UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Empirical methods for systematic reviews and evidence-based medicine

van Enst, W.A.

Publication date
2014

Document Version
Final published version

Link to publication

Citation for published version (APA):
van Enst, W. A. (2014). *Empirical methods for systematic reviews and evidence-based medicine*. [Thesis, fully internal, Universiteit van Amsterdam].

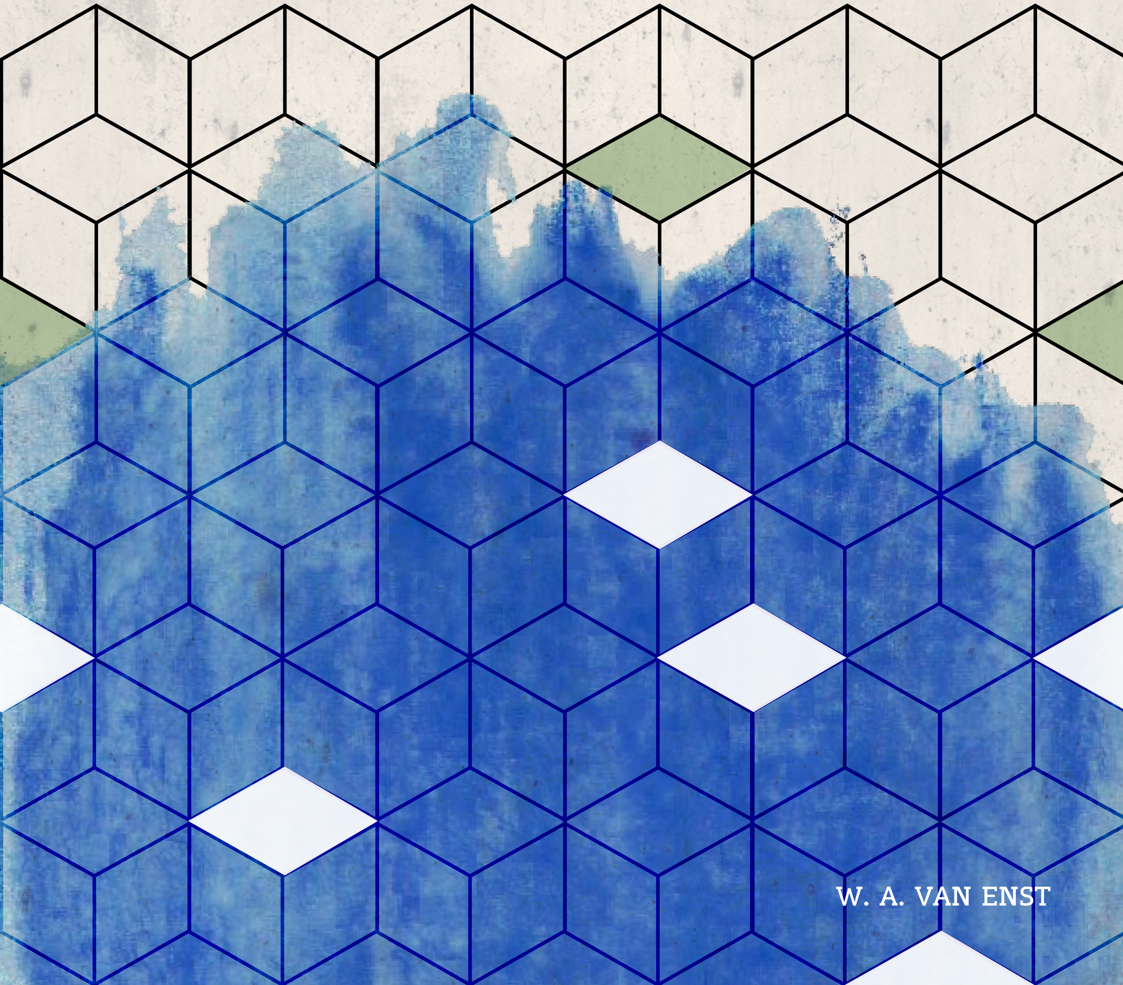
General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

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: https://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.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:17 dec. 2024

EMPIRICAL METHODS
FOR SYSTEMATIC REVIEWS AND EVIDENCE-BASED MEDICINE



W. A. VAN ENST

EMPIRICAL METHODS FOR SYSTEMATIC REVIEWS AND EVIDENCE-BASED MEDICINE

W. A. VAN ENST

COLOFON

Copyright 2014 © W.A. van Enst

Cover and lay out:	Karolien Rijksen, www.karolienrijksen.com
Printed by:	Buijten & Schipperheijn, Amsterdam

Background of the cover: the cover represents several themes discussed in this thesis. The blue cloud pictures the enormous amount of growing information. To answer a research question, information should be systematically structured possible into one simple diamond (pictured in green) which represents the effect that is being studied. Due to invalid methods or publication bias, the pooled estimate may deviated from the true effect (pictured in white).

The research described in this thesis was carried out at the Dutch Cochrane Centre at the Department of Clinical Epidemiology, Biostatistics and Bioinformatics of the Academic Medical Center, Amsterdam, the Netherlands. Financial support by the Dutch Cochrane Centre for publication of this thesis is gratefully acknowledged.

Additional financial support was generously supplied by the Stichting ter Bevordering van de Klinische Epidemiologie Amsterdam, the University of Amsterdam, and the Kennisinstituut van Medisch Specialisten.

EMPIRICAL METHODS FOR SYSTEMATIC REVIEWS AND EVIDENCE-BASED MEDICINE

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. D.C. van den Boom
ten overstaan van een door het college voor promoties ingestelde
commissie, in het openbaar te verdedigen in de Agnietenkapel
op dinsdag 16 september 2014, te 10:00 uur

door **Wynanda Annefloor van Enst**
geboren te Arnhem

PROMOTIECOMMISSIE

Promotor:	Prof. dr. R.J.P.M. Scholten
Copromotor:	Dr. L. Hooft
Overige leden:	Prof. dr. A.H. Zwinderman
	Prof. dr. C.T.J. Hulshof
	Prof. dr. P.I. Spuls
	Prof. dr. S. Middeldorp
	Prof. dr. ir. H.C.W. de Vet
	Prof. dr. R.W.J.G. Ostelo
	Prof. dr. M.M. Rovers

Faculteit der Geneeskunde

TABLE OF CONTENTS

Identification of studies

Quality of studies

Heterogeneity

Dissemination bias

Appendices

CHAPTER I

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).

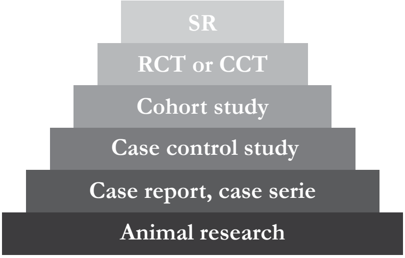
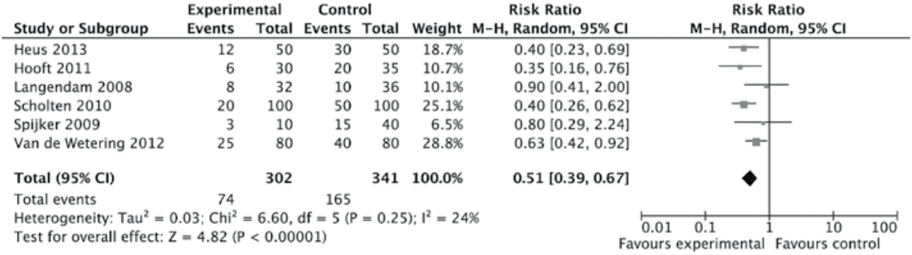


Figure 1. *Evidence-based medicine pyramid. Study designs are hierarchically ordered based on their relevance that the design presents unbiased results to guide patients' care.*

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 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.



	Experimental		Control			Risk Ratio
Study or Subgroup	Events	Total	Events	Total	Weight	M-H, Random, 95% CI
Heus 2013	12	50	30	50	18.7%	0.40 [0.23, 0.69]
Hooft 2011	6	30	20	35	10.7%	0.35 [0.16, 0.76]
Langendam 2008	8	32	10	36	10.1%	0.90 [0.41, 2.00]
Scholten 2010	20	100	50	100	25.1%	0.40 [0.26, 0.62]
Spijker 2009	3	10	15	40	6.5%	0.80 [0.29, 2.24]
Van de Wetering 2012	25	80	40	80	28.8%	0.63 [0.42, 0.92]
Total (95% CI)		**302**		**341**	**100.0%**	**0.51 [0.39, 0.67]**
Total events	74		165			

Heterogeneity: $Tau^2 = 0.03$; $Chi^2 = 6.60$, df = 5 (P = 0.25); $I^2 = 24\%$

Test for overall effect: Z = 4.82 (P < 0.00001)

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).

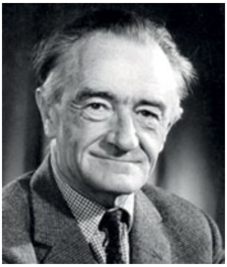


Figure 3. *Archie Cochrane and the logo of The Cochrane Collaboration (12).*

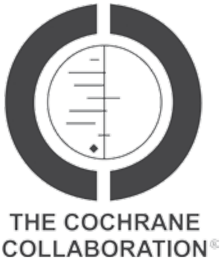


Figure 4. *The logo represents two Cs referring to its name, whilst the inner part shows the first published meta-analysis (13).*

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 enrolment 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

1. Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. Evidence-Based Medicine: how to practise and teach EBM. Second Edition ed. Edinburgh: Churchill Livingstone; 2000.

2. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008 Apr 26;336(7650):924-6.

3. Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? PLoS Med 2010

4. Sep;7(9):e1000326.
Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0 [updated March 2011] ed. The Cochrane Collaboration; 2011.

5. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. 1996. Clin Orthop Relat Res 2007 Feb;455:3-5.

6. Chalmers I. Academia's failure to support systematic reviews. Lancet 2005 Feb 5;365(9458):469.

7. Meerpohl JJ, Herrle F, Antes G, von EE. Scientific value of systematic reviews: survey of editors of core clinical journals. PLoS One 2012;7(5):e35732.

8. Mulrow CD. Rationale for systematic reviews. BMJ 1994 Sep 3;309(6954):597-9.

9. Cochrane AL. Effectiveness and Efficiency. Random Reflections on Health Services. London: Nuffield Provincial Hospitals Trust; 1972.

10. Cochrane AL. 1931-1971: a critical review, with particular reference to the medical profession. London: Office of Health Economics; 1979.

11. The Cochrane Collaboration. Archie Cochrane: the name behind Cochrane. The Cochrane Collaboration 2013 [cited 2014 Mar 24];Available from: URL: http://www.cochrane.org/about-us/history/archie-cochrane.

12. Cochrane A.L., Blythe M. One Man's Medicine: An Autobiography of Professor Archie Cochrane'. Cardiff: Cardiff University; 2009.

13. The Cochrane Collaboration. The Logo of The Cochrane Collaboration. 3-28-0003. 5-26-2014.

14. Review Manager (RevMan) [computer program]. Version 5.1. Copenhagen: The Nordic Cochrane Centre: The Cochrane Collaboration; 2011.

15 Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C, et al. Development of AMSTAR: a measurement tool to assess the

methodological quality of systematic reviews. BMC Medical Research Methodology 2007;15(7):10.

16. Kunz R, Neumayer HH, Khan KS. When small degrees of bias in randomized trials can mislead clinical decisions: an example of individualizing preventive treatment of upper gastrointestinal bleeding. Crit Care Med 2002 Jul;30(7):1503-7.

17. Sheldon TA, Guyatt GH, Haines A. Getting research findings into practice. When to act on the evidence. BMJ 1998 Jul 11;317(7151):139-42.

18. Leeflang MM, Debets-Ossenkopp YJ, Visser CE, Scholten RJ, Hooft L, Bijlmer HA, et al. Galactomannan detection for invasive aspergillosis in immunocompromized patients. Cochrane Database Syst Rev 2008;(4):CD007394.

19. Macaskill P, Gatsonis C, Deeks JJ, Harbord RM, Takwoingi Y. Analysing and Presenting Results. In: Deeks JJ, Bossuyt PM, Gatsonis C, editors. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. The Cochrane Collaboration; 2010. p. 46-7.

20. Beynon R, Leeflang MM, McDonald S, Eisinga A, Mitchell RL, Whiting P, et al. Search strategies to identify diagnostic accuracy studies in MEDLINE and EMBASE. Cochrane Database Syst Rev 2013;9:MR000022.

21. Hopewell S, Clarke M, Lefebvre C, Scherer R. Handsearching versus electronic searching to identify reports of randomized trials. Cochrane Database Syst Rev 2007;(2):MR000001.

22. Sampson M, McGowan J, Cogo E, Grimshaw J, Moher D, Lefebvre C. An evidence-based practice guideline for the peer review of electronic search strategies. J Clin Epidemiol 2009 Sep;62(9):944-52.

23. Dickersin K. The existence of publication bias and risk factors for its occurrence. JAMA 1990 Mar 9;263(10):1385-9.

24. Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. Cochrane Database Syst Rev 2009;(1):MR000006.

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.

26. Thornton A, Lee P. Publication bias in meta-analysis: its causes and consequences. J Clin Epidemiol 2000 Feb;53(2):207-16.

27. Dickersin K, Chalmers I. Recognising, investigating and dealing with incomplete and biased reporting of clinical research: from Francis Bacon to the World Health Organisation. James Lind Library 2010 [cited 2014

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

28. DeAngelis CD, Drazen JM, Frizelle FA, Haug C, Hoey J, Horton R, et al. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. JAMA 2004 Sep 15;292(11):1363-4.

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.

30. Reitsma JB, Rutjes AW, Whiting P, Vlassov W, Leeflang MM, Deeks JJ. Assessing methodological quality. In: Deeks JJ, Bossuyt PM, Gatsonis C, editors. Cochrane Handbook for Systematic Reviews of Diagnostic Test Acuracy. Version 1.0.0 ed. The Cochrane Collaboration; 2009.

31. Herbert RD, Bo K. Analysis of quality of interventions in systematic reviews. BMJ 2005 Sep 3;331(7515):507-9.

32. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med 2004 Feb 3;140(3):189-202.

33. Whiting PF, Rutjes AW, Westwood ME, Mallett S. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. J Clin Epidemiol 2013 Oct;66(10):1093-104.

34. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011 Oct 18;155(8):529-36.

35. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol 2003 Nov 10;3:25.

36. Dahabreh IJ, Chung M, Kitsios GD, Terasawa T, Raman G, Tatsioni A, et al. Comprehensive Overview of Methods and Reporting of Meta-Analyses of Test Accuracy. Methods Research Reports 2012 Mar.

37. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008 Apr 26;336(7650):924-6.

38. Ochodo EA, Bossuyt PM. Reporting the accuracy of diagnostic tests: the STARD initiative 10 years on. Clin Chem 2013 Jun;59(6):917-9.

39. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Fam Pract 2004 Feb;21(1):4-10.

40. Smidt N, Rutjes AW, van der Windt DA, Ostelo RW, Bossuyt PM, Reitsma JB, et al. The quality of diagnostic accuracy studies since the STARD statement: has it improved? Neurology 2006 Sep 12;67(5):792-7.

41. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. Biometrics 1994 Dec;50(4):1088-101.

42. Duval S, Tweedie R. Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics 2000 Jun;56(2):455-63.

43. Egger M, Davey SG, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997 Sep 13;315(7109):629-34.

44. Engels EA, Schmid CH, Terrin N, Olkin I, Lau J. Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. Stat Med 2000 Jul 15;19(13):1707-28.

45. van Enst WA, Scholten RJ, Hooft L. Identification of additional trials in prospective trial registers for Cochrane systematic reviews. PLoS One 2012;7(8):e42812.

46. Sampson M, Barrowman NJ, Moher D, Klassen TP, Pham B, Platt R, et al. Should meta-analysts search Embase in addition to Medline? J Clin Epidemiol 2003 Oct;56(10):943-55.

47. Leeflang MM, Scholten RJ, Rutjes AW, Reitsma JB, Bossuyt PM. Use of methodological search filters to identify diagnostic accuracy studies can lead to the omission of relevant studies. J Clin Epidemiol 2006 Mar;59(3):234-40.

48. EQUATOR Network: Enhancing the QUAlity and Transparency Of health Research. 2006. 5-26-2014.

CHAPTER 2

Identification of additional trials in prospective trial registers for Cochrane systematic reviews

W. Annefloor van Enst
Rob J.P.M. Scholten
Lotty Hooft

PLoS One 2012;7(8):e42812

ABSTRACT

Background Publication and selective outcome reporting bias are a threat to the validity of systematic reviews. Extensive searching for additional trials in prospective trial registers could reduce this problem. We have evaluated how authors of Cochrane systematic reviews currently make use of trial registers as an additional source for the identification of potentially eligible trials.

Methods A search was performed in the Cochrane Library of Systematic Reviews to identify systematic reviews with a published protocol since 2008 and a published review between 2008 and 2010. It was assessed of authors had used prospective trial registers, the aim to search prospective trial registers and the yield of searching in prospective trial registers.

Results We included 210 systematic reviews. In 80 reviews (38.1%) the authors had searched in one or more prospective trial register(s) of which 55% had searched in overlapping search portals and individual registers. Most frequently assessed were the MetaRegister (66.3%) and Clinicaltrials.gov (60%) which is in sharp contrast of other registers or portals like the WHO ICTRP Search Portal (20%). Reported motives to use registers were to identify ongoing trials (83.3%), to identify unpublished outcomes or trials (23.5%), to identify recently published trials (11.8%), or to identify any relevant trial (3.9%). In 28 reviews (35%) the authors had selected (ongoing) trials identified in trial registers as potentially eligible.

Discussion/Conclusion Trial registers as an additional source of information are gaining acknowledgement amongst Cochrane reviewers. Nevertheless, searches seem to be inefficient as overlapping databases are frequently consulted, while the WHO ICTRP Search Portal that includes the data from all approved registers worldwide is being underused. Moreover, the emphasis is now on the identification of ongoing trials, although the prospective registers offer a broader potential. Further familiarity of registers and guidance how to search and to report will help to implement this as a common method and utilize the full potential of prospective trial registers for systematic reviews.

INTRODUCTION

Systematic reviews of randomized controlled trials (RCTs) are regarded as the highest level of evidence to guide decisions in healthcare. Cochrane reviews in particular are of high quality because these reviews follow explicit, transparent and systematic methods (1;2). The results of systematic reviews, however, can be biased when the included evidence does not offer a fair representation of all existing evidence. Empirical evidence consistently suggests that statistically significant and positive findings are more likely to be published than non-significant or negative findings (3;4) and will take shorter time to be submitted and to get published after completion of the study (5-8). When publication of trials or outcomes depends on the results, publication bias and selective outcome reporting bias may arise. This can affect the results from the meta-analysis of the review and possibly also the results of a review without any meta-analysis (9).

To minimize the effects of publication bias and outcome reporting bias review authors should perform a comprehensive search to identify all relevant trials (10;11). Most trials can be identified in well known biomedical databases like MEDLINE or EMBASE. Nevertheless, some trials can only be identified by the use of additional strategies like contacting experts, checking the reference lists of eligible trials, handsearching of conference proceedings, searching the Internet with web search engines like Google or searching the websites of relevant organizations (12). These strategies may be very time consuming and still do not guarantee that all relevant trials will be found.

Recently searching in prospective trial registers can be added as another strategy to identify relevant trials. Already in 1986 it was suggested that prospective registration of trials could reduce or even resolve the problems resulting from publication bias and outcome reporting bias (11;13;14). However, for a long period of time trials were not systematically registered. In September 2004, however, 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 enrolment of the first patient. In November 2004, the World Health Organization (WHO) was asked by the international scientific and political community to facilitate the establishment of a network of these national clinical trials registers and to develop strict criteria for 'registry approval' concerning the content, quality and accessibility (15). Currently, there are 15 registries that meet these strict

international requirements (16).
The prospective registration policy of the ICMJE was adopted by many biomedical journals. Trial registration has become common and the number of registered trials has grown considerably (17). Authors can search in the prospective trial registers for ongoing trials, for completed trials that have not published the results (yet) or to check whether the primary outcome has changed or if all outcomes have been reported. The various national or regional trial registers can be searched individually or simultaneously through search portals that include various other registers e.g. the WHO International Clinical Trial Registry Platform (ICTRP) and the MetaRegister of Current Controlled Trials. Our objective was to evaluate how authors of Cochrane systematic reviews make use of trial registers as an additional source for the identification of potentially eligible trials.

METHODS

Selection of reviews

For this study, we included reviews with a protocol published in 2008 that had been converted into a full Cochrane Review by February 2010. The Cochrane Collaboration's Information Management System (ARCHIE at archie.cochrane.org) was searched to identify all Cochrane protocols that were published in 2008 and the Database of Systematic Reviews was searched for full Cochrane reviews in February 2010. Cochrane Diagnostic Test Accuracy reviews, Cochrane Methodology reviews and Cochrane Overviews of reviews were excluded.

The publication date of protocols (2008) was chosen, because the Cochrane Handbook for Systematic Reviews of Interventions which authors are required to read and follow as a guide had added the statement that: 'Trial registers, are the best solution to unpublished trials and the conduct of all systematic reviews should be much simplified when the use of registers becomes widespread' at the end of 2006 (18).

Data extraction

We extracted information on three subjects. First, of each included review we extracted the applied strategies for the identification of additional potentially eligible trials and emphasized on the use of prospective trial registers. We distinguished between three methods to assess trials in a prospective trial registers:

- Search portals with which one can search in various trial registers,
- National or regional registers that are approved by the ICMJE or WHO,
- Non-approved registers (e.g. registers of the pharmaceutical industry,
- Non-approved national registers or registers of specialized foundations).

Second, in case any of these three methods was applied, we searched the review for a particular motive the author had for searching in the prospective trial registers. These motives were retrospectively classified as identification of ongoing trials, identification of unpublished trials or outcomes, identification of recently completed yet unpublished trials, or identification of any relevant trial.

Third, for every review for which prospective trial registers were searched, we registered if the searches had yielded trials from prospective trial registers. We classified trials as identified in a prospective trial register when the trial identification number was reported or when the reference of the trial included a link or a reference to a prospective trial register without any other reference to a publication in a journal. Cochrane reviews distinguish among four types of references: included studies, excluded studies, ongoing studies and studies awaiting assessment. The number of reviews that had identified trials from prospective trial registers was registered for each type of reference. Two reviewers independently extracted data. Disagreements were resolved by discussion and in case of persistent disagreement a third expert was asked to make a decision.

