty or subspecialty, adapted periodically, of all relevant randomised controlled trials” (10). Twenty years later the first Cochrane Centre was established in Oxford, founded by Sir Ian Chalmers and colleagues but named after the progressive Archie Cochrane. One year after its establishment the international Cochrane Collaboration was launched (11).
The Cochrane Collaboration (official logo presented in Figure 4) is a global independent network of health practitioners, researchers, patient advocates and others, involved in the preparation, dissemination, updating and promotion of systematic reviews. To assist review authors, the Cochrane Handbook for Systematic Reviews of Interventions was developed that describes in detail how to undertake a Cochrane review (4), review software was developed (14), and training is made available for all authors. After completion, the high-quality Cochrane reviews are published in The Cochrane Database of Systematic Reviews (CDSR) in The Cochrane Library (CLIB). Besides the CDSR, the CLIB includes five other databases of high-quality, independent evidence, amongst others a database that provides references to clinical trials in all fields of medicine.

For RCTs much empirical research has been performed to establish biases associated with particular characteristics of the study design and conduct, which provided guidance for the Cochrane Handbook (4). However, saturation of knowledge regarding these methods has still not been reached. Therefore, the Cochrane network iteratively contributes to the development of new methods for the preparation of reviews or to improve existing methods. This task is of great importance as the followed methodology determines the quality of a systematic review (15). If the reviews are undertaken without proper knowledge of the methodology, the review may deliver biased results. This is particularly unwanted when they are used to guide clinical practice or policy (16;17).

Within The Cochrane Collaboration the focus has long been on reviews of interventions. In the past few years, Cochrane has broadened its field of interest and has started to cover the field of diagnostic test accuracy (DTA) research. In 2007, the implementation of systematic reviews of diagnostic test accuracy studies was officially launched. In 2008, the first Cochrane Diagnostic
review was published (18) and a draft Cochrane Handbook for DTA reviews was published online (19). In contrast to the knowledge of methods for intervention reviews, the knowledge about biases associated with characteristics of diagnostic accuracy study designs are less explicit and methods on how to undertake the steps of a DTA review evolve rapidly.

Challenges in systematic reviews
Selective publication is the Achilles heel of any systematic review. Consequently, reviewers are challenged to extensively search for studies to identify all relevant evidence (20-22). This can be rather complicated for two reasons. Firstly, not all evidence is published in journals indexed in biomedical databases or is not published at all. Often, the non-published studies are not missing at random, but concern a specific group of evidence: none significant and negative findings (23-25). Missing these studies in a meta-analysis can seriously affect the results. Without negative or null results, the true effect will be overestimated (26). Consequently, review authors need to make efforts to identify unpublished study results. Possible strategies to identify unpublished studies are to search in conference abstracts, contact experts in the field or searching in the recently established prospective trial registers (27). In September 2004 the International Committee of Medical Journal Editors (ICMJE) announced that they would only accept manuscripts for publication as of September 2005 if essential information about the underlying trial design had been deposited into an accepted prospective trial register before enrollment of the first patient (28). This enables review authors to track down all studies that have been initiated and to assess whether or not the results have been published.

A second challenge is that the search strategy has to be very sensitive to ensure that no relevant studies will be missed. However, a sensitive search has the potential to identify a high number of hits that need to be screened for inclusion. Therefore, review authors need to find a balance between a strategy that is sensitive enough to minimize the risk of missing relevant studies, and a strategy that is specific enough to yield a low number of hits. Research has been undertaken to optimize search methods, for example development of search filters for specific medical topics or study designs, such as RCTs. However, filters for identifying DTA studies in the realm of a systematic review seem to fail as they may miss a considerable number of relevant studies (29).

Challenges regarding diagnostic test accuracy reviews
Reviews of DTA aim to summarize all evidence regarding the accuracy of a
diagnostic test that is used to discriminate between diseased and non-diseased patients. These reviews usually do not address RCTs, but other study designs, such as cross-sectional studies, cohort studies or case-control studies. The design of these studies has many variations, including differences in the way patients are selected, in test protocol, in the verification of patients, and in the way the results of the index test and reference standard are assessed. Some of these differences may bias the results of a study, whereas others may have implications for the applicability of the results. An essential step in the review process is therefore to evaluate the risk of bias (30;31). Limitations in the design and conduct of a study may lead to overestimation of the accuracy of the test under study (32;33). The Quality Assessment for Diagnostic Accuracy Studies tool (QUADAS, with QUADAS-2 as the current version (34;35) was developed to assess the risk of bias in DTA studies. This enables authors to draw the conclusions about the results in the light of the risk of bias and concerns regarding the applicability (36). For example, highly biased studies lead to low confidence in the reported results (37), which should be clearly presented to readers of the review.

To enable proper assessment of study quality, complete and accurate reporting of primary studies is necessary (32). Poor reporting of accuracy studies impedes objective assessment of methodological quality and limits assessment of the applicability of the study results (38). Suboptimal reporting therefore hampers the interpretation of the results and generalizability. To improve and promote accurate and complete reporting, the Standards for Reporting of Diagnostic Accuracy Studies (STARD) were developed and published in 2003 (39). Although the quality of reporting of DTA studies has improved significantly after the introduction of STARD, reporting appears to remain suboptimal and could be further improved (40).