RESULTS

We identified 519 protocols for Cochrane reviews that were published in 2008. Of these protocols 212 were converted into a full systematic review in the Cochrane Database of Systematic Reviews by February 2010. Two reviews were excluded because they were Diagnostic Test Accuracy reviews. The final set of systematic reviews consisted of 210 Cochrane reviews. These reviews were published in 2008 (n = 7), 2009 (n = 147) and 2010 (n = 56).

Most applied strategies to identify additional potentially eligible trials were checking the reference lists (83.3%) and contacting experts (49.0%) (Table 1). In 80 of the included reviews (38.1%), the authors had searched in at least one prospective trial register either by using a search portal, a national or regional register approved by the ICMJE or WHO, or a non-approved register. Of those 80 reviews the MetaRegister of Current Controlled Trials was the most frequently used search portal (66.3%) and the WHO ICTRP search portal was used in only 20.0% (Table 2). Clinicaltrials.gov was the most searched individual register (60.0%) which is in sharp contrast with other registers (Table 2). In 75 reviews (93.8%) the authors had searched in a search portal or register that is approved by the ICMJE or WHO, leaving 5 reviews in which only non-approved registers were assessed.

Table 1. *Methods applied for identification of trials in addition to searching in biomedical databases in 210 Cochrane reviews*

Method	Number of reviews (%)*
Checking reference lists	175 (83.3%)
Contacting experts	103 (49.0%)
Searching in prospective trial registers	80 (38.1%)
Handsearching of conference abstracts	78 (37.1%)
Searching the Internet	11 (5.2%)
No additional methods applied	10 (4.8%)

* *Most review authors applied multiple strategies to indentify additional trials. Therefore, the summation of percentages exceeds 100%.*

The combinations of usage are presented in Table 3. In 44 reviews (55%) both a search portal and one or more individual trial registers that are already included in the search portal had been consulted, ignoring the overlap.

In 51 of the 80 reviews (63.8%) one or more motives for searching in prospective trial registers were reported. Of the 51 reviews the motives were to identify ongoing trials (83.3%), to identify unpublished outcomes or trials (23.5%) (they either searched for unpublished outcomes (9.8%), unpublished trials (11.8%), or both (1.9%)), to identify recently published trials (11.8%), and to identify any relevant trial (3.9%).

In 28 of the 80 in which a search portal or register was used reviews (35.0%) the authors had yielded potentially eligible trials from a prospective trial register: in 4 reviews (14.3%) trials were actually included in the review, in 8 reviews (28.6%) the potentially eligible trials ended in the excluded category, in 20 reviews (71.4%) in the ongoing studies category and in 4 reviews (14.3%) in the category of studies awaiting classification (the total percentage exceeds 100% because some reviews found trials for multiple categories). In 34 there were no trials from prospective trial registers mentioned in the reference lists. Additionally, in 18 of the 80 reviews (22.5%) the results from extended strategies were some what confusingly documented such that we were not sure whether the reviewer had or had not identified the trial in a prospective register. None of the reviews explored the possible impact of publication bias.

Table 2. *Overview of trial registers that were searched in 80 Cochrane reviews*

Type of register	Number of reviews (%)*
Search portals	56 (70%)
MetaRegister of Current Controlled Trials	53 (66.3%)
WHO ICTRP Search Portal	16 (20.0%)
International Federation of Pharmaceutical Manufacturers & Associations (IFPMA)	0 (0%)
Registers approved by the WHO or ICMJE	52 (65%)
Clinicaltrials.gov	48 (60.0%)
Australian New Zealand Clinical Trials Registry (ANZCTR)	8 (10%)
International Standard Randomised Controlled Trial Number Register (ISRCTN)	4 (5%)
Netherlands Trial Register (NTR)	3 (3.8%)
Chinese Clinical Trial Register (ChiCTR)	2 (2.5%)
Japan Primary Registries Network	1 (1.3%)
Non-approved registers	44 (55%)

** Most review authors searched in more than one register. Therefore, the summation of percentages exceeds 100%.*

Table 3. *Overview of combinations of trial registers/search portals that were searched in 80 Cochrane reviews*

Combination of usage	Number of reviews (%)
Portal only	12 (15%)
Portal and approved register	13 (16.3%)
Portal and non-approved register	11 (13.7%)
Approved register only	11 (13.7%)
Approved register and non-approved register	8 (10.0%)
Non-approved register only	5 (6.2%)
Combination of all (portal, approved register and unapproved register)	20 (25.0%)
Search strategy assessed overlapping portal and register	44 (55%)

DISCUSSION

This study indicates that the majority of Cochrane authors tried to identify additional trials through extended search strategies. In 38.1% of the reviews this extended search involves consulting any prospective trial registers. This number is a good start but should be improved in the coming years. The emphasis of the use of prospective trial registers is now on the identification of ongoing studies, but this could be much more extensive, for example to compare the outcomes of the protocol to the outcomes in the publication.

The proportion of authors that had searched in prospective trial registers to identify additional trials is promising since trial registration and its possibility to minimize bias in systematic reviews has only received major attention since 2005. In this year the ICMJE required trials to be registered in publicly accessible databases. However, compared to other strategies, like contacting experts or checking the reference lists of eligible trials, this source seems to be underused and there still seems to be room for improvement. First, search portals can be seen as the most efficient way to identify trials as these searches in multiple registers at once. In our evaluation we found that most authors searched the MetaRegister of Current Controlled Trials although the WHO ICTRP Search Portal has more underlying registers it was only consulted by 20% of the review authors. Furthermore, we found that many authors consult Clinicaltrials.gov, which is also accessible through both the WHO ICTRP Search Portal and the MetaRegister. The popularity of the MetaRegister and Clinicaltrials.gov may be the consequence of the guidance of the Cochrane Handbook of 2006 which emphasized on consulting this portal and register (18). Currently, many more registers are mentioned in the Handbook, but clear guidance on which search methods are most efficient is still lacking (19). Second, in 55% of the searches in prospective registers, redundant work was performed by searching a portal and a register that is already incorporated by the portal. This could be the result of a lack of knowledge amongst review authors but it could also be an indication that authors have doubts about the sensitivity of a search in a search portal. Especially in more extensive and complex search strategies a search in a portal could miss studies from underlying registers and the reverse (20). Moreover, some search portals update the registered trials only weekly or monthly (21). To ensure completeness of the search review authors might have decided to search all sources and ignore their overlap. Future improvement of the search options and more frequent updating of the various registers could make the search portals more efficient. Finally, approved registers were used in 52 reviews, whereof only three non-western registers (not American, Australian or European). Authors should actively try to search in all registers to prevent a geographical skewed distribution of trials. The WHO ICTRP Search Portal is helpful for this purpose as it searches in western and non-western registers.

The main motive for searching in prospective trial registers was to identify ongoing trials. This is not surprising because the Cochrane Handbook for Systematic Reviews of Interventions recommends consulting prospective registers for the identification of ongoing trials (22). However, some authors seem to work ahead of guidance and had broader purposes like identification of unpublished trials or unpublished outcomes, which enable controlling for or assessing publication bias, or selective outcome reporting bias. This strong feature of trial registers seems to be used only occasionally and deserves more emphasis and guidance. In addition to this subject, prospective trial registers can also be consulted to compare the primary outcomes as stated in the register to the actual published primary outcomes (23). However, this strong feature of trial registers seems to be used only occasionally and deserves more emphasis and guidance. On the other hand, to improve this feature, the quality of trial data provided in trial registers has deficiencies and needs to be improved (24). The approved registers are most appropriate to compare the outcomes as they fulfill strict criteria on reporting. The approved registers can easily be assessed using the WHO ICTRP Search Portal that incorporates all approved registers.

Searching in prospective trial registers seems to be worthwhile. In 35 reviews (43.8%), at least one or more trials had been identified in a trial register as potentially eligible for the review. Most of those were included in the ongoing trials section. This may alert readers and enable them to track down such trials and update the results of the reviews for their own purposes. This also applies to trials identified in trial registers that were listed in the excluded studies section. Those trials might still be important for the reader if their study question differs from the study questions of the review. Therefore, searching prospective trial registers can help to identify relevant outcomes or trials for the reviews and contribute to the completeness of the evidence and quality of the review.

Our study has some limitations. First, we studied a cohort of Cochrane reviews which apply uniform methods and include detailed reports of the results. We assume that our results do not apply to non-Cochrane reviews. Secondly, although Cochrane authors follow strict methodological and reporting criteria, it could be that not all our items of interest were transparently reported in the review. For example, according to our data, checking the reference list occurred in 83% of the reviews. This seems low for Cochrane reviews where it is standard methodology. This implies that also other items, for example the use of prospective trial registers, could have been not transparently reported, thereby underestimating the results. Incomplete reporting can also apply for the reported motives. Finally the yield from searches in prospective trial registers was poorly and inconsistent documented in almost a quarter (22.5%) of the reviews that had consulted prospective registers or search portals. Therefore, the yield of trials retrieved from prospective trial registers in Cochrane reviews is possibly underestimated. Third, it would be very interesting to measure the effect of the inclusion of trials from prospective trial registers on the results of the review but unfortunately we had too low power (n=4) to perform sensible analysis (25;26). Future research should try to measure this effect.

Our study indicates that many Cochrane authors did search in prospective trial registers, which has led to the identification of relevant trials for the review. However, there seems to be room for improvement. More reviewers should search prospective trial registers and search more efficiently utilizing the full potential of prospective registers instead of focusing on identification of ongoing trials. The Cochrane Collaboration should promote the use of prospective trial registers more intensively and give more guidance to authors to increase the frequency of using prospective registers. This especially applies to the usefulness of trial registers beyond the identification of ongoing trials and to the efficiency to search the WHO ICTRP Search Portal that includes all approved national or regional registers. Coordinators of prospective trial registers and search portals could help authors and trials search coordinators of Cochrane Review Groups to make their search portals more user-friendly. These measures may ensure more frequent and efficient use of current search portals and prospective trial registers using all its potential with the ultimate goal restricting biased trial results and thereby improving evidence-based decisions in healthcare.

ACKNOWLEDGMENTS

We would like to thank Cynthia Ho for her contribution to the data-extraction and René Spijker for his contribution to develop a search strategy for this study.

AUTHOR CONTRIBUTIONS

Conceived and designed the experiments: WE RS LH. Performed the experiments: WE. Analyzed the data: WE LH. Contributed reagents/materials/

analysis tools: WE RS LH. Wrote the paper: WE RS LH.

REFERENCE LIST

1. Fleming PS, Seehra J, Polychronopoulou A, Fedorowicz Z, Pandis N. Cochrane and non-Cochrane systematic reviews in leading orthodontic journals: a quality paradigm? Eur J Orthod 2012 Apr 24.
2. Higgins JPT, Green S. Cochrane Handbook for Systematic Reviews of Interventions. 3th ed. Chichester, England: John Wiley & Sons Ltd.; 2008.
3. Dickersin K. The existence of publication bias and risk factors for its occurrence. JAMA 1990 Mar 9;263(10):1385-9.
4. Dwan K, Altman DG, Arnaiz JA, Bloom J, Chan AW, Cronin E, et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. PLoS One 2008;3(8):e3081.
5. Hopewell S, McDonald S, Clarke M, Egger M. Grey literature in meta-analyses of randomized trials of health care interventions. Cochrane Database of Systematic Reviews 2007;(2):MR000010.
6. Hutton JL, Williamson PR. Bias in meta-analysis due to outcome variable selection within studies. Journal of Applied Statistics 49[3], 359-370. 2000.
7. Ioannidis JP. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. JAMA 279[4], 281-286. 1998.
8. Williamson PR, Gamble C, Altman DG, Hutton JL. Outcome selection bias in meta-analysis. Statistical Methods in Medical Research 14[5], 515-524. 2005.
9. Kirkham JJ, wan KM, Altman DG, Gamble C, Dodd S, Smyth R, et al. The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. BMJ 2010;340(c365).
10. Egger M, Juni P, Bartlett C, Holenstein F, Sterne J. How important are comprehensive searches and the assessment of trial quality in systematic reviews? Health Technology Assessment 2003;7:1-76.
11. Song F, Parekh S, Hooper L, Loke YK, Ryder A, Sutton AJ, et al. Dissemination and publication of research findings: an updated review of related bias. Health Technol Assess 2010;14(8).
12. Blackhall K. Finding studies for inclusion in systematic reviews of interventions for injury prevention - the importance of grey and unpublished literature. Injury Prevention 2007;13:359.
13. Simes RJ. Publication bias: the case of an international registry of clincial trials. Journal of Clinical Oncology 1986;4:1529-41.
14. Dickersin K. The existence of publication bias and the risk factors of its

occurence. JAMA 263[10], 1385-1389. 1990.

15. DeAngelis CD, Drazen JM, Frizelle FA, Haug C, Hoey J, Horton R, et al. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. JAMA 2004 Sep 15;292(11):1363-4.

16. World Health Organisation. How to cite a record on a clinical trials register. http://www.who.int/ictrp/faq/en/index.html . 2012. 1-3-2012.

17. Ghersi D, Pang T. From Mexico to Mali: four years in the history of clincal trial regsitration. Journal of Evidence Based Medicine 2009;2(1):1-7.

18. Higgins JPT, Green S. Cochrane Handbook for Systematic Reviews of Interventions. Chichester, England: John Wiley & Sons Ltd.; 2006.

19. Higgings JPT, Green S. The Cochrane Handbook for Systematic Reviews of Interventions. 5.1.0 ed. Chichester, England: John Wiley & Sons Ltd.; 2011.

20. Duffy J, Glanville J, Mccool R, Varley D, Blackhall K. Clinical trial registers as a source of trials for HTAs. 2012. 8-6-2012 BC.

21. World Health Organisation. International Clinical Trials Registry Platform. http://www who int/ictrp/en/ 2011 [cited 2011 May 18];

22. Lefebvre C, Manheimer E, Glanville J. Searching for studies. In: Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions.Chichester; UK: John Wiley & Sons Ltd; 2008. p. 95-150.

23. Mathieu S, Boutron I, Moher D, Altman DG, Ravaud P. Comparison of registered and published primary outcomes in randomized controlled trials. JAMA 2009 Sep 2;302(9):977-84.

24. Viergever RF, Ghersi D. The quality of registration of clinical trials. PLoS One 2011;6(2):e14701.

25. Hart B, Lundh A, Bero L. Effect of reporting bias on meta-analyses of drug trials: reanalysis of meta-analyses. BMJ 2012;344:d7202.

26. Trinquart L, Abbe A, Ravaud P. Impact of reporting bias in network meta-analysis of antidepressant placebo-controlled trials. PLoS One 2012;7(4):e35219.

CHAPTER 3

MEDLINE studies are sufficient for meta-analyses of Diagnostic Test Accuracy

W. Annefloor van Enst
Rob J.P.M. Scholten
Penny Whiting
Aeilko H. Zwinderman
Lotty Hooft

*Journal of Clinical Epidemiology 2014, Jul 1.
pii: S0895-4356(14) 00200-5*

ABSTRACT

Background To investigate how the summary estimates in diagnostic test accuracy (DTA) systematic reviews are affected when searches are limited to MEDLINE.

Methods A systematic search was performed to identify DTA reviews that had conducted exhaustive searches and included a meta-analysis. Primary studies included in selected reviews were assessed to determine whether they were indexed on MEDLINE. The effect of omitting non-MEDLINE studies from meta-analyses was investigated by calculating the summary ratio of DORs (RDORs): DOR MEDLINE-only/DOR all studies. We also calculated the summary difference in sensitivity and specificity between all studies and only MEDLINE-indexed studies.

Results Ten reviews contributing 15 meta-analyses met inclusion criteria for quantitative analysis. The RDOR comparing MEDLINE only studies to all studies was 1.04 (95% CI 0.95 to 1.15). Summary estimates of sensitivity and specificity remained almost unchanged (difference in sensitivity -0.08%; 95% CI -1% to 1%; difference in specificity: -0.1%; 95% CI -0.8% to 1%).

Discussion/Conclusion Restricting to studies indexed on MEDLINE did not influence the summary estimates of the meta-analyses in our sample. In certain circumstances, for instance when resources are limited, it may be appropriate to restrict searches to MEDLINE. However, the impact on individual reviews cannot be predicted.

What is new?

Key findings

- Less than half of the DTA systematic reviews (43%) included stud ies that are not endexed in MEDLINE.
- Omitting non-MEDLINE studies from the meta-analysis did not significantly hamper the diagnostic odds ratio, sentivity or specificity.

What this adds to what was known

- This is the first meta-epidemiological evidence on the impact of search strategies for DTA systematic reviews.

What is the implication, what should change now

- Empirical evidence indicates that searching in databases beyond MEDLINE for a DTA systematic review may no longer be regarded an absolute neccessity tot produce valid outcomes.

INTRODUCTION

Systematic reviews of diagnostic test accuracy (DTA) studies are important to inform evidence based use of diagnostic tests in clinical practice. A comprehensive search across multiple databases combined with screening the search results to identify studies for inclusion in the review is a key part of any systematic review.(1;2) This process can be time consuming and costly, especially for DTA reviews which often involve screening several thousand references. Methods for efficient searching are therefore needed without introducing bias by missing relevant studies.

There are many electronic bibliographic databases that can be used to identify biomedical studies.(3) Most reviewers only search a small subset of the available databases, even in a comprehensive search. The best-known databases include MEDLINE and EMBASE. As from January 2010, MEDLINE records are included in EMBASE, while some EMBASE records are not covered by MEDLINE. EMBASE, covers other journals especially drug therapy journals, more European journals, and more non-English journals compared to MEDLINE.(4) Regional databases like PASCAL and LILACS or specialized databases like PsychINFO may include studies additional to EMBASE and MEDLINE. Thus if one of these databases is not searched when conducting a systematic review there is a risk that some relevant studies will be missed.

When time or financial resources are limited, simplifying the searches can be a practical solution. However, this may compromise the quality of the review by missing relevant studies. Much research has been done to develop search filters to enhance the precision of the search, defined as the number of relevant records identified by a search divided by the number of records identified. Therefore the number needed to read (NNR), defined as the number of records needed to read to find one relevant additional paper, can be reduced.(5;6) However, empirical evidence has found that even the most sensitive methodological filters for searching for DTA studies miss relevant studies.(7;8)

Reducing the number of databases to be searched could reduce the amount of work involved in searching and also the NNR for screening search results and so be time- and cost effective. In particular, costs will be reduced if only MEDLINE is searched as this database is freely accessible through the PubMed interface. Empirical research has shown that excluding EMBASE when searching for randomized controlled trials (RCTs) will affect the results of intervention reviews. This is the consequence of a systematic difference between the two databases for RCTs. Trials that are indexed on MEDLINE on average find larger effects and have more significant results compared to studies indexed on EMBASE. Searching exclusively in MEDLINE may lead to an overestimation of the magnitude of treatment effects, which could affect patient management.(9) While the publication process of trials is often dependent on identification of a significant effect, there is no such effect in diagnostic studies as the main outcomes are accuracy measures such as the diagnostic odds ratio (DOR), sensitivity and specificity. Due to the nature of these outcomes it is not obvious to specify a hypothesis and test for it. Other factors may influence the publication process, but it is not clear whether these factors are associated with particular databases.

A previous review has shown that failure to search multiple databases to identify studies for inclusion in DTA reviews misses relevant studies.(2) However, this review did not investigate the impact of these missing studies on the results of the review. Restricting a review to studies indexed on a single database, for example MEDLINE, is only problematic if this leads to biased results. We would assume that reviews based exclusively on studies indexed on MEDLINE could have biased results if the results of those studies differ systematically from relevant studies indexed on other databases. We therefore aimed to assess whether restriction of databases influences the estimation of measures of accuracy in DTA reviews.

METHODS

Identification of reviews

MEDLINE was searched through the PubMed interface to identify DTA reviews published between January 2006 and January 2011. The methodological filter of Devillé (10) was applied to identify reviews covering diagnostic test accuracy combined with the review filter that is available in PubMed to identify systematic reviews. Search results were limited to 622 journals that had an impact factor ≥4 in 2010 (11) and were accessible through the medical library of the University of Amsterdam. The complete search strategy can be found in Appendix 1. In addition, the Cochrane Database of Systematic Reviews (CDSR) was searched in March 2011 for all DTA reviews. The literature search and the presentation of the review was structured according to the PRISMA guidelines.(12)

Inclusion criteria

We included reviews in which the authors evaluated the diagnostic accuracy of one or more tests against a reference standard and reported measure of accuracy: the Diagnostic Odds Ratio (DOR), sensitivity and specificity. We only included DTA reviews that had conducted a meta-analysis and that had searched MEDLINE and at least one other biomedical database. We excluded narrative reviews, genomic reviews, animal reviews, reviews that had applied a language or quality restriction, reviews that had assessed the analytical validity of tests and reviews that only evaluated other measures of diagnostic performance such as reproducibility and reliability. Two reviewers independently assessed titles and abstracts of the references identified by the electronic search for relevance. Inclusion screening of full text articles was conducted independently by two reviewers. A third reviewer was consulted in case of disagreement. Only meta-analyses that included both studies indexed on MEDLINE and studies not indexed on MEDLINE were included for further analysis.

Data extraction

Descriptive characteristics (author, publication year, test under evaluation, and purpose of the test) and full references for each included primary study were extracted from each review. Data to populate two-by-two tables (the number of true positives, false positives, false negatives, and true negatives) were extracted for all individual studies of all included meta-analyses. We contacted the authors of reviews when two-by-two tables were not reported in the review. In case of no response we sent two reminders requesting the missing data. When no reply was received from the authors we extracted data from the primary studies ourselves.

Comprehensiveness of the searches

We aimed to assess whether the comprehensiveness of the search was associated with finding studies not indexed on MEDLINE.(2) We assessed the comprehensiveness of the search according to the AMSTAR checklist.(13) AMSTAR, a measurement tool to assess the methodological quality of a systematic review, has several items that determine the comprehensiveness of a search for a systematic review: 1. at least two electronic sources should been searched, 2. the years and names of databases should be reported, 3. key words and/or MeSH or EMTREE index terms should be provided, and 4. extra effort should be made to identify extra studies. Each of these items was scored individually, although in AMSTAR all are scored as a single item (item 3). We also added two extra items that we believe contribute to the comprehensiveness of a search: 5. whether the search was performed without using a search filter for DTA studies as we know that can lead to missing studies (8;14) and 6. whether the full search strategy was available. Each item could be scored with 1 when the item was fulfilled or with 0 when the item was not fulfilled or unclear. Altogether, each search could thus score between 0 and 6.

Data analysis

We assessed which of the primary studies included in each review were indexed on MEDLINE by means of a known item search. A known-item search implies a user who is looking for one particular study; in our case the study included in the review for which we had the full reference extracted from the review.(9) References that could be identified in MEDLINE were labeled as 'in MEDLINE' (iM) and those that were not as 'Not in MEDLINE' (NiM).

We selected the primary meta-analysis from each systematic review which we defined as the (sub)set of clinically relevant studies on which the conclusions of the review were based. The selection process was double checked by a second reviewer for a subset of reviews (20%) and those considered most complicated to classify (AF, PW). For reviews that assessed more than one index test we selected all meta-analyses that contributed to the conclusion.

First, the impact of not including NiM studies was measured quantitatively by redoing the meta-analyses and calculating the DOR, sensitivity and specificity: once with all studies and once without NiM studies. Analyses were performed using the bivariate random effects model (15) in Stata version 10.0.(16) The DOR, sensitivity and specificity were calculated with corresponding 95% confidence intervals (CIs). Per meta-analysis we quantified the effect of excluding NiM studies by estimating the summary relative diagnostic odds ratio (RDOR): DORiM / DORiM+NiM. We also calculated the asymptotic variance of the RDOR. The log(RDORs) were then pooled by the use of a random effects generic inverse variance model to estimate the average effect of restricting analyses to NiM studies. Similarly to the RDOR, the summary of the difference in sensitivity and specificity between iM and NiM were estimated. The statistical methods are explained further in Appendix 1.

We used a sensitivity analysis to assess the impact of leaving out NiM studies for the meta-analyses with largest proportion of NiM studies (NiM/(iM+NiM to maximise the likelihood of finding an effect of missing studies by a search limited to MEDLINE).

We assessed if there was an association between the comprehensiveness of the search and finding all studies in MEDLINE. The comprehensiveness of the search was summarized by calculating a score for the search characteristics (1 point for each item fulfilled). This score (range 0 to 6) was used as a continuous independent variable in a logistic regression

RESULTS

The searches identified 615 hits of which 116 were considered potentially relevant and were assessed for inclusion based on full text papers. We identified 42 reviews but in 24 (57%) of these, including two Cochrane reviews, all included primary studies were indexed on MEDLINE and so these were not investigated further. The 18 reviews with contrast presented 52 meta-analyses, 39 meta-analyses were considered as primary meta-analyses as they contributed to the conclusions of the reviews. Eight reviews including 18 primary meta-analyses were excluded because of the characteristics of the primary meta-analyses (all studies included in the primary meta-analyses were indexed on MEDLINE, lack of specification of which primary studies were included in the meta-analyses, or lack of information to populate the two-by-two tables). The final ten reviews included 18 meta-analyses of which three did not have contrast, leaving 15 meta-analyses for inclusion. The selection process is presented graphically in Figure 1, and the basic characteristics of the included reviews are presented in Table 1.

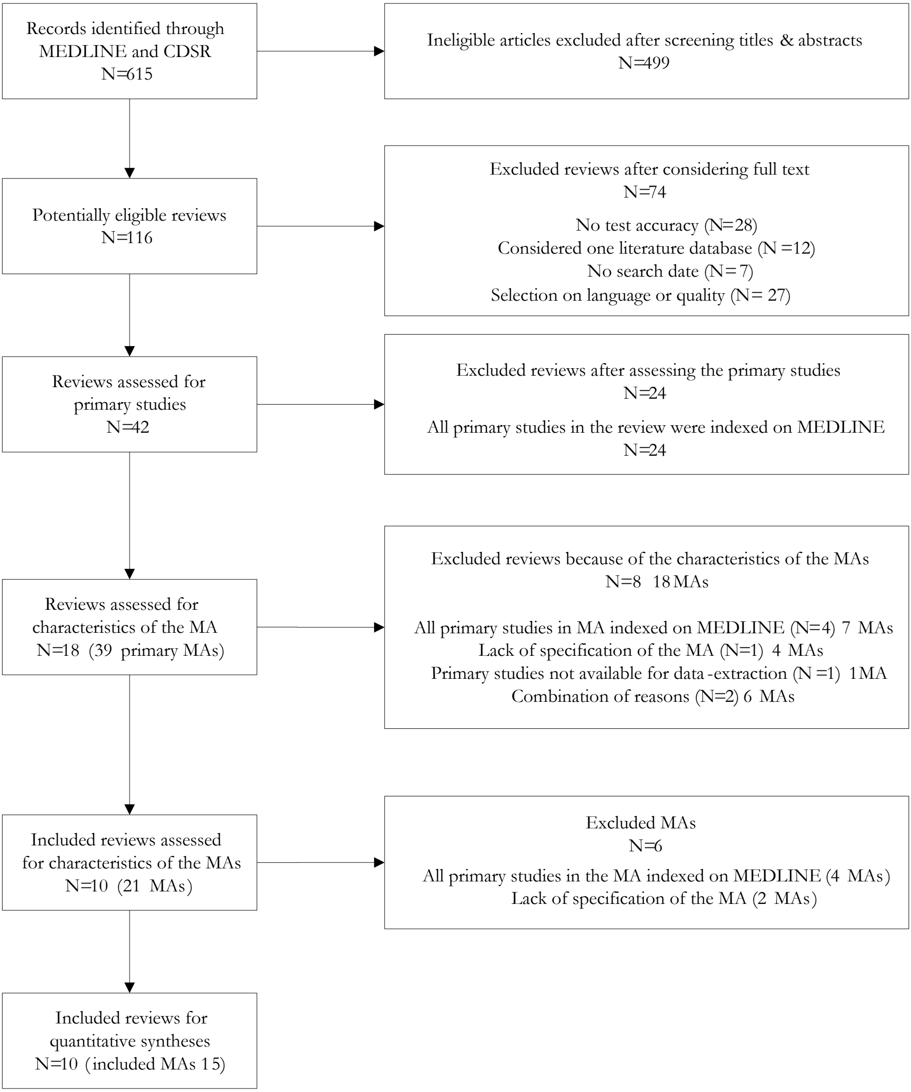

Records identified through MEDLINE and CDSR N=615
Ineligible articles excluded after screening titles & abstracts N=499
Potentially eligible reviews N=116
Excluded reviews after considering full text N=74
No test accuracy (N=28)
Considered one literature database (N =12)
No search date (N= 7)
Selection on language or quality (N= 27)
Reviews assessed for primary studies N=42
Excluded reviews after assessing the primary studies N=24
All primary studies in the review were indexed on MEDLINE N=24
Reviews assessed for characteristics of the MA N=18 (39 primary MAs)
Excluded reviews because of the characteristics of the MAs N=8 18 MAs
All primary studies in MA indexed on MEDLINE (N=4) 7 MAs
Lack of specification of the MA (N=1) 4 MAs
Primary studies not available for data-extraction (N =1) 1MA
Combination of reasons (N=2) 6 MAs
Included reviews assessed for characteristics of the MAs N=10 (21 MAs)
Excluded MAs N=6
All primary studies in the MA indexed on MEDLINE (4 MAs)
Lack of specification of the MA (2 MAs)
Included reviews for quantitative syntheses N=10 (included MAs 15)

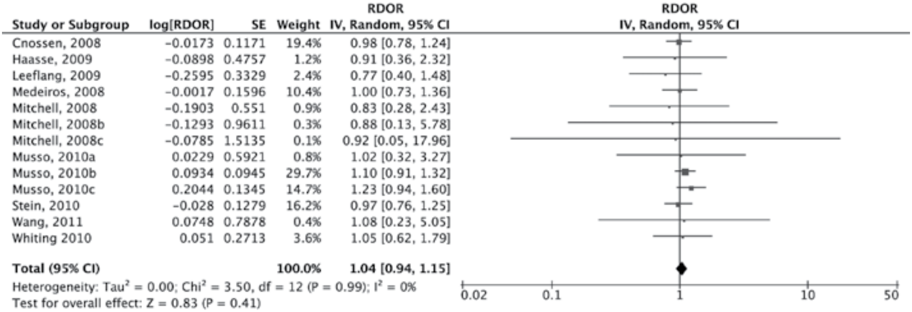
Figure 1. *Flow chart of selection process of systematic reviews and their primary meta-analyses*

Table 1. *Characteristics of included systematic reviews (n=10) and meta-analyses (n=15)*

Review author (publication year)	Diagnostic test	Diagnostic test	Search score (1 to 5)	Studies in the review N (% NiM)	Included MA N	Studies in MA N (%NiM)
Cnossen (2008) (17)	All uterine artery Doppler indices	Pre-ecplamsia and intrauterine growth restriction	5	132 (10%)	1	17 (18%)
Diel (2010) (18)	Interferon- y release assays	Active tuberculosis	1	124 (0.8%)	1	19 (11%)†
Haase (2009) (19)	Neutrophil Gelatinase-Associated Lipocalin	Acute kidney injury	2	19 (21%)	1	19 (21%)
Leeflang (2009) (20)	Serum galactomannan ELISA	Invasive aspergillosis (IA)	5	42 (7%)	1	18 (11%)
Medeiros (2008) (21)	Dynamic contrast-enhanced breast magnetic resonance	Breast lesion and cancer	3	69 (3%)	1	69 (1%)
Mitchell (2008) (22)	One or two simple verbal questions	Depression in cancer settings	3	10 (10%)	3	9 (11%) 5 (20%) 3 (33%)
Musso (2010) (23)	ELISA (Cytokeratin-18), NAFLD fibrosis score and Fibroscan	Non alcoholic fatty liver disease	1	32 (25%)	3	11 (27%) 9 (22%) 6 (17%)
Stein (2009) (24)	Pelvic ultrasonography	Ectopic pregnancy	2	10 (10%)	1	10 (10%)
Wang (2011) (25)	Myocardial perfusion scintigraphy and dobutamine stress echocardiography	Coronary artery disease	5	17 (6%)	1	7 (14%)
Whiting (2010) (26)	Anti–Citrullinated Peptide Antibodies	Rheumatoid Arthritis	3	151 (17%)	2	15 (7%) 138 (15%)*
Total	-	-	-	**482 (11%)**	**15**	**355 (16%)**

*N=number; MA=meta-analysis; NiM = not in Medline; †data limited to diseased participants; * meta-analysis only included to study the robustness of the results in a sensitivity analysis*

The mean percentage of NiM studies in the included meta-analyses was 16% (range 1.0 to 33%). The RDORs comparing the DOR for iM studies with the DOR of all studies ranged from 0.77 to 1.23 with a pooled RDOR of 1.04 (95% CI 0.94 to 1.15) suggesting that restricting searches to MEDLINE may slightly overestimate the results compared to searching a broader range of databases (Figure 2). However, the point estimate is very close to 1 and not significant. None of the individual meta-analyses were significantly affected by leaving out NiM studies. There were no statistically significant differences between the sensitivity and specificity of iM studies and all studies (difference in sensitivity -0.08%; 95% CI -1% to 1% range between
-5.8% to 4.8% and difference in specificity: -0.1%; 95% CI -0.8% to 1% range between -2.2% and 2.3%) (Figure 3 and 4). Heterogeneity was minimal for all meta-analyses.



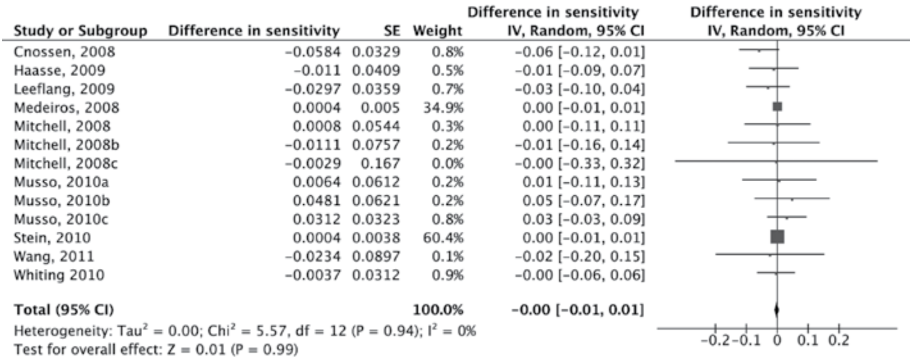
Study or Subgroup	log[RDOR]	SE	Weight	RDOR IV, Random, 95% CI
Cnossen, 2008	−0.0173	0.1171	19.4%	0.98 [0.78, 1.24]
Haasse, 2009	−0.0898	0.4757	1.2%	0.91 [0.36, 2.32]
Leeflang, 2009	−0.2595	0.3329	2.4%	0.77 [0.40, 1.48]
Medeiros, 2008	−0.0017	0.1596	10.4%	1.00 [0.73, 1.36]
Mitchell, 2008	−0.1903	0.551	0.9%	0.83 [0.28, 2.43]
Mitchell, 2008b	−0.1293	0.9611	0.3%	0.88 [0.13, 5.78]
Mitchell, 2008c	−0.0785	1.5135	0.1%	0.92 [0.05, 17.96]
Musso, 2010a	0.0229	0.5921	0.8%	1.02 [0.32, 3.27]
Musso, 2010b	0.0934	0.0945	29.7%	1.10 [0.91, 1.32]
Musso, 2010c	0.2044	0.1345	14.7%	1.23 [0.94, 1.60]
Stein, 2010	−0.028	0.1279	16.2%	0.97 [0.76, 1.25]
Wang, 2011	0.0748	0.7878	0.4%	1.08 [0.23, 5.05]
Whiting 2010	0.051	0.2713	3.6%	1.05 [0.62, 1.79]
Total (95% CI)			**100.0%**	**1.04 [0.94, 1.15]**

Heterogeneity: Tau² = 0.00; Chi² = 3.50, df = 12 (P = 0.99); I² = 0%
Test for overall effect: Z = 0.83 (P = 0.41)

Figure 2. *Forest plot for the relative Diagnostic Odds Ratio (RDOR) indicating the difference between including only MEDLINE studies in the primary meta-analysis versus including all studies (NB. For some reviews multiple meta-analyses were included as they considered different tests).*

The meta-analyses selected as the primary meta-analyses were also those with the largest contrasts in NiM/(iM+NiM) with the exception of one review. For this one review a greater number of studies would have been missed by a different meta-analysis (NiM 7% primary meta-analyses versus NiM 15% for other meta-analysis). A sensitivity analysis showed that including this meta-analysis would not have led to different conclusions (RDOR of 1.07; 95% CI 0.98 to 1.17).

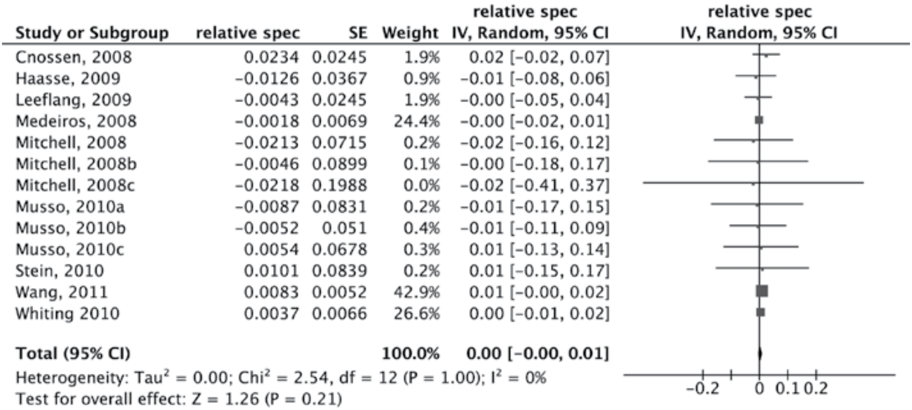
There was a non-significant association between the comprehensiveness of the search and finding all studies in MEDLINE (OR 0.82 (95% CI 0.53 – 1.26)).
A score below 1 indicates that a higher search score is associated with reducing the odds of finding all studies in MEDLINE.



Study or Subgroup	Difference in sensitivity	SE	Weight	Difference in sensitivity IV, Random, 95% CI
Cnossen, 2008	−0.0584	0.0329	0.8%	−0.06 [−0.12, 0.01]
Haasse, 2009	−0.011	0.0409	0.5%	−0.01 [−0.09, 0.07]
Leeflang, 2009	−0.0297	0.0359	0.7%	−0.03 [−0.10, 0.04]
Medeiros, 2008	0.0004	0.005	34.9%	0.00 [−0.01, 0.01]
Mitchell, 2008	0.0008	0.0544	0.3%	0.00 [−0.11, 0.11]
Mitchell, 2008b	−0.0111	0.0757	0.2%	−0.01 [−0.16, 0.14]
Mitchell, 2008c	−0.0029	0.167	0.0%	−0.00 [−0.33, 0.32]
Musso, 2010a	0.0064	0.0612	0.2%	0.01 [−0.11, 0.13]
Musso, 2010b	0.0481	0.0621	0.2%	0.05 [−0.07, 0.17]
Musso, 2010c	0.0312	0.0323	0.8%	0.03 [−0.03, 0.09]
Stein, 2010	0.0004	0.0038	60.4%	0.00 [−0.01, 0.01]
Wang, 2011	−0.0234	0.0897	0.1%	−0.02 [−0.20, 0.15]
Whiting 2010	−0.0037	0.0312	0.9%	−0.00 [−0.06, 0.06]
Total (95% CI)			**100.0%**	**−0.00 [−0.01, 0.01]**

Heterogeneity: Tau² = 0.00; Chi² = 5.57, df = 12 (P = 0.94); I² = 0%
Test for overall effect: Z = 0.01 (P = 0.99)

Figure 3. *Forest plot for the relative sensitivity indicating the difference between including only MEDLINE studies in the primary meta-analysis versus including all studies (NB. For some reviews multiple meta-analyses were included as they considered different tests).*



Study or Subgroup	relative spec	SE	Weight	relative spec IV, Random, 95% CI
Cnossen, 2008	0.0234	0.0245	1.9%	0.02 [−0.02, 0.07]
Haasse, 2009	−0.0126	0.0367	0.9%	−0.01 [−0.08, 0.06]
Leeflang, 2009	−0.0043	0.0245	1.9%	−0.00 [−0.05, 0.04]
Medeiros, 2008	−0.0018	0.0069	24.4%	−0.00 [−0.02, 0.01]
Mitchell, 2008	−0.0213	0.0715	0.2%	−0.02 [−0.16, 0.12]
Mitchell, 2008b	−0.0046	0.0899	0.1%	−0.00 [−0.18, 0.17]
Mitchell, 2008c	−0.0218	0.1988	0.0%	−0.02 [−0.41, 0.37]
Musso, 2010a	−0.0087	0.0831	0.2%	−0.01 [−0.17, 0.15]
Musso, 2010b	−0.0052	0.051	0.4%	−0.01 [−0.11, 0.09]
Musso, 2010c	0.0054	0.0678	0.3%	0.01 [−0.13, 0.14]
Stein, 2010	0.0101	0.0839	0.2%	0.01 [−0.15, 0.17]
Wang, 2011	0.0083	0.0052	42.9%	0.01 [−0.00, 0.02]
Whiting 2010	0.0037	0.0066	26.6%	0.00 [−0.01, 0.02]
Total (95% CI)			**100.0%**	**0.00 [−0.00, 0.01]**

Heterogeneity: Tau² = 0.00; Chi² = 2.54, df = 12 (P = 1.00); I² = 0%
Test for overall effect: Z = 1.26 (P = 0.21)

Figure 4. *Forest plot for the relative specificity indicating the difference between including only MEDLINE studies in the primary meta-analysis versus including all studies (NB. For some reviews multiple meta-analyses were included as they considered different tests).*

DISCUSSION

Our results suggest that restricting searches to MEDLINE would only have had a negligible, non-significant impact on the summary estimates of the meta-analyses included in our study. In addition, for the majority of the reviews (57%) assessed in our meta-epidemiological study, all of the included studies were indexed on MEDLINE although the review authors had searched at least one other database. These reviews would not have been impacted at all had the search been restricted to MEDLINE.

Our results differ from evidence found for intervention reviews. For intervention reviews it has been shown that studies that are uniquely indexed on EMBASE on average yield 27% lower effect estimates than studies from other databases.(9) These, and similar results from other studies (27;28), are the founding to dismiss reviews that have searched in only one electronic database as low quality.(13) Our study indicates that the summary estimates of the meta-analyses in DTA reviews that searched only MEDLINE are not significantly affected. Consequently, DTA reviews may therefore not automatically be dismissed as 'low quality' if they have searched only one electronic database.

Although our results indicate that limiting the search to MEDLINE studies did not alter the summary estimates of the reviews, there are other consequences of a more restrictive search that should be considered. Missing relevant studies that could potentially have been included in the review will decrease the power of the meta-analysis and make the confidence intervals around summary measures of accuracy wider. This will potentially lower the confidence in the results of the meta-analysis. Heterogeneity is a feature of almost all DTA reviews.(29) Investigation of heterogeneity is important to determine the most reliable estimate of accuracy. Heterogeneity can be investigated using meta-regression or subgroup analysis but this requires sufficient power. Missing studies will therefore also decrease the potential to investigate heterogeneity. Second, different databases use different indexing systems and almost no search strategy has perfect sensitivity; even strategies designed to be very sensitive miss relevant studies.(14) Therefore, searching in multiple databases may increase the likelihood of identifying studies also available in MEDLINE.

This study has some limitations. The number of meta-analyses that could be included in our analysis was small, also because most meta-analyses had no NiM studies and could therefore not add information to our analysis. A second reason was poor reporting of studies, proving too few details to include the study for analysis. Poor reporting of DTA studies is a common problem in DTA studies.(30) To prevent from this, we selected reviews with an impact factor of ≥4 that have been shown to have higher quality of reporting. Still, this seemed to be insufficient. Due to the small number of included meta-analyses, we had low power for assessing the effect of limiting the meta-analysis to iM studies. Moreover, the fact that we did not identify a significant difference for the summary estimates does not exclude that no difference does exist. In addition, if we had been able to estimate the difference between the various diagnostic parameters assessed in the NiM and iM separately, the contrast between these two subgroups might have been bigger, but unfortunately the number of NiM studies was too low to perform a separate meta-analysis upon. Further, the low power of our analysis and poor reporting of reviews also limited our ability to investigate the relationship between the comprehensiveness of the search and finding studies outside of MEDLINE. It would have been more interesting to assess the specific features of the searches individually to provide insight into which strategies are most likely to identify studies beyond MEDLINE but the analysis did not have sufficient power to do so. It would been preferable to have used a validated checklist for this purpose like the Canadian Agency for Drugs and Technologies in Health peer review checklist which is used by information specialists to peer review search strategies.(31) However, these tools require access to the full strategy, but these are very rarely published in a review.