Accuracy is usually expressed as the proportion of correctly identified diseased patients (sensitivity) and the proportion of correctly identified non-diseased patients (specificity). As mentioned previously, publication of RCTs may rely on the direction and significance of the effect, thus causing publication bias in meta-analyses (23). For RCTs methods are developed to evaluate whether meta-analyses are affected by this type of bias by investigating the relationship between treatment effect and study size (41-43). These methods, however, don’t seem to be suitable for assessment of selective publication of DTA studies and these tests are therefore not strongly promoted (19). Despite the possibility that DTA studies are also affected by publication bias, empirical studies about possible underlying. Currently, it is challenging
for DTA review authors to assess the possible impact of publication bias on their meta-analysis and how they should interpret the results produced by the different methods to explore publication bias.

Similar to meta-analyses of RCTs, meta-analyses of DTA studies are challenged by heterogeneity arising from a diversity of clinical and non-clinical factors (44). An additional source of heterogeneity in meta-analyses of DTA studies is introduced by the correlated outcomes of interest: sensitivity and specificity. Sensitivity and specificity are negatively correlated due to implicit or explicit differences in the index test threshold for positivity. This threshold effect adds additional heterogeneity in the already complex bivariate meta-analyses for DTA (19). So, dealing with heterogeneity in a sophisticated manner in DTA reviews can be quite complex.

Outline of this thesis

This thesis was undertaken to contribute to the development of methods of systematic reviews. It addresses various steps of systematic reviews. The first step is identification of studies. Chapter 2 evaluates to what extent prospective trial registers are used to identify additional studies for Cochrane systematic reviews. Searching in prospective trial registers is particularly important for the identification of unpublished studies. For this reason, searching prospective trial registers may contribute to the validity of the review. We present current practice of trial identification in prospective trial registers and the results thereof in 210 Cochrane reviews of intervention studies (45).

MEDLINE is a major source for study identification and is freely available via the search engine PubMed. For intervention reviews it has been demonstrated that it is necessary to search studies in multiple databases (46). Searching for published diagnostic studies, however, is complex. Searches for DTA studies must be very sensitive and search filters do not perform satisfactorily (47). Sensitive searches result in high numbers of references needed to read, increasing the workload. In Chapter 3, we investigate the effect on the pooled estimates of meta-analyses of DTA studies, when the search is limited to MEDLINE. If the search could be limited to MEDLINE, this will reduce the workload and screening time as the number of references needed to read will decrease. Additionally, this strategy may also reduce costs because MEDLINE is freely available through the interface PubMed.

In Chapter 4 we evaluate how authors of DTA reviews assess the quality of primary studies and how they incorporate study quality in the conclusions of their reviews. Evaluating the quality of underlying evidence is vital to
understand and interpret the results. Quality assessment for DTA reviews is challenging and demands substantial knowledge of DTA methodology. We describe which tool the review authors use, how they present the results and if and how they incorporate the results of the quality assessment when formulating conclusions. We also present what is reported about the quality of the included studies in the abstracts of the reviews. This is of particular importance since many clinicians usually rely only on the abstracts as they have limited time to read complete articles.

To enable quality assessment of included primary studies, sufficient details regarding the design and conduct of these studies must be. For reporting purposes several guidelines are available (48). For adequate reporting of DTA studies the Standards for Reporting of Diagnostic Accuracy Studies (STARD) have been developed (39). In Chapter 5 we investigate whether the quality of reporting of DTA studies has improved since the introduction of STARD and which items are particularly well or poorly reported.

Studies included in a systematic review may differ with respect to setting, patient or test characteristics, test thresholds, reference standards, or study design. Such differences may cause heterogeneity. In DTA reviews heterogeneity is the rule rather than the exception and to enable optimal interpretation of the results of a DTA review, exploring heterogeneity is an important component of a DTA review. For diagnostic reviews, however, assessment and exploration of heterogeneity is more complex due to the bivariate nature of the outcomes. In Chapter 6 we investigate how review authors deal with heterogeneity in DTA reviews and how they present the results.

As previously mentioned, selective reporting is the Achilles heel of any systematic review. Selective reporting usually leads to an overestimation of the effect and an underestimation of the adverse effects. For intervention reviews the mechanisms behind dissemination biases are well understood, but for DTA reviews those mechanisms are still unclear. Chapters 7 and 8 are both focused on reporting biases in DTA reviews. In Chapter 7 we study whether and how reviewer authors deal with the possible threat of publication bias in their diagnostic reviews. We summarize which methods are used, and how their results are used to formulate the conclusions. We also compare the results of various commonly used tests that all aim to identify publication bias. In Chapter 8, we examine whether small study effects or time lag effects affect DTA meta-analyses. Small study effects refer to the association between the sample size of a study and the outcome of the study. Time lag effect refers to
an association between the time since first publication and the size of the effect of a study. Both of these effects are known to be present in meta-analyses of intervention studies, but to date it is unknown whether these effects are also present in meta-analyses of diagnostic studies.

Finally, this thesis ends with Chapter 9 comprising the summary, concluding remarks and suggestions for further research.
REFERENCE LIST


25. Ioannidis JP. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. JAMA 1998 Jan 28;279(4):281-6.


Mar 12]; Available from: URL: www.jameslindlibrary.org


29. Leeflang MMG, Scholten RJPM, Rutjes AWS, Reitsma JB, Bossuyt PMM. Use of methodological search filters to identify diagnostic accuracy studies can lead to the omission of relevant studies. Journal of Clinical Epidemiology 2006;59:234-40.