A known-item search was used to assess whether studies were indexed on MEDLINE. This type of search distinguishes itself from searches used in systematic reviews. As the name indicates, the study that is searched for is known in a known-item search, while for a systematic review explorative searches are used. Explorative searches depend on the indexing system of the databases. Therefore it is possible that studies indexed on MEDLINE had not been found by the review authors when using an explorative search, although we have identified it using the known-item search.(2) Explorative searches in MEDLINE are likely to miss studies due to inadequate assignment of MeSH terms.(27) Performing the known-item search for this study therefore may have resulted in an overestimation of the number of iM studies. Access to the full search strategies would have enabled us to replicate the original searches and to determine which studies would be identified by searching MEDLINE but these full search strategies are rarely reported. Due to this limitation we might have overestimated the number of studies a reviewer will identify on MEDLINE but did not affect our result that accuracy results of iM studies are not significantly different than accuracy results in studies uniquely indexed on other electronic databases.

Despite the fact that the results of our study are based on a small sample of meta-analyses, our confidence in the results are strengthen by the fact that no individual included review had significant differences between the DOR based on all studies compared to the DOR based on the studies that were iM studies only. In addition, when the meta-analyses with the largest proportion of NiM studies were included, the analyses with the greatest number of missing studies, still no significant difference was found. Another strength of our study is that the inclusion of reviews was restricted to those that had no inclusion limitations on language or quality. A selective inclusion of primary studies for the reviews could otherwise have interfered with our results. Additionally, we investigated the association between the search characteristics of the included reviews and finding all studies in MEDLINE; no association was found. A further strength is that the analyses to estimate the impact of leaving out NiM studies were controlled for the dependency that exist for every meta-analysis with all studies included and the meta-analysis including the subset of MEDLINE studies.

Although this study has been undertaken to answer challenges with respect to searching for primary studies, it also gives insights into the underlying topic of selective reporting. It is well known that publication and dissemination bias can influence the results of a meta-analysis.(32) However, evidence for its existence and manifestation in the field of diagnostic test accuracy is scarce.(33;34) Our results did not identity a systematic difference in accuracy measures between studies indexed on MEDLINE and other database indexed studies, but this does not exclude that there is a systematic difference between MEDLINE and other databases. It would be worthwhile to investigate if our results also apply for specific fields of medical test accuracy research as the results of this study may differ between fields.

CONCLUSION

Restricting the search to MEDLINE did not have a significant impact on the summary estimates of the reviews included in our study. Missing relevant studies, however, may lower the precision of the summary measures of accuracy and the power of the analyses to investigate heterogeneity. When financial resources are low, restricting searches to MEDLINE does not seem to affect the summary estimates of the review. However, the impact for individual reviews is still unpredictable.

ACKNOWLEDGEMENTS

We would like to thank René Spijker for his contributions to the development of the search strategy and his comments on the manuscript. We also thank Cynthia Ho for her contributions to the data-extraction. Help with translation of the Chinese studies by Yaolong Chen of the Chinese GRADE Center was greatly appreciated. Finally, we like to thank Conny Steenman-Marcusse for her grammar suggestions throughout the manuscript.

AUTHORS'CONTRIBUTIONS

Wynanda A van Enst contributed to the development of the protocol, study selection, data-extraction and analysis and wrote the manuscript. Rob J P M Scholten contributed to the development of the protocol, data-analysis and commented on the manuscript. Penny Whiting contributed to the development of the protocol, study selection and commented on the manuscript. Aeilko H Zwinderman developed the statistical analysis, contributed to the data-analysis and wrote the statistical methods for the manuscript. Lotty Hooft contributed to the development of the protocol, data extraction and commented on the manuscript.

REFERENCE LIST

1. de Vet HCW, Eisinga A, Riphagen II, Aertgeerts B, Pewsner D. Chapter 7: Searching for Studies. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. Version 0.4 [updated September 2008] ed. The Cochrane Collaboration; 2008.
2. Whiting P, Westwood M, Burke M, Sterne J, Glanville J. Systematic reviews of test accuracy should search a range of databases to identify primary studies. Journal of Clinical Epidemiology 2008;61:357-64.
3. Indiana University. Ulrichsweb Global Serials Directory. 2013.
4. Lefebvre C, Manheimer E, Glanville J. Chapter 6. Searching for studies. In: Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions. 5.1.0 ed. Wiley; 2008.
5. Bachmann LM, Coray R, Estermann P, Ter RG. Identifying diagnostic studies in MEDLINE: reducing the number needed to read. JAMIA 2002 Nov;9(6):653-8.
6. Vincent S, Greenley S, Beaven O. Clinical Evidence diagnosis: Developing a sensitive search strategy to retrieve diagnostic studies on deep vein thrombosis: a pragmatic approach. Health Info Libr J 2003 Sep;20(3):150-9.
7. Beynon R, Leeflang MM, McDonald S, Eisinga A, Mitchell RL, Whiting P, et al. Search strategies to identify diagnostic accuracy studies in MEDLINE and EMBASE. Cochrane Database Syst Rev 2013;9:MR000022.
8. 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.
9. Sampson M, Platt R, StJohn PD, Moher D, Klassen TP, Pham B, et al. Should meta-analysts search Embase in addition to Medline? Journal of Clinical Epidemiology 2003;56:943-55.
10. Deville WL, Bezemer PD, Bouter LM. Publications on diagnostic test evaluation in family medicine journals: an optimal search strategy. J Clin Epidemiol 2000 Jan;53(1):65-9.
11. Smidt N, Rutjes AW, van der Windt DA, Ostelo RW, Reitsma JB, Bossuyt PM, et al. Quality of reporting of diagnostic accuracy studies. Radiology 2005 May;235(2):347-53.
12. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for

systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097.

13. Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Medical Research Methodology 2007;15(7):10.

14. Whiting P, Westwood M, Beynon R, Burke M, Sterne JA, Glanville J. Inclusion of methodological filters in searches for diagnostic test accuracy studies misses relevant studies. Journal of Clinical Epidemiology 2011 Jun;64(6):602-7.

15. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. Journal of Clinical Epidemiology 2005 Oct;58(10):982-90.

16. Stata Statistical Software: Release 10 [computer program]. College Station: TX: StataCorp LP; 2007.

17. Cnossen JS, Morris RK, ter RG, Mol BW, van der Post JA, Coomarasamy A, et al. Use of uterine artery Doppler ultrasonography to predict pre-eclampsia and intrauterine growth restriction: a systematic review and bivariable meta-analysis. CMAJ 2008 Mar 11;178(6):701-11.

18. Diel R, Loddenkemper R, Nienhaus A. Evidence-based comparison of commercial interferon-gamma release assays for detecting active TB: a metaanalysis. Chest 2010 Apr;137(4):952-68.

19. Haase M, Bellomo R, Devarajan P, Schlattmann P, Haase-Fielitz A. Accuracy of neutrophil gelatinase-associated lipocalin (NGAL) in diagnosis and prognosis in acute kidney injury: a systematic review and meta-analysis. Am J Kidney Dis 2009 Dec;54(6):1012-24.

20. Leeflang MM, Debets-Ossenkopp YJ, Visser CE, Scholten RJ, Hooft L, Bijlmer HA, et al. Galactomannan detection for invasive aspergillosis in immunocompromized patients. Cochrane Database Syst Rev 2008;(4):CD007394.

21. Medeiros LR, Duarte CS, Rosa DD, Edelweiss MI, Edelweiss M, Silva FR, et al. Accuracy of magnetic resonance in suspicious breast lesions: a systematic quantitative review and meta-analysis. Breast Cancer Res Treat 2011 Jan 8.

22. Mitchell AJ. Are one or two simple questions sufficient to detect depression in cancer and palliative care? A Bayesian meta-analysis. Br J Cancer 2008 Jun 17;98(12):1934-43.

23. Musso G, Gambino R, Cassader M, Pagano G. Meta-analysis: Natural history of non-alcoholic fatty liver disease (NAFLD) and diagnostic accuracy of non-invasive tests for liver disease severity. Ann Med 2010 Nov 2.

24. Stein SC, Fabbri A, Servadei F, Glick HA. A critical comparison of clinical decision instruments for computed tomographic scanning in mild closed traumatic brain injury in adolescents and adults. Ann Emerg Med 2009 Feb;53(2):180-8.

25. Wang LW, Fahim MA, Hayen A, Mitchell RL, Lord SW, Baines LA, et al. Cardiac testing for coronary artery disease in potential kidney transplant recipients: a systematic review of test accuracy studies. Am J Kidney Dis 2011 Mar;57(3):476-87.

26. Whiting PF, Smidt N, Sterne JA, Harbord R, Burton A, Burke M, et al. Systematic review: accuracy of anti-citrullinated Peptide antibodies for diagnosing rheumatoid arthritis. Ann Intern Med 2010 Apr 6;152(7):456-64.

27. Suarez-Almazor ME, Belseck EF, Homik J, Dorgan M, Ramos-Remus C. Identifying clinical trials in the medical literature with electronic databases: MEDLINE alone is not enough. Controlled Clinical Trials 2000;21:476-87.

28. Dickersin K, Manheimer E, Wieland S, Robinson KA, Lefebvre C, McDonald S. Development of the Cochrane Collaboration's CENTRAL Register of controlled clinical trials. Eval Health Prof 2002 Mar;25(1):38-64.

29. Dinnes J, Deeks J, Kirby J, Roderick P. A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy. Health Technol Assess 2005 Mar;9(12):1-113, iii.

30. Korevaar DA, van Enst WA, Spijker R, Bossuyt PM, Hooft L. Reporting quality of diagnostic accuracy studies: a systematic review and meta-analysis of investigations on adherence to STARD. Evid Based Med 2013 Dec 24.

31. Canadian Agency for Drugs and Technologies in Health. CADTH peer review checklist for search strategies [Internet]. 2013. Ottawa, The Agency.

32. Song F, Parekh S, Hooper L, Loke YK, Ryder J, Sutton AJ, et al. Dissemination and publication of research findings: an updated review of related biases. Health Technol Assess 2010 Feb;14(8):iii, ix-iii,193.

33. DeAngelis CD, Drazen JM, Frizelle FA, Haug C, Hoey J, Horton R, et al. Clinical trial registration: a statement from the International Committee of

Medical Journal Editors. JAMA 2004 Sep 15;292(11):1363-4.

34. Song F, Khan KS, Dinnes J, Sutton AJ. Asymmetric funnel plots and publication bias in meta-analyses of diagnostic accuracy. Int Journal of Epidemiology 2002 Feb;31(1):88-95.

CHAPTER 4

Incorporating quality assessments of primary studies in the conclusions of diagnostic accuracy reviews: a cross-sectional study

Eleanor A. Ochodo
Wynanda A. van Enst
Christiana A. Naaktgeboren
Joris A.H. de Groot
Lotty Hooft
Karel G.M. Moons
Johannes B. Reitsma
Patrick M. Bossuyt
Mariska M.G. Leeflang

BMC Medical Research Methodology, 2014;14(1):70

ABSTRACT

Background Drawing conclusions from systematic reviews of test accuracy studies without considering the methodological quality (risk of bias) of included studies may lead to unwarranted optimism about the value of the test(s) under study. We sought to identify to what extent the results of quality assessment of included studies are incorporated in the conclusions of diagnostic accuracy reviews.

Methods We searched MEDLINE and EMBASE for test accuracy reviews published between May and September 2012. We examined the abstracts and main texts of these reviews to see whether and how the results of quality assessment were linked to the accuracy estimates when drawing conclusions.

Results We included 65 reviews of which 53 contained a meta-analysis. Sixty articles (92%) had formally assessed the methodological quality of included studies, most often using the original QUADAS tool (n = 44, 68%). Quality assessment was mentioned in 28 abstracts (43%); with a majority (n = 21) mentioning it in the methods section. In only 5 abstracts (8%) were results of quality assessment incorporated in the conclusions. Thirteen reviews (20%) presented results of quality assessment in the main text only, without further discussion. Forty-seven reviews (72%) discussed results of quality assessment; the most frequent form was as limitations in assessing quality (n = 28). Only 6 reviews (9%) further linked the results of quality assessment to their conclusions, 3 of which did not conduct a meta-analysis due to limitations in the quality of included studies. In the reviews with a meta-analysis, 19 (36%) incorporated quality in the analysis. Eight reported significant effects of quality on the pooled estimates; in none of them these effects were factored in the conclusions.

Discussion/Conclusion While almost all recent diagnostic accuracy reviews evaluate the quality of included studies, very few consider results of quality assessment when drawing conclusions. The practice of reporting systematic reviews of test accuracy should improve if readers not only want to be informed about the limitations in the available evidence, but also on the associated implications for the performance of the evaluated tests.

INTRODUCTION

Systematic reviews of diagnostic test accuracy form a fundamental part of evidence-based practice (1;2). An essential part of a systematic review is the evaluation of the risk of bias (3) also referred to as assessment of methodological quality (4). Limitations in the design and conduct of the study may lead to overestimation of the accuracy of the test under study (5;6). This is of concern, because tests introduced in practice based on weak evidence may lead to misdiagnosis, improper management of patients and, subsequently, poor health outcomes (7-9). Such limited evidence could also lead to unnecessary testing and avoidable health care costs (7).

The Quality Assessment for Diagnostic Accuracy Studies tool (QUADAS) has been developed and introduced to evaluate the methodological quality of studies included in systematic reviews of test accuracy (10). A revised version, QUADAS-2, was introduced in 2011. The revised instrument considers methodological quality in terms of risk of bias and concerns regarding the applicability of findings to the research question. It does so in four key domains: patient selection, index test, reference test, and patient flow (11). The QUADAS-2 tool is recommended by the U.K National Institute for Health and Clinical Excellence, the Cochrane Collaboration, and the U.S. Agency for Healthcare Research and Quality.

The use of QUADAS in test accuracy reviews to assess the methodological quality of included primary studies is increasing. Willis and Quigley reported that 40% of diagnostic reviews published between 2006 and 2008 used the QUADAS tool (12), while Dahabreh and colleagues reported that, in 2004, about 2% of diagnostic reviews used QUADAS, while 44% did so in 2009 (13).

Simply assessing quality without interpreting and using the results to draw conclusions is not sufficient in evidence synthesis. The results from quality assessment should be used to make inferences about the validity of the results.

The challenge of incorporating quality assessments of the included studies into the overall findings of a review is well known in intervention reviews. Moja and colleagues (14) reported that just about half of the 965 reviews they examined had incorporated the results of quality assessment in their analysis and interpretation of the results of their studies. A similar study done almost 10 years later by Hopewell and colleagues (15) reported that only 41% of the 200 reviews they examined incorporated the risk of bias assessment into the interpretation of their conclusions. The challenge of incorporating results of quality assessment in the conclusions may also be present in diagnostic accuracy reviews.

Readers, who usually have limited or basic knowledge of the methodological process involved in diagnostic reviews, often focus exclusively on the conclusion sections of a review when arriving at a judgment about a test's performance (16). In this regard, drawing conclusions without considering the risk of bias in included studies may lead to unwarranted optimism about the value of the test(s) under study. We sought to identify to what extent – and, if so, how – quality assessment is incorporated in the conclusions of diagnostic accuracy reviews.

METHODS

This study was part of a larger meta-epidemiological study to examine the methodology used in recent test accuracy reviews. Since diffusion of methods takes time, we focused on recently published reviews. On 12th September 2012, we identified a convenience sample of test accuracy reviews indexed in the databases MEDLINE and EMBASE between 1st May and 11th September 2012 using the search strategy available in Additional file 1.

Eligible were reviews with a systematic search and methodology in appraising and summarising studies that evaluated a medical test against a reference standard. These reviews could present summary accuracy measures generated in a meta-analysis or present a range of accuracy measures without a summary measure. We included reviews published in English and which evaluated human studies dealing with patient data (as opposed to specimen data). We excluded individual patient data reviews and reviews evaluating the accuracy of prognostic tests in predicting future events. The methodology for evaluating quality in reviews of prognostic tests is less well developed than that for diagnostic tests.

The data extraction form was pilot tested by performing double data extraction on a third of the articles (by E.O., W.E., C.N., J.G., L.H., and M.L.). Discrepancies were discussed and unclear questions on the form were made more specific. Data extraction was then performed by one researcher (by E.O., W.E., C.N., and M.L.) using the standardized form and checked by another researcher (by E.O., W.E., C.N., and M.L.). Disagreements were resolved through discussion and when necessary by including a third reviewer (P.B.).

As conclusions are influenced largely by the methods used and the results produced in a review, we first examined every included review to check if methodological quality of included studies had been assessed using the recommended tool, QUADAS or QUADAS-2 (10;11), or any other tool that the authors specified as a system to assess risk of bias.

We examined the abstracts to check if methodological quality was mentioned in any of the sections (background, methods, results and conclusions). Abstracts are the most commonly read part of articles and readers often rely on abstracts to give them a snapshot of the content of reviews; where full texts cannot be accessed, judgments of a test's performance may be made on abstracts alone (17-19).

We examined the main body of the review to check if the methodological quality of included studies was assessed, which tool had been used to assess quality, how results of quality assessments were presented, if the quality of studies had influenced the decision to perform a meta-analysis, if and how an assessment of quality was incorporated into the analysis, and if and how the results of quality assessment were discussed and eventually used in drawing conclusions about the test.

We regarded quality as being incorporated into the conclusions of the review when results of quality assessment of the included studies, or limitations surrounding quality assessment, were considered together with the accuracy estimates of the diagnostic tests in drawing conclusions about the performance of the test(s) under evaluation. We distinguished between drawing conclusions about test performance and making recommendations for future research. Conclusions of test performance are usually based solely on the results of the review and could be used as guidance for clinical practice, whereas recommendations for research are generally made after considering additional information not necessarily investigated in the review itself.

RESULTS

Search results

The initial search identified 1,273 articles. We excluded 1,184 articles after screening titles and abstracts, and had to exclude 24 more articles after reading the full text. Sixty-five reviews were eventually included in this study on quality assessment. Of these reviews, 53 contained a meta-analysis (see Figure 1).

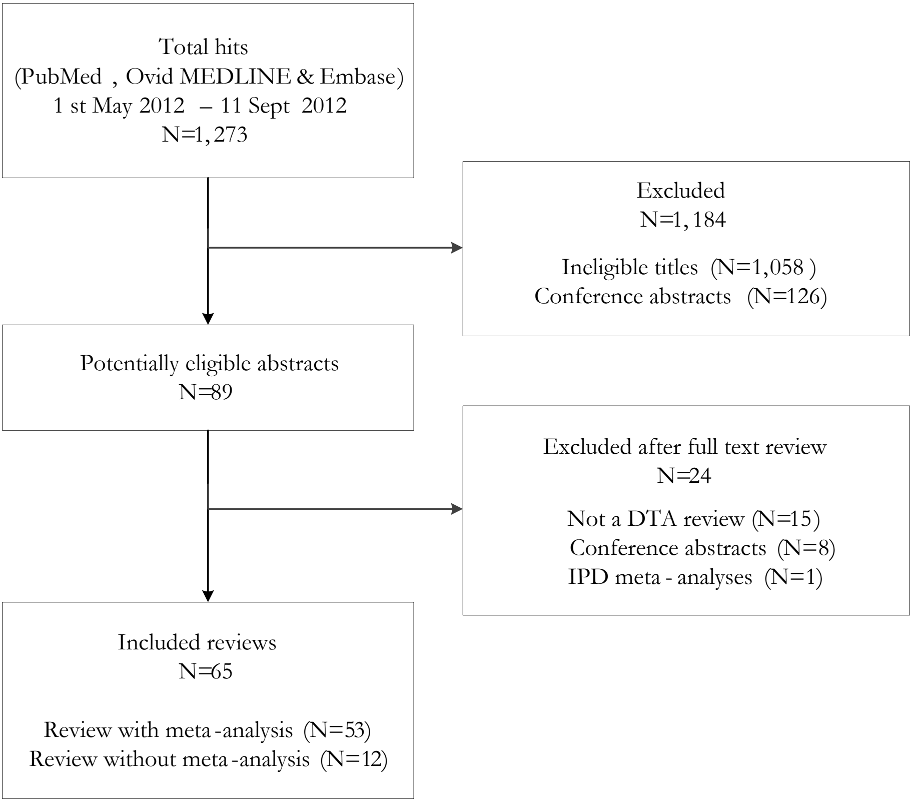

Total hits
(PubMed , Ovid MEDLINE & Embase)
1 st May 2012 – 11 Sept 2012
N=1,273
Excluded
N=1,184
Ineligible titles (N=1,058)
Conference abstracts (N=126)
Potentially eligible abstracts
N=89
Excluded after full text review
N=24
Not a DTA review (N=15)
Conference abstracts (N=8)
IPD meta - analyses (N=1)
Included reviews
N=65
Review with meta-analysis (N=53)
Review without meta-analysis (N=12)

Figure 1. *Legend: Flow Chart of Study Inclusion*

Characteristics of included reviews

Details of the study characteristics are outlined in Table 1. In summary, this sample of 65 reviews included one Cochrane review and 64 reviews published in other peer-reviewed journals. The median impact factor of the journals in which the included reviews were published in was 3.1 [Interquartile range, 2.4 to 4.1]. Of all the tests evaluated in the included reviews, imaging tests formed the largest group (n=36, 55%).

Instruments used to assess methodological quality

Of the included reviews, 60 (92%) had formally assessed the methodological quality of included studies. Most reviews had used QUADAS to assess the quality of included studies (n=44) and most presented their results as tables of individual quality items (n=31). Details of this assessment are outlined in Table 1.

Table 1. *Characteristics of included reviews*

Characteristic		Number (%) N=65
Number of primary studies in reviews, median [IQ range]		16 [10-24]
Journal Impact Factor, median [IQ range]		3.1 [2.4-4.1]
Type of test evaluated		
	Imaging test	36 (55)
	Laboratory test	17 (26)
	Other	12 (18)
Publication		
	Cochrane library	1 (1)
	Other peer reviewed journals	64 (99)
Quality assessment tools		
	No quality assessment	5 (8)
	QUADAS	44 (68)
	QUADAS-2	1 (1)
	STARD	3 (5)
	Both QUADAS and STARD	4 (6)
	Quality Assessment of Reliability Studies	1 (1)
	Other checklists of quality criteria	6 (9)
	Unclear	1 (1)
Presentation of quality results*		
	Table of individual quality items	31 (48)
	Summary score	18 (28)
	Summary graph	12 (18)
	Narrative explanation	7 (11)
	Other	5 (8)

** One review could have one or more ways of presenting results.*

Incorporation of assessments of quality in the review

a. Abstract

Table 2 summarizes the approaches used to mention quality in the abstract of the review with examples. Quality assessment was only mentioned in 28 abstracts (43%); a majority of these referred to it in the methods section (n=21). Only 5 reviews (20-24) linked results of quality assessment to accuracy estimates in the conclusion of the abstract. Three of these had not performed a meta-analysis (21-23) due to the poor quality of included studies.

b. Main Text

Table 3 summarizes, with examples, the approaches used to incorporate quality in the main text of the review. The detailed breakdown of how quality was incorporated in the analysis, discussion and eventually to the conclusions in the main text of the review is presented below.

Table 2. *Incorporation of quality assessment in abstracts of diagnostic reviews*

Approach	Overall quality of included studies	Number N=65	Example
Quality mentioned in abstract		28 (43%) [a]	
Quality in methods		21 (32%)	The quality of the studies was assessed using the guidelines published by the QUADAS (quality assessment for studies of diagnostic accuracy, maximum score 14) (25).
Quality in results		12 (19%)	"The sensitivity analysis of 10 high quality studies (a score of >=4) showed a pooled sensitivity of 94% and pooled specificity of 0.95" (26). "The quality of the included studies was poor to mediocre" (27).
Quality results considered in conclusion		5 (8%)	[œ]"The observed high sensitivity of the punch biopsy derived from all studies is probably the result of verification bias" (24). [ß]"The quality of the studies investigating these tests is too low to provide a conclusive recommendation for the clinician" (23).

[a] *Quality was mentioned in one or more sections in the abstract*

[œ] *Example of conclusion in a review with a meta-analysis*

[ß] *Example of conclusion in a review without a meta-analysis*

Table 3. *Incorporation of quality assessment in main text of diagnostic reviews*

Approach	Number N=65	Example
Quality mentioned in the main text	60 (92%) [b]	
Results of quality assessment reported, no mention in discussion or conclusion	13 (20%)	Results presented as table of individual QUADAS items. No further discussion or interpretation of results (28).
Results of quality assessment reported and discussed, but quality not linked to conclusion	41 (63%)	Assessed quality using criteria of internal and external validity. Overall quality clearly not stated. Discussion as limitation only: "Fourth, the variability in the quality of the primary studies may introduce important limitations for the interpretation of this review study" Conclusion: "Based on the results of this systematic review, F-18 FDG PET (PET/CT) was useful in ruling in extrahepatic metastases of HCC and valuable for ruling out the recurrent HCC" (29).
Results of quality assessment reported and discussed, and conclusions regarding test accuracy linked to results of quality assessment	6 (9%)	[œ] "In conclusion, the observed high sensitivity and low specificity of the colposcopy-directed punch biopsy for high grade CIN might be a result of verification bias. The sensitivity looks high but is probably a spurious finding caused by the fact that most studies restricted excision mainly to women with a positive punch biopsy" (24). [ß] " There exists a wide range of physical diagnostic tests for FAI and/or labral pathology and little information on the diagnostic accuracy and validity. The methodological quality of the diagnostic accuracy studies is moderate to poor " (23).
Results of quality assessment reported and discussed, and recommendations based on general unspecified quality items	12 (18%)	Assessed quality with Original QUADAS Only included high quality studies based on a summary score (>9/14) "In conclusion, T2WI combined with DWI is superior to T2WI alone in the detection of prostate cancer. High-quality prospective studies regarding the combination of T2WI plus DWI in detecting prostate carcinoma still need to be conducted" (30).

[b] *Quality was mentioned in one or more sections in the abstract*

[œ] *Example of conclusion in a review with a meta-analysis*

[ß] *Example of conclusion in a review without a meta-analysis*

Incorporation in the analysis

Twelve of the included reviews did not contain a meta-analysis. Four reviews (21-23;31) cited the poor quality of the identified studies as a reason for not conducting a meta-analysis, three (21-23) of which further factored the poor quality of studies in their conclusion. Other reasons for not conducting a meta-analysis were heterogeneity in test executions or study populations (n=5) and not meeting inclusion criteria (n=1); 2 reviews did not give an explanation.

Among the reviews with a meta-analysis (n=53), nineteen (36%) incorporated quality in the analysis. Quality was incorporated in the analysis using meta-regression (n=6), sensitivity analysis (n=4), subgroup analysis (n=2), both meta-regression and subgroup analysis (n=2) or through unspecified methods, (n=5). Eight found significant effects of quality on accuracy; in none of them these effects were factored in the conclusions.

Incorporation in the discussion

Thirteen reviews (20%) only presented results of quality assessment, without further discussion; most of these (n=12) contained a meta-analysis. In total, 47 reviews (72%) discussed the results of quality assessment but only 6 (9%) further linked these results to their conclusions.

Ten reviews without a meta-analysis discussed their results but only four (20-23) linked these results to their conclusions. Quality was discussed as a study limitation (n=7), as a strength of the review (n=2) and as potentially influencing the accuracy estimates (n=1).

For the reviews with a meta-analysis, the results of the quality assessment were discussed 35 times in the discussion section, and twice in the results section. In the discussion section, quality was discussed as a study limitation (n=21), as a strength of the study (n=7), as a summary of results of the analysis (n=11), and as potentially influencing the summary estimates of the review (n=4). Eight studies discussed quality in more than one way. In the results section, quality was discussed as potentially influencing the summary estimates of the review (n=1) and as strength of the review (n=1). Twenty of the reviews that did not incorporate quality in their analysis (n=30) discussed their results of quality assessment. They did so mostly as limitations in assessing the quality of included studies (n=14, 70%).

Incorporation in conclusions

In total, only six reviews (9%) incorporated the results of quality assessment in their conclusions in the main text of the review (20-24;32). Most of which (n=4) were reviews without a meta-analysis (20-23). Three reviews cited poor quality as a reason for not conducting a meta-analysis (21-23).

Of these six reviews that incorporated quality in the conclusions, three were published in a journal with an impact factor above the median impact factor (3.1) of the included reviews. In addition, two reviews were imaging studies and four reviews evaluated tests belonging to the category 'other'.

For the reviews with a meta-analysis, one acknowledged the limitations in assessing the quality of included studies (32), and one other considered the potential effect of the quality item 'verification bias' on the test's accuracy estimates (24). These reviews did not highlight the quality of included studies (high or low quality) in the main text and had not performed any statistical analysis to investigate the effect of quality differences on pooled estimates.

Of these two reviews, one also incorporated results of quality assessment in the conclusion in the abstract (24). The other review (32) encouraged authors in the conclusion of the main text to be cautious when interpreting the results of the review, because of the methodological limitations, but did not highlight this limitation in the conclusion of the abstract. An abstract that presents overly optimistic conclusions compared to the main text may lead to overinterpretation of the test's accuracy results (33).

Twelve reviews made recommendations about the test in the main text, based on general unspecified quality items not linked to the results of quality assessment, and using phrases such as 'high quality studies are needed' or 'large prospective studies are needed'. These were all reviews with a meta-analysis.

DISCUSSION

In a sample of 65 recently published diagnostic accuracy reviews of which 53 contained a meta-analysis, we found that almost all (92%) had assessed the methodological quality of included studies. Yet only 6 reviews (9%) considered results of quality assessment when drawing conclusions about the test's performance. Three of these had decided not to perform a meta-analysis because of limitations in quality of the available evidence.

Whiting and colleagues (34) have previously reviewed existing quality assessment tools for diagnostic accuracy studies, two years after the introduction of the original QUADAS tool. They examined to what extent quality had been assessed and incorporated in diagnostic systematic reviews. Just about half of the 114 systematic reviews examined had assessed the methodological quality of included studies; 91 different quality assessment tools were identified. In contrast, only 5 different quality assessment tools could be identified in our study, with QUADAS being used in about 8 in 10 reviews assessed. This reinforces the existing evidence on the rapid uptake of QUADAS (12,13).

Whiting and colleagues observed that 11 reviews (10%) used study quality as a basis for recommendations for future research. Yet it was unclear if these recommendations were based on the quality as documented in the reviews. Recommendations for future research can also be based on aspects not necessarily investigated in the review. Our study showed that twelve reviews made recommendations about the test based on general unspecified quality items not linked to the results of quality assessment, using rather general phrases, such as 'high quality studies are needed' or 'large prospective studies are needed'.

The specific reasons for not considering the assessments of quality of included studies in the overall findings of reviews are unclear. The absence of quality considerations could be partly explained by the parallel absence of clear recommendations on how to do so: guidance on how to incorporate quality into the conclusions of a review is scarce and vague.

Key guidance papers on reporting and evaluating systematic reviews, such as the Cochrane handbook (3;4;35), the statements on preferred reporting items for systematic reviews and meta-analyses (PRISMA) (36), on the assessment of multiple systematic reviews (AMSTAR) (37), and on the grading of recommendations assessment, development and evaluation (GRADE) (38;39) recommend that the methodological quality of included studies is discussed and factored into the overall findings of the review, but all of these fall more or less short on clearly explaining how to do so.

For instance, the Cochrane handbook for reviews of diagnostic accuracy studies (4;35) recommends that quality is assessed, included in the analysis, and used to generate recommendations for future research. It does not explicitly state how to discuss the results and incorporate the findings into the conclusions. The PRISMA guideline (36) is explicit in recommending that authors present the results of the risk of bias assessment and highlight, in the discussion section, any limitations encountered during risk of bias assessment. About the conclusion section, the recommendation in PRISMA is more vague; it advises authors to 'provide a general interpretation of the results in the context of other evidence, and implications for future research'. AMSTAR (37) is a scoring system for evaluating the quality of a systematic review, rather than that of the studies included in such a review. One item it recommends, as a measure of the quality of a review, is whether the review used the quality of included studies in formulating conclusions (Item 8). GRADE (38;39) provides a framework for making evidence based recommendations by rating the quality of the evidence and grading the strength of recommendations. In this process risk of bias assessment is a key component. The strength of GRADE lies in providing guidance on how to make recommendations; it does not stipulate how risk of bias assessment can be incorporated in evidence synthesis.

Another aspect to be held responsible for the absence of quality considerations in the conclusions of systematic reviews may be the multi-dimensional nature of evaluations of risk of bias. Since there are multiple quality or risk of bias items to consider, review authors may find it difficult to select the most important quality items to assess, analyse, discuss and draw conclusions from. Some authors use a summary score, a quantitative estimate of quality items evaluated. However, the use of such simple summary scores is discouraged because they fail to consider differences in importance of quality items (40;41).

Poor reporting of relevant items in primary diagnostic accuracy studies, as stipulated by the Standards for Reporting of Diagnostic Accuracy initiative (STARD) (7), limits the assessments of quality of these studies. Authors may find it challenging to draw conclusions about the quality of the included studies and their impact on the test accuracy estimates when their assessments of quality or risk of bias are unclear. Many authors of reviews in our study discussed the challenges in assessing the quality of included studies as a review limitation.

Our study has one main limitation. Given that QUADAS-2 was recently introduced - just one year before the time of our search - and that uptake of novel methods takes time, we did not expect to find many articles utilizing the new version. This limited our evaluation of how results using QUADAS-2 are incorporated into the conclusions. Nonetheless, we anticipate that drawing conclusions from the multiple domains of risk of bias recommended by QUADAS-2 will still be challenging.

Although most reviews in our study did not consider quality in drawing conclusions, the ones that did show that it is possible to consider the strength of the evidence in making statements about a test's performance based on a systematic review of test accuracy studies. If there is no quality evidence, one can refrain from meta-analysis, and make no firm statements about test performance. Alternatively, one can explicitly qualify the results from a meta-analysis of poor quality studies as evidence with limited credibility. If there are studies with and studies without deficiencies, one can limit the analysis to high quality studies, and add explicit statements to that extent to the conclusions. If there are studies with high risk of bias and studies at low risk, one can explore the effects of this variability on the summary estimates. If there are systematic effects, one could and should factor this finding into the conclusions. The dominant practice seems the worst possible scenario: to evaluate the quality of included studies without considering the findings from that exercise in drawing conclusions.

Guidance is needed in assisting authors to incorporate results of quality assessment in the conclusions. Such guidance should come from concerted actions of methodologists. It could be presented in the form of simple and practical online tutorials or tutorials published in scientific journals. Such tutorials could guide authors with examples on how to draw conclusions, especially in light of challenges such as the multiple domains of risk of bias recommended by QUADAS-2, when quality of included studies has no statistical effect on the pooled accuracy estimates, or when the risk of bias assessment is hampered by poor reporting of included studies, or when poor quality of studies precludes a meta-analysis.

CONCLUSION

We found it disturbing that quality of the included evidence was evaluated in almost all diagnostic reviews, but that almost no authors had incorporated the results of quality assessment in the conclusions of their reviews. The practice of reporting systematic reviews of test accuracy should improve if readers not only want to be informed about the limitations in the available evidence, but also on the associated implications for the performance of the evaluated tests in clinical practice. Reviewers and readers of test accuracy reviews need to check that the results or limitations of quality assessment are incorporated in the abstract and conclusion of the review. Simply relying on the review results, without considering the quality of the underlying research, could lead to the uptake of poorly performing tests in practice and, consequently, to suboptimal patient management.

COMPETING INTERESTS

No funding was received for this project.
JR and PB were involved in the development of both the original and revised QUADAS tool.

KGM and ML were involved in the development of the revised QUADAS tool.

ACKNOWLEDGEMENTS

We thank René Spijker, MSc (Dutch Cochrane Centre, University of Amsterdam) for assisting in the development of the search strategy of this project.

REFERENCE LIST

1. Methods Guide for Medical Test Reviews. AHRQ Publication No. 12-EC017. Rockville, MD: Agency for Healthcare Research and Quality. 2012 Jun.
2. Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM. Systematic reviews of diagnostic test accuracy. Ann Intern Med 2008 Dec 16;149(12):889-97.
3. Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. 2011.
4. Reitsma J, Rutjes A, Whiting P, Vlassov VV, Leeflang M, Deeks J. Chapter9: Assessing methodological quality, Cochrane Handbook for Systematic Reviews of Diagnostic Test Acuracy Version 1.0.0. Deeks J, Bossuyt P, Gatsonis C, editors. 2009. The Cochrane Collaboration.
5. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med 2004 Feb 3;140(3):189-202.
6. Whiting PF, Rutjes AW, Westwood ME, Mallett S. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. J Clin Epidemiol 2013 Oct;66(10):1093-104.
7. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. Ann Intern Med 2003 Jan 7;138(1):40-4.
8. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med 2003 Jan 7;138(1):W1-12.
9. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999 Sep 15;282(11):1061-6.
10. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol 2003 Nov 10;3:25.
11. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011 Oct 18;155(8):529-36.
12. Willis BH, Quigley M. Uptake of newer methodological developments and the deployment of meta-analysis in diagnostic test research: a systematic

review. BMC Med Res Methodol 2011;11:27.

13. Dahabreh IJ, Chung M, Kitsios GD, Terasawa T, Raman G, Tatsioni A, et al. Comprehensive Overview of Methods and Reporting of Meta-Analyses of Test Accuracy. Methods Research Report. (Prepared by the Tufts Evidence-based Practice Center under Contract No. 290-2007-10055-I.) . AHRQ, Rockville, MD: Agency for Healthcare Research and Quality ; 2012 Mar. Report No.: 12.
14. Moja LP, Telaro E, D'Amico R, Moschetti I, Coe L, Liberati A. Assessment of methodological quality of primary studies by systematic reviews: results of the metaquality cross sectional study. BMJ 2005 May 7;330(7499):1053.
15. Hopewell S, Boutron I, Altman DG, Ravaud P. Incorporation of assessments of risk of bias of primary studies in systematic reviews of randomised trials: a cross-sectional study. BMJ Open 2013;3(8):e003342.
16. Zhelev Z, Garside R, Hyde C. A qualitative study into the difficulties experienced by healthcare decision makers when reading a Cochrane diagnostic test accuracy review. Syst Rev 2013;2:32.
17. Beller EM, Glasziou PP, Altman DG, Hopewell S, Bastian H, Chalmers I, et al. PRISMA for abstracts: reporting systematic reviews in journal and conference abstracts. PLoS Med 2013;10(4):e1001419.
18. Boutron I, Dutton S, Ravaud P, Altman DG. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. JAMA 2010 May 26;303(20):2058-64.
19. Yavchitz A, Boutron I, Bafeta A, Marroun I, Charles P, Mantz J, et al. Misrepresentation of randomized controlled trials in press releases and news coverage: a cohort study. PLoS Med 2012;9(9):e1001308.
20. Atluri S, Singh V, Datta S, Geffert S, Sehgal N, Falco FJ. Diagnostic accuracy of thoracic facet joint nerve blocks: an update of the assessment of evidence. Pain Physician 2012 Jul;15(4):E483-E496.
21. Cook C, Mabry L, Reiman MP, Hegedus EJ. Best tests/clinical findings for screening and diagnosis of patellofemoral pain syndrome: a systematic review. Physiotherapy 2012 Jun;98(2):93-100.
22. Mejare IA, Axelsson S, Davidson T, Frisk F, Hakeberg M, Kvist T, et al. Diagnosis of the condition of the dental pulp: a systematic review. Int Endod J 2012 Jul;45(7):597-613.
23. Tijssen M, van CR, Willemsen L, de VE. Diagnostics of femoroacetabular impingement and labral pathology of the hip: a systematic review of the accuracy and validity of physical tests. Arthroscopy 2012 Jun;28(6):860-71.
24. Underwood M, Arbyn M, Parry-Smith W, De Bellis-Ayres S, Todd R,

Redman CW, et al. Accuracy of colposcopy-directed punch biopsies: a systematic review and meta-analysis. BJOG 2012 Oct;119(11):1293-301.

25. Chang K, Lu W, Wang J, Zhang K, Jia S, Li F, et al. Rapid and effective diagnosis of tuberculosis and rifampicin resistance with Xpert MTB/RIF assay: a meta-analysis. J Infect 2012 Jun;64(6):580-8.

26. Chen J, Yang R, Lu Y, Xia Y, Zhou H. Diagnostic accuracy of endoscopic ultrasound-guided fine-needle aspiration for solid pancreatic lesion: a systematic review. J Cancer Res Clin Oncol 2012 Sep;138(9):1433-41.

27. van Teeffelen AS, Van Der Heijden J, Oei SG, Porath MM, Willekes C, Opmeer B, et al. Accuracy of imaging parameters in the prediction of lethal pulmonary hypoplasia secondary to mid-trimester prelabor rupture of fetal membranes: a systematic review and meta-analysis. Ultrasound Obstet Gynecol 2012 May;39(5):495-9.

28. Wu L, Dai ZY, Qian YH, Shi Y, Liu FJ, Yang C. Diagnostic value of serum human epididymis protein 4 (HE4) in ovarian carcinoma: a systematic review and meta-analysis. Int J Gynecol Cancer 2012 Sep;22(7):1106-12.

29. Lin CY, Chen JH, Liang JA, Lin CC, Jeng LB, Kao CH. 18F-FDG PET or PET/CT for detecting extrahepatic metastases or recurrent hepatocellular carcinoma: a systematic review and meta-analysis. Eur J Radiol 2012 Sep;81(9):2417-22.

30. Wu LM, Xu JR, Ye YQ, Lu Q, Hu JN. The clinical value of diffusion-weighted imaging in combination with T2-weighted imaging in diagnosing prostate carcinoma: a systematic review and meta-analysis. AJR Am J Roentgenol 2012 Jul;199(1):103-10.

31. Quatman CE, Quatman-Yates CC, Schmitt LC, Paterno MV. The clinical utility and diagnostic performance of MRI for identification and classification of knee osteochondritis dissecans. J Bone Joint Surg Am 2012 Jun 6;94(11):1036-44.

32. Smith TO, Drew B, Toms AP, Jerosch-Herold C, Chojnowski AJ. Diagnostic accuracy of magnetic resonance imaging and magnetic resonance arthrography for triangular fibrocartilaginous complex injury: a systematic review and meta-analysis. J Bone Joint Surg Am 2012 May 2;94(9):824-32.

33. Ochodo EA, de Haan MC, Reitsma JB, Hooft L, Bossuyt PM, Leeflang MM. Overinterpretation and misreporting of diagnostic accuracy studies: evidence of "spin". Radiology 2013 May;267(2):581-8.

34. Whiting P, Rutjes AW, Dinnes J, Reitsma JB, Bossuyt PM, Kleijnen J. A systematic review finds that diagnostic reviews fail to incorporate quality despite available tools. J Clin Epidemiol 2005 Jan;58(1):1-12.

35. Macaskill P, Gatsonis C, Deeks JJ, Harbord R, Takwoingi Y. Chapter 10: Analysing and Presenting results. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0. The Cochrane Collaboration. In: Deeks JJ, Bossuyt PM, Gatsonis C, editors. 1 ed. The Cochrane Collaboration; 2010.

36. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097.

37. Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol 2007;7:10.

38. Brozek JL, Akl EA, Jaeschke R, Lang DM, Bossuyt P, Glasziou P, et al. Grading quality of evidence and strength of recommendations in clinical practice guidelines: Part 2 of 3. The GRADE approach to grading quality of evidence about diagnostic tests and strategies. Allergy 2009 Aug;64(8):1109-16.

39. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008 Apr 26;336(7650):924-6.

40. Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA 1999 Sep 15;282(11):1054-60.

41. Whiting P, Harbord R, Kleijnen J. No role for quality scores in systematic reviews of diagnostic accuracy studies. BMC Med Res Methodol 2005;5:19.

CHAPTER 5

Reporting quality of diagnostic accuracy studies: a systematic review and meta-analysis of investigations on adherence to STARD

Daniël A. Korevaar
W. Annefloor van Enst
Rene Spijker
Patrick M.M. Bossuyt
Lotty Hooft

Journal of Evidence Based Medicine
2014;19(2):47-54

ABSTRACT

Background Poor reporting of diagnostic accuracy studies impedes an objective appraisal of the clinical performance of diagnostic tests. The Standards for Reporting of Diagnostic Accuracy Studies (STARD) statement, first published in 2003, aims to improve the reporting quality of such studies.
The objective was to investigate to which extent published diagnostic accuracy studies adhere to the 25-item STARD checklist, whether the reporting quality has improved after STARD's launch and whether there are any factors associated with adherence.

Methods We performed a systematic review and searched MEDLINE, EMBASE and the Methodology Register of the Cochrane Library for studies that primarily aimed to examine the reporting quality of articles on diagnostic accuracy studies in humans by evaluating adherence to STARD. Study selection was performed in duplicate; data were extracted by one author and verified by the second author.

Results We included 16 studies, analysing 1496 articles in total. Three studies investigated adherence in a general sample of diagnostic accuracy studies; the others did so in a specific field of research. The overall mean number of items reported varied from 9.1 to 14.3 between 13 evaluations that evaluated all 25 STARD items. Six studies quantitatively compared post-STARD with pre-STARD articles. Combining these results in a random-effects meta-analysis revealed a modest but significant increase in adherence after STARD's introduction (mean difference 1.41 items (95% CI 0.65 to 2.18)).

Discussion/Conclusion The reporting quality of diagnostic accuracy studies was consistently moderate, at least through halfway the 2000s. Our results suggest a small improvement in the years after the introduction of STARD. Adherence to STARD should be further promoted among researchers, editors and peer reviewers.

INTRODUCTION

In 2003, the Standards for Reporting of Diagnostic Accuracy Studies (STARD) statement was published in 13 biomedical journals (1;2). Diagnostic accuracy studies provide estimates of a test's ability to discriminate between patients with and without a predefined condition, by comparing the test results against a clinical reference standard. The STARD initiative was developed in response to accumulating evidence of poor methodological quality and poor reporting among test accuracy studies in the prior years (3;4). The STARD checklist contains 25 items which invite authors and reviewers to verify that critical information about the study is included in the study report. In addition, a flowchart that specifies the number of included and excluded patients and characterises the flow of participants through the study is strongly recommended. Since its launch, the STARD checklist has been adopted by over 200 biomedical journals (http://www.stard-statement.org/).

Over the past 20 years, reporting guidelines have been developed and evaluated in many different fields of research. Although a modest increase in reporting quality is sometimes noticed in the years following the introduction of such guidelines (5;6), improvements in adherence tend to be slow (7). This makes it difficult to make statements about the impact of such guidelines. For STARD, there has been some controversy around its effect (8). While one study noticed a small increase in reporting quality of diagnostic accuracy studies shortly after the introduction of STARD (9), another study could not confirm this (10).

Systematic reviews can provide more precise and more generalizable estimates of effect. A recently published systematic review evaluated adherence to several reporting guidelines in different fields of research, but STARD was not among the evaluated guidelines (11). To fill this gap, we systematically reviewed all the studies that aimed to investigate diagnostic accuracy studies' adherence to the STARD checklist in any research field. Our main objective was to find out how diagnostic accuracy studies adhere to (specific items on) the STARD checklist. Our research questions were: (1) How is the current (or rather, most recent) quality of reporting of diagnostic accuracy studies? (2) Has the quality of reporting improved after the introduction of STARD? (3) How do diagnostic accuracy studies score on specific items on the checklist? (4) Are there any factors associated with adherence to the checklist?

METHODS

Search and selection

The original protocol of this study can be obtained from the corresponding author. We performed a systematic review and searched MEDLINE and EMBASE, which, to our knowledge, provide the best sources for methodological reviews. To make sure that all relevant data were captured, we also searched the Methodology Register of the Cochrane Library, of which the content is sourced from MEDLINE and additional manual searches. We included studies that primarily aimed to examine the quality of reporting of articles of diagnostic accuracy studies in humans in any field of research, by evaluating their adherence to the STARD statement. Details on the search strategies are provided in Web only file 1. The final search was performed on 13 August 2013. The searches were performed without any restrictions for language, year of publication or study type. We excluded systematic reviews on the accuracy of a single test that had used the STARD checklist to score the quality of reporting in the included articles, as well as studies that investigated the influence of reporting quality on pooled estimates of test accuracy results. Such articles would be on a too specific topic to be able to make statements on the reporting quality of diagnostic accuracy studies in general. Studies focusing on reports about analytical rather than clinical performance were also excluded. Although the design of these two types of studies show many similarities, STARD was not designed for studies on analytical test performance and several items on the lists do not apply in this setting. We also excluded studies that evaluated less than 10 STARD items and studies that had not presented their results quantitatively (as a mean number of reported items or a score per individual item) because this would make an objective comparison between studies impossible.

Two authors (DK and WvE) independently screened the titles and abstracts of the search results to identify potentially eligible studies. If at least one author identified an abstract as potentially eligible, the full text of the article was assessed by both authors. Disagreements were resolved through discussion, whenever possible. If agreement could not be reached, the case was discussed with a third author (LH). One author (DK) also reviewed reference lists of included studies for additional relevant papers.

Data collection

An extraction form was created before the literature search was performed, and piloted on three known eligible studies. After the pilot, the form was slightly modified. One author (DK) extracted relevant data from the included studies, which were verified by the second author (WvE). Disagreements were resolved through discussion. If necessary, a third author (LH) made the final decision.

Of each included article, the first author, country, year of publication and journal were extracted. We also identified the inclusion and exclusion criteria, research field, primary aims, the number of studies included, which STARD items were evaluated and how they had been scored. In addition, we retrieved (descriptive) statistics regarding overall and item-specific STARD adherence, and adherence comparisons between articles published post-STARD versus those published pre-STARD. Any additional study characteristics mentioned to be associated with STARD adherence were extracted. We also extracted any statistics on inter-rater agreement in evaluating STARD items, and conclusions, interpretation and recommendations of the authors.

We assessed the quality of included studies by using the 11-item AMSTAR (Assessment of Multiple Systematic Reviews) tool (12). As several items on this list do not apply to the studies included in our review, we omitted four items and only assessed item 1 (was an 'a priori' design provided?), item 2 (was there duplicate study selection and data extraction?), item 3 (was a comprehensive literature search performed?), item 4 (were inclusion and exclusion criteria provided?), item 5 (was a list of included and excluded studies provided?), item 6 (were the characteristics of included studies provided?) and item 9 (was the conflict of interest included?).

Analysis: overall adherence to STARD

We calculated k statistics to assess inter-reviewer agreement for the two phases of study selection. For each included study, we calculated the overall STARD score, defined as the mean number of items reported by articles included in that study, and the proportion of articles adhering to each specific STARD item. For each STARD item, we calculated the median and range of these proportions.

Some studies also counted how often an item was partially reported. To be able to make comparisons between studies, we counted partially reported items as half in calculating proportions. Some STARD items pertain to the index test and the reference standard. Whenever these were analysed separately, half a point was allocated per reported item. If a study reported that an item on the STARD checklist was not applicable to all evaluated articles, that study was not included in our overall analysis for that specific item. If a study reported that a STARD item was applied to less than 100% of the evaluated articles, the score was calculated for the number of articles for which the item applied and the calculated proportions were adjusted.

Analysis: adherence to STARD before and after its launch

To obtain a summary estimate and the corresponding 95% CI of the difference in adherence before and after its launch, we used inverse variance random-effects meta-analysis (13). Only studies specifically reporting pre-STARD and post-STARD results were included in this analysis. We explored statistical heterogeneity using the I^2 test (14). We performed a subgroup analysis by separately analysing studies examining a general sample of diagnostic accuracy studies, rather than those investigating adherence in a specific field of research.

One included study only reported standard deviations (SDs) for (equally sized) subgroups of STARD-adopting and non-adopting journals (10). We calculated their overall SD by taking the square root of the pooled variances. SDs of one other study were obtained after contacting the authors (15).

We used inverse variance random-effects meta-analysis to calculate summary ORs and 95% CIs for item-specific adherence in the pre-STARD versus post-STARD groups. Only studies specifically reporting the proportion of evaluated articles adhering to each individual item for both the pre-STARD and post-STARD groups were included in this analysis.

RESULTS

Search results and characteristics of included studies

Five hundred and eighteen studies were identified through the search, of which 35 were deemed potentially eligible after screening titles and abstracts (Figure 1). After studying the full texts, we were able to include 16 studies (9;10;15-28). Reasons for exclusion of potentially eligible studies are provided in Figure 1. No additional studies were identified through reference lists. Inter-reviewer agreement was substantial for the screening of titles and abstracts (=0.77 (95% CI 0.66 to 0.88)), and was perfect for the subsequent assessment of full-texts (k=1.0).

The characteristics of the included studies are provided in Table 1. Three studies investigated adherence to STARD in a general sample of articles on diagnostic accuracy studies, the other 13 had performed so in a specific field of research. None of the included studies had evaluated a recent sample of articles: one study evaluated articles published through 2010, one study through 2008, two studies through 2007, and four studies through 2006. All other studies included only articles published before 2006. Twelve studies included articles published before and after STARD's launch. One study investigated only articles published pre-STARD and three studies investigated only articles published post-STARD.

The number of evaluated articles varied markedly between the included studies, with a median of 55 (range: 16-300). Most of the studies (n=13) evaluated all 25 STARD items. However, among three of these, one item was found not applicable to all included articles. The other three studies had evaluated 24, 22 and 13 items of the 25 items, respectively. Kappa-values for overall inter-rater agreement on the STARD-items were reported by nine studies: moderate agreement (k=0.41-0.6) was achieved in one study, substantial agreement (k=0.61-0.8) in six studies, and almost perfect agreement (k=0.81-1.0) in two other studies (29). An overall percentage agreement was reported by seven studies; this varied between 81% and 95%. Four studies did not report on inter-rater agreement.

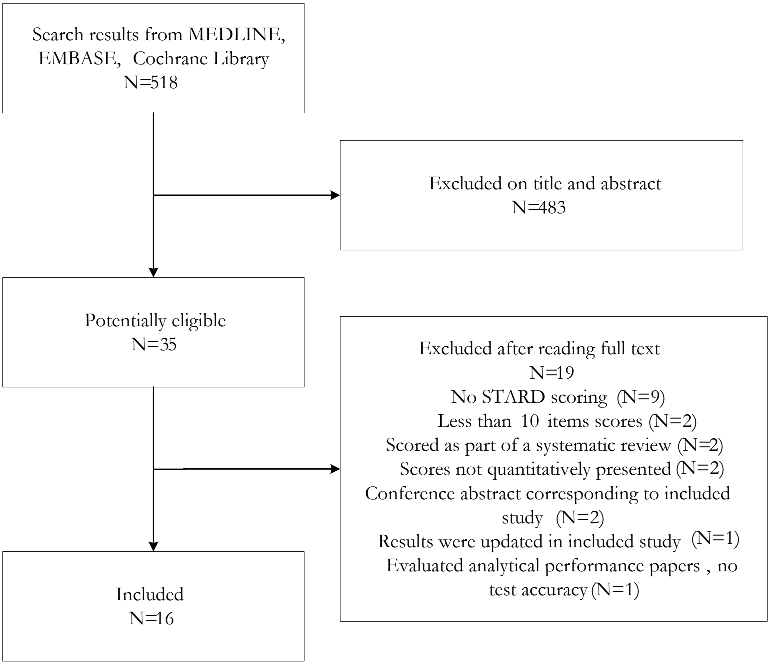

Search results from MEDLINE, EMBASE, Cochrane Library
N=518
Excluded on title and abstract
N=483
Potentially eligible
N=35
Excluded after reading full text
N=19
No STARD scoring (N=9)
Less than 10 items scores (N=2)
Scored as part of a systematic review (N=2)
Scores not quantitatively presented (N=2)
Conference abstract corresponding to included study (N=2)
Results were updated in included study (N=1)
Evaluated analytical performance papers , no test accuracy (N=1)
Included
N=16

Figure 1. *Flowchart for selection of studies.*

An a priori study design was provided by only one included study. Seven studies performed the complete study selection in duplicate, while three did so in part. Eleven studies evaluated the reporting quality of all the included studies in duplicate, and three did so for a part of the included studies. All the included studies provided comprehensive data on the literature searches and the inclusion and exclusion criteria. Although more than half (n=9) of the studies provided a list of included studies, only two provided a list of excluded studies. Characteristics of included studies were provided, to some extent, by all studies: all gave information on the research field in which included articles were performed and 12 studies gave information on the type of tests used. Only three studies gave information on the included studies' design.

Overall adherence to STARD

The overall mean STARD score varied from 9.1 to 14.3 for the 13 studies that had evaluated all 25 STARD items, with a median of 12.8 items (Table 1). Fifteen (94%) of the included studies concluded that the adherence to STARD was poor, medium, suboptimal or needed improvement. One study used more conservative language and concluded that adherence of included articles was highly variable. Seven studies evaluating all 25 items only reported post-STARD results or reported pre-STARD and post-STARD results separately. The overall mean number of items reported in these post-STARD results varied from 12.0 to 15.5, with a median of 13.6. Most of the included studies recommended the use of STARD as a guideline to improve the quality of reporting of diagnostic accuracy studies, and no study discouraged it.

The medians and ranges of the proportions of adherence to individual STARD-items reported by included studies are provided in Table 2. There was a large between-study variation in adherence to specific items. Overall, only 12 items had a median proportion exceeding 50%; only three items had a median proportion above 75%. When only evaluating post-STARD results, these median proportions were slightly better: 15 items exceeding 50% and 6 items exceeding 75%. Six items (8, 9, 10, 11, 13 and 24) concern the index test as well as the reference standard. Reporting of the index test was better than reporting of the reference standard for all of these items.

Several studies reported on factors potentially associated with quality of reporting. One study found that adherence to STARD was significantly better for cohort studies compared with case-control studies (9), but another study could not confirm this (24). Other factors reported to be significantly associated with higher STARD scores were sample size (higher scores among larger studies (15)) and research field (obstetric studies scored better than gynaecological studies (15), and tuberculosis and malaria studies scored better than HIV studies (18)). Factors that did not show a significant difference were geographical area (15), level of evidence (24) and pooled sensitivity and specificity (28), but these findings were not replicated in a subsequent study.

Table 1. *Characteristics of included studies*
**One of the 25 evaluated STARD-items was not applicable to all the articles included in this study*

1st author	Country	Year	Journal	Research field	# of articles included
Areia (16)	Portugal	2010	Endoscopy	Endoscopy.	110
Coppus(17)	The Netherlands	2006	Fertility and Sterility	Reproductive medicine.	51
Fontela(18)	Canada	2009	PlosONE	Commercial tests for tuberculosis, HIV, malaria.	90
Freeman(19)	U.K.	2009	European Journal of Obstetrics & Gynecology and Reproductive Biology	Non-invase prenatal diagnostic tests for Rhesus D genotyping.	27
Gómez Sáez(20)	Spain	2009	Medicina Clinica	Any research field, 4 Spanish journals.	58
Johnson(21)	U.K.	2007	Ophthalmology	Optical coherence tomography (OCT) in glaucoma.	30
Lumbreras(22)	Spain	2006	Gaseta Sanitària	Genetic-molecular research.	44
Paranjothy(23)	U.K.	2007	Journal of Glaucoma	Scanning laser polarimetry (SLP) for diagnosing glaucoma.	20
Rama(24)	U.K.	2006	Clinical Orthopaedics and Related Research	Orthopedics.	37
Selman(15)	U.K.	2011	BMC Women's Health	Obstetrics and gynaecology.	300
Shunmugam(25)	U.K.	2006	Investigative Ophthalmology & Visual Science	Heidelberg retina tomography (HRT) for glaucoma detection.	29
Siddiqui(26)	U.K.	2005	British Journal of Ophthalmology	Ophthalmology.	16
Smidt(9)	The Netherlands	2006	Neurology	Six general and 6 disease/ discipline-specific journals.	265
Wilczynski(10)	Canada	2008	Radiology	Twelve journals on radiology, internal medicine or general medicine.	240
Zafar(27)	U.K.	2008	Clinical and Experimental Ophthalmology	Diabetic retinopathy (DR) screening.	76
Zintzaras(28)	Greece	2012	BMC Musculoskeletal Disorders	Anti-CCP2 for the diagnosis of rheumatoid arthritis.	103

Time frame	# of STARD items evaluated	Mean STARD score (% of items evaluated)	Authors' conclusions on quality of reporting
1998-2008	25	12.9 (52%)	"Recent publications in diagnostic endoscopy achieve only medium quality."
1999 vs 2004	25	12.3 (49%)	"The quality of reporting in articles on test accuracy in reproductive medicine is poor to mediocre."
2004-2006	25	13.6 (54%)	"Diagnostic studies on TB, malaria and HIV commercial tests [...] were often poorly reported."
1996-2006	25	9.1 (36%)	"Articles have consistent weaknesses in their reporting."
2004-2007	25	12.0 (48%)	"Despite efforts by different groups of research to achieve higher methodological quality in the diagnostics field, on average, they follow less than half of the items proposed by STARD."
2001-2006	25*	13.2 (53%)	"Quality of reporting of the diagnostic accuracy of OCT in glaucoma is suboptimal."
2002-2005	24	9.8 (41%)	"The articles on genetic-molecular diagnostic tests [...] fail to satisfy most of the quality requirements assembled in the STARD proposal."
1997-2000 vs 2004-2005	25*	13.5 (54%)	"The quality of reporting of diagnostic accuracy tests for glaucoma with SLP is suboptimal."
2002-2004	25	14.2 (57%)	"Current standards of reporting of diagnostic accuracy studies in orthopaedic journals are suboptimal."
1977-2007	25	12.5 (50%)	"The reporting of included studies in this review overall was poor."
1995-2004	25*	14.3 (57%)	"The quality of reporting of diagnostic accuracy tests for glaucoma with HRT is suboptimal."
2002	25	11.6 (47%)	"The current standards of reporting of diagnostic accuracy tests are highly variable."
2000 vs 2004	25	12.8 (51%)	"After publication of STARD, the quality of reporting of diagnostic accuracy studies has slightly improved. There is still room for improvement."
2001-2002 vs 2004-2005	13	8.2 (63%)	"We found low rates of adherence to the STARD checklist items."
1995-2006	25	9.9 (40%)	"The quality of diagnostic accuracy reports in DR screening is suboptimal."
2003-2010	22	14.0 (64%)	"The overall reporting quality was relatively good but needs further improvement."

Table 2. *Proportions of adherence to individual STARD items.*

STARD item	Overall		
	Studies evaluating item n	Median of proportions %	Range %
25. Clinical applicability of findings	14	98%	41-100%
4. Participant recruitment	16	85%	55-100%
2. Research questions/aims	14	84%	24-100%
8. Technique of:	16	73%	31-98%
8a. Index test	5	92%	49-95%
8b. Reference standard	5	63%	13-86%
15. Characteristics of study population	16	73%	42-90%
7. Reference standard and rationale	16	70%	28-98%
9. Units/cut-offs/categories for:	16	70%	0-98%
9a. Index test	5	84%	68-95%
9b. Reference standard	5	73%	55-76%
3. Study population	16	68%	23-92%
6. Data collection	16	68%	21-100%
19. Cross tabulation of results	15	65%	2-99%
18. Distribution of severity of disease	16	62%	0-97%
21. Estimates of diagnostic accuracy	15	56%	12-97%
12. Methods for statistics used	15	49%	8-90%
14. Dates of study	16	47%	6-73%
1. Study identified as test accuracy study	13	40%	8-100%
5. Participant sampling	16	40%	12-89%
23. Estimates of variability of accuracy	15	37%	0-100%
17. Time interval between tests	15	34%	0-77%
11. Blinding of results of:	16	29%	14-54%
11a. Index test	5	43%	33-72%
11b. Reference test	5	23%	12-48%
22. How uninterpretable results were handled	15	28%	8-62%
10. Persons exectuting:	16	26%	2-73%
10a. Index test	5	33%	7-46%
10b. Reference standard	5	20%	0-35%
16. Eligible patients not undergoing either test	16	24%	5-78%
16a. Flow diagram	12	5%	0-16%
13. Methods for test reproducibility for:	15	16%	0-88%
13a. Index test	4	20%	12-53%
13b. Reference standard	4	7%	0-12%
24. Estimates of test reproducibility, for:	15	8%	0-96%
24a. Index test	4	20%	13-38%
24b. Reference standard	4	3%	0-8%
20. Adverse events	12	7%	0-33%

Post-STARD results only		
Studies evaluating item n	Median of proportions %	Range %
5	98%	84-99%
7	93%	60-98%
5	88%	76-96%
7	74%	40-97%
4	84%	58-97%
4	55%	23-72%
7	70%	60-93%
7	76%	45-98%
7	83%	63-85%
4	91%	71-94%
4	75%	56-80%
7	63%	21-88%
7	83%	43-95%
6	66%	28-99%
7	52%	11-98%
6	56%	22-97%
6	49%	11-90%
7	73%	42-81%
5	24%	18-99%
7	64%	31-89%
6	39%	0-100%
6	38%	25-74%
7	33%	16-55%
4	50%	26-67%
4	25%	15-48%
6	25%	8-57%
7	20%	2-42%
4	26%	4-51%
4	14%	0-33%
7	53%	13-70%
4	8%	0-22%
6	18%	0-88%
3	35%	6-48%
3	4%	0-6%
6	8%	0-96%
3	22%	6-44%
3	0%	0-6%
6	11%	1-18%

Adherence to STARD before and after its launch

Of the 12 studies that had included articles published before and after the publication of STARD, 6 reported results for the pre-STARD and post-STARD groups. Combining these studies in a meta-analysis showed that significantly more items were reported post-STARD, with an estimate difference of 1.41 items (95% CI 0.65 to 2.18) (Figure 2). However, the great majority of the 383 post-STARD articles included in this analysis were published in the 2 years after introduction of STARD (2004 and 2005, n=349); only 34 articles were published after 2005. As expected, I^2 test showed evidence of substantial statistical heterogeneity (66%). Subgroup analysis of the two studies that reported on a general sample of diagnostic accuracy studies(9;10) showed a non-significant increase in the number of reported STARD-items (difference of 1.02 items (95% CI -0.08 to 2.12), I^2 =80%).

	Post-STARD			Pre-STARD				Mean Difference	
Study or Subgroup	Mean	SD	Total	Mean	SD	Total	Weight	IV, Random, 95% CI	Year
Smidt 2006	13.55	3.2	141	11.94	3.3	124	19.4%	1.61 [0.82, 2.40]	2006
Rama 2006	15.5	3.3	19	15.1	3.3	11	7.0%	0.40 [−2.05, 2.85]	2006
Coppus 2006	12.41	3.2	27	12.08	3.3	24	10.5%	0.33 [−1.46, 2.12]	2006
Wilczynski 2008*	8.43	2.22	120	7.94	2.55	120	21.1%	0.49 [−0.11, 1.09]	2008
Areia 2010	14.3	2.9	24	11.6	3.6	86	13.5%	2.70 [1.31, 4.09]	2010
Selman 2011**	12.32	2.92	17	11.05	2.57	88	12.7%	1.27 [−0.22, 2.76]	2011
Selman 2011***	15.33	3.04	35	12.75	3.54	160	15.8%	2.58 [1.43, 3.73]	2011
Total (95% CI)			**383**			**613**	**100.0%**	**1.41 [0.65, 2.18]**	

Heterogeneity: $Tau^2 = 0.63$; $Chi^2 = 17.70$, df = 6 (P = 0.007); $I^2 = 66\%$
Test for overall effect: Z = 3.61 (P = 0.0003)

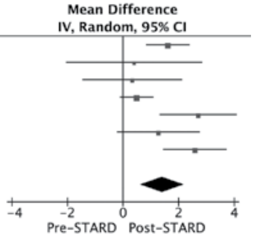


Figure 2. *Forest plot for studies included in meta-analysis comparing adherence post-STARD and pre-STARD. Llegend: *Wilczynski et al evaluated only 13 STARD items; the other studies evaluated 25 STARD items; **Results of the studies on obstetrics; ***Results of the studies on gynaecology.*

Six other studies have reported some form of analysis of STARD adherence over time. One of these noticed an upward trend in the number of items reported pre-STARD and post-STARD (23). Four others could not confirm this: two studies reported that introduction of STARD did not seem to have improved the quality of reporting of articles included in their analysis (21;22), one study observed no improvement of quality of reporting over time (27) and one study noticed a (non-significant) decline in adherence after STARD publication (20).

The pre-STARD versus post-STARD meta-analyses for individual items are reported in Web only file 2. Six items were significantly more reported after the publication of STARD: item 4 (describes participant recruitment), item 5 (describes participant sampling), item 6 (describes data collection), item 14 (reports dates of study), item 15 (reports characteristics of study population) and item 23 (reports estimates of variability of accuracy). Although still rare, the number of studies reporting a flow diagram also increased significantly. None of the STARD items showed a significant decrease in frequency of reporting.

DISCUSSION

In this systematic review we evaluated adherence to the Standards for Reporting of Diagnostic Accuracy Studies (STARD). We were able to include 16 studies, together evaluating 1,496 articles on diagnostic accuracy studies. The overall quality of reporting in these articles, published both in general and in disease-specific journals, was moderate, at least through halfway the 2000s, confirming the necessity of the introduction of STARD. Results of overall adherence were consistent among all included studies, and varied from 9.1 to 14.3 items being reported, of the 25 items on the checklist. Several factors were reported to be associated with STARD adherence by individual studies, but none of these associations was confirmed by a second study.

Although modest, there seemed to be an improvement in reporting quality (1.41 items (95% CI 0.65 to 2.18)) in the first years after STARD's publication in 2003 compared with articles published pre-STARD. Even though the CI is wide, this improvement is significant. The fact that the quality of the seven analyses included in this meta-analysis was acceptable, and that all of them showed an increase in reported items (three of them significant), increases our confidence in the estimates of effect.

Our study has several potential limitations. Most of the studies evaluated articles on diagnostic accuracy studies published before 2006; none evaluated articles published after 2010. Therefore, we cannot comment on how diagnostic accuracy studies currently adhere to STARD. Most of the included studies reported substantial inter-rater agreement on individual items, with marked differences between studies in reported frequencies of adherence to specific items (Table 2). There was also considerable heterogeneity in our meta-analysis comparing pre-STARD and post-STARD adherence. It is likely that this can, at least partially, be explained by between-study differences in scoring for specific items. For example, while some studies indicated that for item 3 at least the inclusion and exclusion criteria had to be reported, others only considered this item as fully reported when the setting and locations were also described. Only seven studies specifically reported how often an item was judged not to be applicable to the evaluated articles, while the others did not. Therefore, we were not always able to do a mathematical correction for non-applicable items. It is difficult to say whether between-study differences in scores of specific items were caused by a great diversity in adherence in the respective research fields, by heterogeneity in methods of scoring, or both. We would have liked to compare the differences in compliance between STARD-adopting and non-adopting journals, and between high-impact and low-impact journals, but were unable to do so, because this information was almost never available in the included studies.

Although the overall quality of reporting was moderate, several items scored relatively good, with a median proportion of 70% or higher: item 2 (research questions/aims), item 4 (participant recruitment), item 7 (reference standard), item 8 (technique of index test and reference standard), item 9 (units/cut-offs/categories of tests), item 15 (study group characteristics) and item 25 (clinical applicability of findings). Worrisome is the fact that more than half of the 25 STARD items had median proportions of adherence under 50%. Especially, the reporting of test methods and results was suboptimal

Seven items scored remarkably poor, with a median proportion of 30% or lower: 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 particularly alarming because several of these items can be related to biased results. If no or incomplete information on such items is reported, the potential for bias cannot be determined. Review bias, which can result when readers of a test have knowledge of the outcome of other tests or additional clinical information (item 11) (3), and verification bias, which occurs when a patient is only tested by the reference standard in case of a positive index test (item 16) (30), are likely to give inflated estimates of diagnostic accuracy. Limited test reproducibility (items 13 and 24), an effect of instrumental and/or observer variability, and not including missing responses or outliers (item 22), can also introduce biased or imprecise accuracy estimates (2). Interestingly, for all the six items that apply to the index test and reference standard, adherence was better for the index test. Since accuracy estimates of an index test completely depend on the reference standard, authors should be encouraged to provide all the relevant information of both tests. Finally, flowcharts were rarely reported, both pre-STARD and post-STARD. Since these highly facilitate a reader's assessment of study design, their use should be further promoted.

Owing to a constant increase in technological and scientific innovations, the number of available diagnostic tests has been growing exponentially over the past decades. Diagnostic tests are indispensable in patient management since many clinical decisions depend on their results. Implementation and proper usage of a test in any given clinical setting should be based on a thorough consideration of its costs, safety and clinical performance and utility. High-quality diagnostic accuracy studies are crucial in this consideration. Compared with other forms of research, diagnostic accuracy studies are probably more sensitive to bias (3;31). The STARD checklist facilitates a complete and transparent reporting of diagnostic accuracy studies and, consequently, allows readers (clinicians, editors, reviewers, policy makers, etc.) to identify sources of bias that may influence the clinical value and generalizability of a test. While reviews of diagnostic studies often struggle with high heterogeneity, complete and transparent reporting would facilitate an identification of potential sources of heterogeneity.

Although we have presented evidence that the quality of reporting of diagnostic accuracy studies is slowly increasing, it seems that there is still significant room for improvement. A recent study showed that adherence to guidelines is also suboptimal in other fields of research (11). Although the scientific community seems to become more and more aware of the importance of transparent reporting, further enforcement of reporting guidelines among researchers, editors and peer reviewers is a necessity. We strongly recommend authors of diagnostic accuracy studies to take STARD into account from the stage of designing the study and onwards. This way, the items can easily be incorporated in the final article. In addition, this may lead to an increased awareness among authors about potential sources of bias, which allows them to take preventive measures and, consequently, also increases the methodological quality of their study. In addition, we recommend that an evaluation of adherence to STARD should be performed on a more recent cohort of diagnostic accuracy studies. A systematic review has recently shown that, after the introduction of the CONSORT (Consolidated Standards of Reporting Trials) statement, adopting journals had a larger increase in reporting quality of randomised controlled trials than non-adopting journals(32). Such information may be useful in the effort to convince journal editors of the necessity of adopting reporting guidelines. Future evaluations can compare reporting quality of diagnostic accuracy studies between STARD-adopting and non-adopting journals. This way, an estimation of the impact of adopting STARD on reporting quality can be made.

COMPETING INTERESTS

The authors declare that they have no competing interests.

FUNDING

No external funding.

REFERENCE LIST

1. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Standards for Reporting of Diagnostic Accuracy. Clin Chem 2003 Jan;49(1):1-6.
2. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Clin Chem 2003 Jan;49(1):7-18.
3. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999 Sep 15;282(11):1061-6.
4. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. JAMA 1995 Aug 23;274(8):645-51.
5. Plint AC, Moher D, Morrison A, Schulz K, Altman DG, Hill C, et al. Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. Med J Aust 2006 Sep 4;185(5):263-7.
6. Moher D, Jones A, Lepage L. Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. JAMA 2001 Apr 18;285(15):1992-5.
7. Turner L, Shamseer L, Altman DG, Weeks L, Peters J, Kober T, et al. Consolidated standards of reporting trials (CONSORT) and the completeness of reporting of randomised controlled trials (RCTs) published in medical journals. Cochrane Database Syst Rev 2012;11:MR000030.
8. Bossuyt PM. STARD statement: still room for improvement in the reporting of diagnostic accuracy studies. Radiology 2008 Sep;248(3):713-4.
9. Smidt N, Rutjes AW, van der Windt DA, Ostelo RW, Bossuyt PM, Reitsma JB, et al. The quality of diagnostic accuracy studies since the STARD statement: has it improved? Neurology 2006 Sep 12;67(5):792-7.
10. Wilczynski NL. Quality of reporting of diagnostic accuracy studies: no change since STARD statement publication--before-and-after study. Radiology 2008 Sep;248(3):817-23.
11. Samaan Z, Mbuagbaw L, Kosa D, Borg D, V, Dillenburg R, Zhang S, et al. A systematic scoping review of adherence to reporting guidelines in health care literature. J Multidiscip Healthc 2013;6:169-88.
12. Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C,

et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol 2007;7:10.

13. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986 Sep;7(3):177-88.

14. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003 Sep 6;327(7414):557-60.

15. Selman TJ, Morris RK, Zamora J, Khan KS. The quality of reporting of primary test accuracy studies in obstetrics and gynaecology: application of the STARD criteria. BMC Womens Health 2011;11:8.

16. Areia M, Soares M, Dinis-Ribeiro M. Quality reporting of endoscopic diagnostic studies in gastrointestinal journals: where do we stand on the use of the STARD and CONSORT statements? Endoscopy 2010 Feb;42(2):138-47.

17. Coppus SF, van d, V, Bossuyt PM, Mol BW. Quality of reporting of test accuracy studies in reproductive medicine: impact of the Standards for Reporting of Diagnostic Accuracy (STARD) initiative. Fertil Steril 2006 Nov;86(5):1321-9.

18. Fontela PS, Pant PN, Schiller I, Dendukuri N, Ramsay A, Pai M. Quality and reporting of diagnostic accuracy studies in TB, HIV and malaria: evaluation using QUADAS and STARD standards. PLoS One 2009;4(11):e7753.

19. Freeman K, Szczepura A, Osipenko L. Non-invasive fetal RHD genotyping tests: a systematic review of the quality of reporting of diagnostic accuracy in published studies. Eur J Obstet Gynecol Reprod Biol 2009 Feb;142(2):91-8.

20. Gomez SN, Hernandez-Aguado I, Lumbreras B. [Observacional study: evaluation of the diagnostic research methodology in Spain after STARD publication]. Med Clin (Barc) 2009 Sep 5;133(8):302-10.

21. Johnson ZK, Siddiqui MA, Azuara-Blanco A. The quality of reporting of diagnostic accuracy studies of optical coherence tomography in glaucoma. Ophthalmology 2007 Sep;114(9):1607-12.

22. Lumbreras B, Jarrin I, Hernandez A, I. Evaluation of the research methodology in genetic, molecular and proteomic tests. Gac Sanit 2006 Sep;20(5):368-73.

23. Paranjothy B, Shunmugam M, Azuara-Blanco A. The quality of reporting of diagnostic accuracy studies in glaucoma using scanning laser polarimetry. J Glaucoma 2007 Dec;16(8):670-5.

24. Rama KR, Poovali S, Apsingi S. Quality of reporting of orthopaedic diagnostic accuracy studies is suboptimal. Clin Orthop Relat Res 2006 Jun;447:237-46.

25. Shunmugam M, Azuara-Blanco A. The quality of reporting of diagnostic accuracy studies in glaucoma using the Heidelberg retina tomograph. Invest Ophthalmol Vis Sci 2006 Jun;47(6):2317-23.

26. Siddiqui MA, Azuara-Blanco A, Burr J. The quality of reporting of diagnostic accuracy studies published in ophthalmic journals. Br J Ophthalmol 2005 Mar;89(3):261-5.

27. Zafar A, Khan GI, Siddiqui MA. The quality of reporting of diagnostic accuracy studies in diabetic retinopathy screening: a systematic review. Clin Experiment Ophthalmol 2008 Aug;36(6):537-42.

28. Zintzaras E, Papathanasiou AA, Ziogas DC, Voulgarelis M. The reporting quality of studies investigating the diagnostic accuracy of anti-CCP antibody in rheumatoid arthritis and its impact on diagnostic estimates. BMC Musculoskelet Disord 2012;13:113.

29. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986 Feb 8;1(8476):307-10.

30. Reitsma JB, Moons KG, Bossuyt PM, Linnet K. Systematic reviews of studies quantifying the accuracy of diagnostic tests and markers. Clin Chem 2012 Nov;58(11):1534-45.

31. Ochodo EA, Bossuyt PM. Reporting the accuracy of diagnostic tests: the STARD initiative 10 years on. Clin Chem 2013 Jun;59(6):917-9.

32. Turner L, Shamseer L, Altman DG, Weeks L, Peters J, Kober T, et al. Consolidated standards of reporting trials (CONSORT) and the completeness of reporting of randomised controlled trials (RCTs) published in medical journals. Cochrane Database Syst Rev 2012;11:MR000030.

CHAPTER 6

Investigating and reporting of sources of heterogeneity in systematic reviews of diagnostic studies

Christiana A. Naaktgeboren
Wynanda A. van Enst
Eleanor E. Ochodo
Joris A.H. de Groot
Lotty Hooft
Mariska M. Leeflang
Patrick M. Bossuyt
Karel G.M. Moons
Johannes B. Reitsma

Journal of Clinical Epidemiology 2014, Jul 22. pii: S0895-4356(14)00229-7

ABSTRACT

Background To examine how authors explore and report on sources of heterogeneity in systematic reviews of diagnostic accuracy studies.

Methods A cohort of systematic reviews of diagnostic tests was systematically identified. Data was extracted on whether an exploration of the sources of heterogeneity was undertaken, how this was done, the number and type of potential sources explored, and how results and conclusions were reported.

Results Of the 65 systematic reviews, 12 did not perform a meta-analysis and 8 of these gave heterogeneity between studies as a reason not to. Of the 53 reviews containing a meta-analysis, 40 explored potential sources of heterogeneity in a formal manner and 27 identified at least one source of heterogeneity. The reviews not investigating heterogeneity were smaller than those that did (median of 8[IQR:5-15] vs. 14[IQR:11-19] primary studies). Twelve reviews performed a sensitivity analysis, 25 stratified analyses, and 19 meta-regression. Many sources of heterogeneity were explored compared to the number of primary studies in a meta-analysis (median ratio 1:5). Review authors placed importance on the exploration of sources of heterogeneity; 37 mentioned the exploration or the findings thereof in the abstract or conclusion of the main text.

Discussion/Conclusion Methods for investigating sources of heterogeneity varied widely between reviews. Based on our findings of the review, we made suggestions on what to consider and report on when exploring sources of heterogeneity in systematic reviews of diagnostic studies.

What is new?

Key findings

- A wide variety of approaches are used for exploring sources of heterogeneity in reviews of diagnostic studies.
- Exploring and reporting sources of heterogeneity is complex in reviews of diagnostic tests as they typically focus on two potential correlated outcomes (sensitivity and specificity).

What this adds to what was known

- Inspired by the variety of approaches observed in this review, a list of items to consider and report on when exploring sources of heterogeneity in diagnostic reviews was developed.

What is the implication, what should change now

- Further guidance for exploring sources of heterogeneity could improve the strength and usefulness of systematic reviews of diagnostic studies.

INTRODUCTION

As with any review, the results between studies in a review of diagnostic tests are likely to be different, also referred to as variability or heterogeneity in results. Some heterogeneity in the results between studies can be expected simply due to chance variation. Even if studies are methodologically identical and carried out in the same population, their results will vary because each study only observes a finite sample from the total population of interest. This variation is known as chance variation, and is directly linked to the sample size of a study.

Statistical tests and measurements (such as Cochran's Q test or I^2) are often used to conclude whether there is more heterogeneity than is expected due to chance alone (1). If there is more heterogeneity than expected due to chance alone, this is termed systematic differences, statistical heterogeneity, or 'true' heterogeneity. In a random effects model, this 'true' heterogeneity is anticipated and such models then estimate its magnitude with a metric known as τ^2 or the between-study variance (2).

When there are indications that there is 'true' heterogeneity, it is likely that something is causing the heterogeneity (e.g. the index test's performance varies between settings or the study designs differ between studies), and reviewers are encouraged to look into the possible causes of this heterogeneity (3;4).

Unexplained heterogeneity in a review usually results in a downgrading of the quality of the evidence it provides (5;6). Identifying whether there are systematic differences in accuracy of the index test between studies is an important step in translating the results of the review to clinical practice.

The pooling of the results from diagnostic studies in a meta-analysis has an additional level of complexity compared to intervention meta-analyses in that there are usually two correlated analytical outcome measures of interest, namely the sensitivity and specificity of the index test. Similar to intervention reviews, there can be many causes for true heterogeneity, including both clinical and non-clinical factors, such as age, disease spectrum, or study design characteristics. However, a special additional source of heterogeneity that reviews of diagnostic studies may present is related to having two correlated outcomes of interest. Sensitivity and specificity are often negatively correlated due to implicit or explicit differences in the index test threshold. This so-called threshold effect adds an additional layer of complexity to the exploration of sources of heterogeneity in meta-analyses of diagnostic studies.

While guidelines for investigating the sources of heterogeneity in results in systematic reviews of interventions have been established (7), this is not yet the case for systematic reviews of diagnostic studies. The number of systematic reviews of diagnostic studies published each year is rapidly increasing and the methods for performing such studies have seen many technical developments over the past years (3;8).

The aim of this methodological review of the literature is to document how sources of heterogeneity are currently being explored in systematic reviews of diagnostic accuracy studies and to propose a list of items for researchers to consider and report on when performing such an exploration.

METHODS

Overarching project

This study was a part of a meta-epidemiologic project on systematic reviews of diagnostic studies. The goal of this project was to investigate several methodological topics such as small sample size effects, time lag bias, quality assessment, and how to interpret tests and measurements of heterogeneity.

Selection of review articles

Systematic reviews of diagnostic test accuracy studies were identified on September 12th using a systematic search in EMBASE and MEDLINE indexed journals between May 1st and September 11th, 2012 (see Appendix 1). Titles and abstracts were screened and then full texts were read to make a final selection. We distinguished between reviews with and without meta-analyses. A meta-analysis was defined as a review in which a pooled estimate for at least one accuracy estimator was reported or, alternatively, in which a summary ROC curve (sROC) was provided.

As this article is about formal methods for investigating sources of heterogeneity, as opposed to narrative descriptions of heterogeneity, the primary articles of interest were reviews that contained a meta-analysis. However, as one approach to deal with a high amount heterogeneity is to not pool the results in a meta-analysis, we also performed a brief subsidiary examination of systematic reviews without a meta-analysis to document the reasons review authors provided for not pooling. Reviews on prognostic tests (those used to predict a future condition or event rather than to test for the presence or absence of a current one), testing in animals, individual patient data reviews, conference abstracts, and written in languages other than English were excluded.

Data extraction from the reviews

The data extraction form was pilot tested by performing double data extraction on a third of the articles (by C.N., W.E., E.O., J.G., L.H., and M.L.). Discrepancies were discussed and unclear questions on the form were made more specific. Data extraction was then performed by one researcher (by C.N., W.E., and E.O.) using the standardized form and checked by another (by C.N., W.E., or E.O.).

Systematic reviews often contain more than one meta-analysis. In order to prevent the dominance of reviews containing multiple meta-analyses, only information from the main meta-analysis was collected. For objectiveness and clarity in data-extraction, the main meta-analysis was defined as the largest group of studies for which a meta-analysis was performed. We thought that the largest meta-analysis was also most likely to have explored sources of heterogeneity. As we were more interested in the range of possibilities for exploring sources of heterogeneity than in precise counts of methods used, we do not think that this selection of the largest meta-analyses will bias our conclusions.

In addition to general review characteristics gathered for the overarching project, information was extracted on the following: whether sources of heterogeneity were explored, the number and type of sources explored, the methods used to explore these sources, how these results were reported, and how conclusions about sources of heterogeneity were made.

For the systematic reviews without a meta-analysis, we extracted the reasons why review authors refrained from calculating a pooled estimate. We were particularly interested in seeing if heterogeneity was one of the reasons given for not performing a meta-analysis. When the reviews with a meta-analysis did not explore sources of heterogeneity, information was extracted about the reasons why they did not. What review authors reported about how they intended to make a decision on whether to explore sources of heterogeneity was also recorded.

When counting the number of potential sources of heterogeneity explored (which we refer to hereafter as "factors"), the types of factors were counted, rather than the number of subgroups or strata of those factors. For example, if threshold effects were explored and summary estimates were presented for several cutoff points, threshold was only counted as one factor. The relationship (ratio) between the number of factors explored and the number of primary studies in the review was analyzed, as well as the relationship between the number of factors explored and the method used to perform the exploration.

The factors explored were categorized as clinical, quality (i.e. study design characteristics), index test related, or "other". Explanation of some of the quality related factors explored can be found in QUADAS-2, a revised tool for the quality assessment of diagnostic accuracy studies (9). Publication year of the target disease under study, was categorized under the category "other" because it is often difficult to know what it truly measures. For example, over time, technological advances may result in the accuracy of an imaging test improving, but at the same time the patient spectrum could also change. In such a situation, publication year could be categorized as a clinical or an index test related source of heterogeneity. In addition to categorizing the factors, it was also noted whether continuous factors were categorized or sum scores (scores which summarize information about several factors into a single value) were used.

The methods used to explore sources of heterogeneity were classified into three categories: sensitivity analysis, stratification, and meta-regression. We defined sensitivity analysis as the exclusion of one or more studies from the meta-analysis for the purpose of seeing how the summary estimates in the reduced group differ from the overall estimate (10). Stratified analysis was defined as the calculation of summary estimates for subgroups defined by a particular factor (e.g. providing separate estimates of sensitivity and specificity for each gender). Meta-regression was defined as the entering of a factor or factors into a meta-regression model as coefficients to explore how they influenced the summary estimates (11).

Since there are different ways to come to conclusions about whether a specific factor is a source of heterogeneity, what the authors reported about how they made this conclusion was recorded. Additionally, it was noted whether the sources were tested statistically, whether the results were presented per subgroup, and whether the reduction in heterogeneity or remaining heterogeneity (within a subgroup compared to all groups combined) was reported.

Information was extracted on what authors reported in the abstract and conclusion about the exploration of sources of heterogeneity. Studies that discussed this in either the abstract or conclusion were considered to place a high importance on this topic, as these are the sections in which the most important findings are typically discussed and which most readers often base their own conclusions upon (12).

Data extraction from the reviews

As methodological reviews should go beyond only describing what has been done to provide assistance to researchers (11), a list of items for researchers to consider and report on when exploring sources of heterogeneity in a systematic review of diagnostic studies was developed. The domains in the list parallel data extraction (i.e. whether to explore sources of heterogeneity, selection of factors to explore, methods of exploration, and presentation and interpretation of results) and the contents were inspired by the variety of approaches observed in the reviews.

RESULTS

Search results

After exclusion of duplicates, the search resulted in a total of 1273 hits. Upon screening of titles as well as the exclusion of articles that were only conference abstracts, 1058 articles were excluded. The full text of the remaining 89 potentially relevant articles was reviewed to determine whether they met the inclusion criteria. After this process, 65 systematic reviews were identified of which 53 contained at least one meta-analysis. Appendix 2 contains the search results details and Appendix 3 contains a list of the included reviews.

Table 1. *Characteristics of all reviews containing a meta-analysis (n=53) and of the subset in which heterogeneity was investigated statistically (n=40)*

Characteristic	N	%
Number of primary studies (median [IQR])	14	[9.5-18.5]
Size of primary studies (median [IQR])	87	[45-182]
Type of study		
Image	32	60%
Lab	15	28%
Clinical Examination	6	11%
Meta-analyses looking at more than one index test	31	58%
2 tests	14	26%
3-6 tests	16	30%
>6 tests	1	2%
Contained a comparative question	21	40%
Testing and measuring heterogeneity		
Cochran's Q-test	28	53%
I^2	31	58%
τ^2	7	13%
Prediction intervals, ellipses, or bands	3	6%
Method(s) for conducting the meta-analysis		
Univariate analysis only	13	26%
SROC (Moses-Littenberg): linear regression D on S	24	48%
HSROC (Rutter and Gatsonis): accuracy, scale and threshold parameter	5	5%
Bivariate random effects model (Reitsma): random effects sens & spec	13	26%
Studies performing a quality assessment		
QUADAS	40	75%
QUADAS-2	1	2%
STARD	5	9%
Other or own instrument	5	9%
No quality assessment	4	8%
Studies investigating sources of heterogeneity statistically	40	75%
Methods for investigating sources of heterogeneity (n=40)†		
Sensitivity Analysis	12	30%
Stratified Analysis	25	63%
Meta-Regression	19	48%
Number of potential sources of heterogeneity explored/number of primary studies in meta-analysis (n=40) (median [IQR])	0.21	[.09-.46]
Number of reviews identifying sources of heterogeneity (n=40)		
At least 1	29	73%
More than 1	8	20%

† *These numbers do not add up to 40 because some studies used more than one of the methods.*

General characteristics of the reviews

Of the 12 systematic reviews that did not perform a meta-analysis, eight stated that they did not do so because there was too much heterogeneity. Other reasons given for not performing a meta-analysis were low quality of the primary studies (n=4), too few primary studies (n=2), and studies having different cut-offs (n=1).

The general characteristics of the 53 reviews that contained a meta-analysis can be found in Table 1. The meta-analyses contained a median of 14 primary studies [IQR 9.5-18.5]. The majority of reviews were on imaging tests (60%), a large percentage was on lab tests (26%), and a few were on clinical examination procedures (14%). Over half of the meta-analyses investigated more than one index test and the majority of these contained a comparative question.

More than half of the reviews that contained a meta-analysis tested for heterogeneity using Cochran's Q-test (n=28) and more than half of the reviews measured it using I^2 (n=31).(13;14) Very few presented the between-study variance estimate (τ^2) from a random effects model (n=7), and even fewer interpreted this for the reader by presenting prediction intervals or ellipses (n=3). Prediction regions show the range of values where the true value from new comparable study is likely to be found.(2) When obtaining summary estimates of test accuracy, only about a third used a more advanced hierarchical bivariate model (n=13) (15;16), while approximately half used a SROC according to Moses and Littenberg (n=24) (17), and the rest only undertook univariate pooling (n=13).

Almost all of the reviews with a meta-analysis performed a quality assessment of the primary studies (n=49). The vast majority of studies used the formal tool QUADAS (a quality assessment tool for diagnostic accuracy studies) (n=40) (18). As the newer version of this tool, QUADAS-2, was only introduced at the end of 2011, it is logical that it was only used in one of the reviews (19).

Number and type of sources of heterogeneity explored

Forty of the 53 reviews containing a meta-analysis formally explored sources of heterogeneity. There was a large spread in the number of factors that were explored as potential sources of heterogeneity (Figure 1). The median ratio of factors explored relative to the number of primary studies contained in a meta-analysis, was approximately 1 factor for every 5 primary studies. In 33 of the 40 meta-analyses exploring sources of heterogeneity (80%), more than one factor was explored for every ten studies included. Note that we only counted factors that were actually formally explored. Authors often stated in the methods that they would look at many factors as sources of heterogeneity, but for various reasons (e.g. too few studies in the meta-analysis or poor reporting on the factor of interest) some of them could not be explored.

A breakdown of the number of meta-analyses investigating particular types of factors can be found in Table 2. There did not appear to be a difference in the frequency in which any of the categories of factors (i.e. clinical, quality (study design characteristics), or index test related) were explored. The factors categorized under the "other" category of sources of heterogeneity were the percentage of index test positives (n=1), studies that appeared to be outliers based on visual assessment of the ROC curve (n=1), and a leave-one-out (outlier) analyses (n=2).

While some studies reported that they pre-specified what factors they would explore, some reviews may have decided which factors to explore based (partially) on visual exploration of the results presented in forest or ROC plots. This was not information that we extracted from the meta-analyses, as it is often impossible to tell whether authors selected factors prior to or after the gathering of the primary study results. However, we did observe a difference in approaches from comments made by the authors. For example, one reviewer reported that "heterogeneity was evaluated visually through observed differences between study characteristics and methodologies, and by examining for substantial difference in the sensitivities and specificities on the forest plot. Studies that demonstrated considerable heterogeneity were excluded from the meta-analysis"(20).

Of the 25 reviews in which continuous factors were explored (such as mean age, publication year, or prevalence), 15 studies dichotomized these factors. In the other 10 it was unclear how these factors were explored. Only 3 of the 15 reviews that had dichotomized the factor of interest provided the reason behind the chosen cut-off. Of the 25 reviews that explored quality items, quality sum-scores were explored as a potential factor of heterogeneity in 5 reviews.

Methods of investigating sources of heterogeneity

Of the 40 studies investigating heterogeneity, 12 performed sensitivity analysis, 25 stratification, and 19 meta-regression (Table 3). Fifteen studies used two of these methods and 1 used all three.

The number of factors explored varied by methods for investigating heterogeneity. Sensitivity analysis was only conducted on 1 or 2 factors per

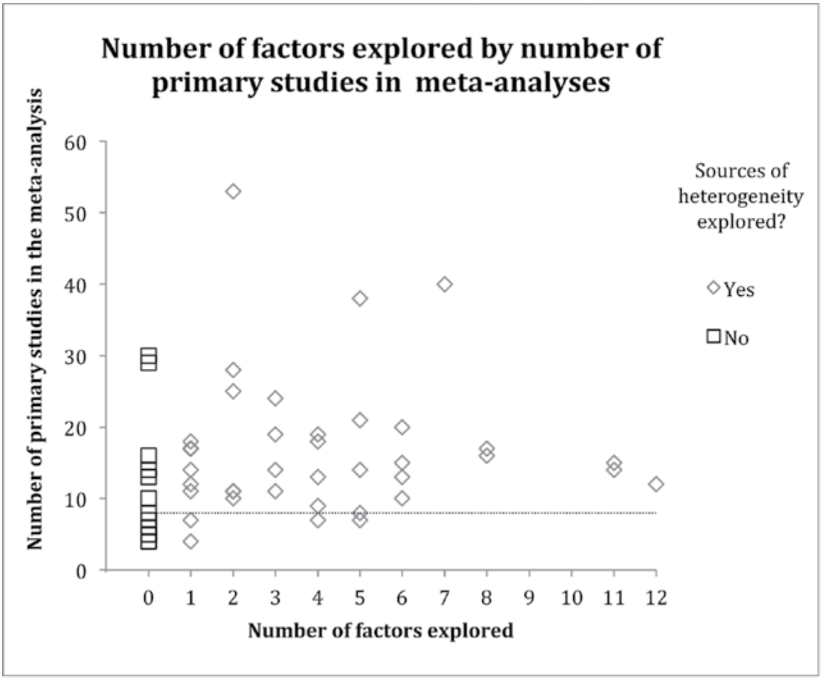


Figure 1. *Relationship between the numbers of potential sources of heterogeneity explored (factors) and the number of primary studies in the meta-analysis. The dashed line is drawn at 8 studies, the median number of studies in meta-analyses that did not explore sources of heterogeneity. One outlying study with 114 primary studies which explored 11 factors of heterogeneity was excluded to improve readability.*

meta-analysis, while a median of 3 factors were explored using stratification, and 4 through meta-regression. Comparing stratification to meta-regression, a high numbers of factors (i.e. >6) were only explored using meta-regression. While the number of factors explored varied between the methods, there was no difference observed by method type in terms of the number of primary studies or the total number of subjects in the meta-analysis.

In the studies using meta-regression, it was often unclear whether they had explored factors one by one (e.g. by fitting multiple models), or whether they entered multiple factors into a single model. In 7 of the 18 studies that had looked at more than one potential factor, it was reported that multiple factors were put into the model at the same time. However, in the other 11 studies, it was unclear how factors had been entered and removed from the model.

Table 2. *Sources of heterogeneity explored (n=40)*

Category	Factor	No. of studies investigating factor (n=40)
Clinical Factors		25
	Age	7
	Sex	2
	Spectrum or other clinical-related factors	18
	Prevalence	8
Study quality items†		25
	Blinding	7
	Sample size	9
	Reference test	9
	Verification biases	5
	Quality score	5
	Prospective vs. Retrospective design	8
	Consecutive vs. Non-consecutive enrolment	4
	Other quality items	8
Test/Threshold		20
	Variation of the index test	15
	Index test threshold	7
Other		
	Publication year	9
	Other not-classified‡	4

† *Explanation of how some of these quality items can introduce bias can be found in QUADAS-2, a revised tool for the quality assessment of diagnostic accuracy studies(9)*

‡*Other: studies that appeared to be outliers based on visual assessment of the ROC curve (n=1) percentage of index test positives (n=1), and a leave-one-out sensitivity analysis (n=2)*

Reviews that did not examine sources of heterogeneity

Although most meta-analyses attempted to explain the variety in study results in a descriptive manner, one fourth (n=13) did not explore sources of heterogeneity formally. Meta-analyses that did not explore sources of heterogeneity were not very different from those that did. Overall, they were slightly smaller in terms of the number of primary studies and participants included in the meta-analysis (Table 3). Still, several studies that explored sources of heterogeneity were smaller than those that had not (Figure 1). Authors concluded that there was significant heterogeneity in about two-thirds of the meta-analyses in both groups (those exploring and those not exploring sources of heterogeneity) (Table 3).

Of the 13 meta-analyses that did not report on the formal exploration of sources of heterogeneity, only one author reported the reason not to, namely, that there were too few studies. In the methods section, 4 articles (out of the 53 meta-analyses) announced that the results of the tests for true heterogeneity would influence the decision of whether to explore sources of heterogeneity.

Interpretation and presentation of results

Although many reviews explored sources of heterogeneity (n=40), the methods to which they came to conclusions about sources of heterogeneity, the thoroughness to which they reported their results, and the importance that was given to the findings of this exploration varied (Table 4).

In total, only 11 (28%) gave a clear description of how they defined sources of heterogeneity. A variety of methods for defining a significant source of heterogeneity was observed, such as comparing the confidence intervals of subgroups, looking at the p-value of the regression coefficient in meta-regression, and testing if 'true' heterogeneity was (still) present within the subgroups. Twenty-nine studies (73%) identified at least one source of heterogeneity and 8 (20%) identified more than one source. Of these 29 studies, only 8 (28%) explained how they came to this conclusion.

Some researchers only performed the exploration of sources of heterogeneity without presenting the results in a form that was easy for readers to interpret (i.e. by only performing statistical testing or presenting coefficients from meta-regression) (n=6, 15%), while others (also) presented stratified results for at least one factor explored (n=34, 85%).

Only 3 (8%) of the 40 meta-analyses that had explored sources of heterogeneity did not mention anything about this exploration, or the findings thereof, in either the abstract or the conclusion (the main findings). Of the 29 studies identifying a source of heterogeneity, 25 (86%) reported this finding in the main findings. On the other hand, 14 (35%) authors reported in the main findings that they were either unable to explore (particular) sources of heterogeneity or unable to come to a conclusion about the cause of heterogeneity. When reporting the findings in the abstract, authors usually (n=13, 93%) presented and gave a clinical or methodological explanation for the subgroup results.

Table 3. *Comparison of methods of investigating sources of heterogeneity (n=40)*

	Only a narrative exploration of heterogeneity (n=13)	**Explored sources of heterogeneity statistically (n=40)**	**Sensitivity N=12†**	**Stratification N=25†**	**Meta-Regression N=19†**
No. of factors explored (median [IQR])	-	-	1[1,1]	3[1,4]	4 [3,6]
1	-	-	10	10	1
2 to 3	-	-	2	6	6
4 to 5	-	-	0	9	5
6 or more	-	-	0	0	7
Study Size (median [IQR])					
No. of primary studies included	8 [5-15]	14.5 [11-19]	14.5 [12.25-22.25]	15 [11,18]	15 [11,20]
No. of subjects in meta-analysis	560 [174-1716]	2106 [636-6495]	2150.5 [766, 8635]	1725 [493, 4828]	2576 [1112,13662]
Ratio of no. of factors explored to the number of studies in the meta-analysis (median [IQR])	-	.21 [.09-.46]	.07 [.06, .12]	.15 [.07, .26]	.27 [.15, .50]
Authors concluded that there was significant/ meaningful true heterogeneity	8 (62%)	27 (68%)	-	-	-

† *The numbers do not add up to 40 because some studies used more than one method: 3 studies performed sensitivity analysis and stratification, 4 sensitivity analysis and meta-regression, 8 meta-regression and stratification, and 1 study used all three methods.*

Table 4. *Meta-analyses differ in how they analyze, report, and present their conclusions about the sources of heterogeneity (n=40)*

Analysis and Reporting			
Defined significant source of heterogeneity	11		
Presented results per subgroup	34		
Reduction in heterogeneity was reported	6		
	In abstract	**In conclusions**	**In either**
No mention of sources of heterogeneity	16	5	3
Unable to explore what causes the heterogeneity	0	6	6
Unable to conclude what causes the heterogeneity	4	8	10
Identified factors as sources of heterogeneity	8	24	25
Presented subgroup results	14	not applicable[†]	14
Interpreted subgroup results	13	not applicable[†]	13

[†] *This information was not applicable as most studies present the results in the results section, not the conclusion.*

Table 5. *Summary items to consider and report on when exploring sources of heterogeneity*

Domain	**Key items to consider and report on**
Whether to explore sources of heterogeneity	• Consider and report how this decision will be made • Report why the exploration was not possible
Selecting potential sources of heterogeneity to explore	• Consider whether to limit the number of factors explored • Consider and report on how potential sources will be selected: o Motivated analysis (a few factors thought to be of most particular clinical interest or cause severe bias) or exploratory analysis (many available factors) o A priori or a posteriori selection of factors • Consider exploring individual quality items instead of quality sum-scores • Consider whether each factor is a patient or study level characteristic. o When it is a patient-level factor, consider whether subgroup estimates can be can be extracted (e.g. separate estimates for male and female) as opposed to study-level characteristics (e.g. percentage male and female)

Table 5. *Summary items to consider and report on when exploring sources of heterogeneity (continuing from page 105)*

Domain	Key items to consider and report on
Methods of exploring sources of heterogeneity	• Consider, for each factor being explored, what method to use to explore sources of heterogeneity: o Sensitivity, stratified analysis, or (bivariate) meta-regression • If there are two main outcomes of interest in the study (i.e. sensitivity and specificity), consider using bivariate meta-regression • When performing (bivariate) meta-regression: o Consider and report the form of the factors being explored (categorical or continuous). If factors are categorized, report the cut-off points and reasoning behind them. o Consider and report how factors are entered into the model (a separate model for each factor, all factors in the same model, etc.)
Interpretation and Presentation of results	• Consider and report how conclusions are drawn about what is a significant source of heterogeneity • Consider whether reporting stratified results will help interpretation • Before concluding that a particular factor causes heterogeneity, consider what other closely related factors could also have caused it • Consider and report why factors identified as sources of heterogeneity could cause heterogeneity

DISCUSSION

Strengths and limitations of this review

In addition to documenting how sources of heterogeneity are currently being explored, this review goes one step further than existing reviews (in the following discussion) to provide a list of items that researchers can consider and report on when investigating sources of heterogeneity (8;21). While we do not provide formal guidance, the list of items we provide will be helpful to researchers in that it raises awareness of the various options available. This list can be found in Table 5.

Although our sample size of meta-analyses was somewhat smaller than prior methodological studies on systematic reviews of diagnostic studies (8;21), we think that it was sufficiently large. We think that our sample size was large enough because our goal was to get an idea of the range of approaches used rather than to precisely estimate how many studies took each approach to investigating sources of heterogeneity (11).

Whether to explore sources of heterogeneity statistically

The first decision to make when looking into why results vary between primary studies is whether to explore the potential factors causing this variability in a formal, statistical manner as opposed to simply providing a narrative description. It is important to acknowledge that there are several insurmountable barriers to formally investigating sources of heterogeneity. In addition to a small number of primary studies included in the review, poor reporting in the primary studies, or high similarity of the studies in terms of study design and study population can make investigating sources of heterogeneity difficult (7).

While it makes sense that there is no need to explore the source of heterogeneity if there is no true heterogeneity detected, defining true heterogeneity is challenging. Authors may judge whether there is heterogeneity from visual inspection of the forest or ROC plots. While viewing data can provide insights into the variability between studies, it is subjective and formal inferences about the presence of true heterogeneity can only be made based on statistical tests and measurements (3;22).

That said, statistical tests and measurements of heterogeneity also have their pitfalls. Tests for heterogeneity, such as the Cochrane's Q-statistic have low power for the typical review of diagnostic tests in which often few and also relatively small studies are included (23;24). Likewise, the confidence interval around I^2 will be very large when there are few studies, meaning that there is large uncertainty about heterogeneity. This high degree of uncertainty makes the I^2 difficult to use when making a decision about whether to explore sources of heterogeneity, regardless of the chosen cut point (22). Furthermore, the I^2 does not take into account the correlation between sensitivity and specificity.

The bivariate random effects model provides metrics for heterogeneity that take into account the correlation between sensitivity and specificity (15). This model provides three parameters: between-study variance (τ^2) in sensitivities and specificities as well as the covariance between them. The combination of these three metrics makes it possible to examine total study variance in sensitivity or specificity as well as conditional variance (the variance in sensitivities at a fixed value of specificity or vice versa). However, the τ^2s are difficult for authors to interpret. More guidance is needed on interpreting the tests and measurements for true heterogeneity in diagnostic studies.

Selecting potential sources of heterogeneity to explore

Overall, the included reviews explored a high number of potential sources of heterogeneity compared to the number of studies in the meta-analysis (median

1:5). We caution against the use of testing a high number of factors compared to the number of studies in the meta-analysis to avoid the well-known problem of multiple testing. When too many factors are tested, the risk of incorrectly concluding a factor causes (some of) the heterogeneity increases. There is no recommended ratio of the number of factors to number of studies which can be explored in a meta-analysis, however a common rule of thumb in regression analysis is that for every covariate (in this case factor) there should be at least 10 observations (in this case primary studies) (25;26).

It is difficult to translate this rule to bivariate meta-regression as the initial model itself, without any covariates, already has five parameters that need to be estimated and each covariate that is explored adds two additional parameters to be estimated, instead of only one as is the case in regression analysis.

Although it is difficult to judge exactly how authors choose which sources to explore, two general approaches to selecting sources of heterogeneity to explore were identified: motivated and exploratory. The motivated approach is to carefully select a few factors to explore for which one has reasons to believe that they may lead to differences in accuracy. The exploratory approach is to explore many potential factors regardless of whether there is a strong reason to believe that each factor could influence test accuracy. It is helpful to communicate to the reader whether the choice of factors was motivated or exploratory as well as whether the factors were selected before or after observing the results (27;28).

Some of the factors explored are categorical by nature, but many are continuous, such as age, prevalence of disease, or publication year. Careful thought should be given to whether to categorize factors and it is important to mention the cut-off value(s) as well as the reasoning behind them (29). Additionally, when performing a meta-regression, it is important to consider and report whether factors are explored one by one or multiple factors were entered into the model at once.

Sometimes authors calculate a quality sum-score to get a better feel for which studies are of higher quality than others. Because sum-scores give equal weighting to unequal factors, exploring sum-scores as factors of heterogeneity is generally discouraged (30). Post-hoc exclusion of studies based only visual inspection of the ROC plot analysis or a leave-one-out analysis is discouraged as well.

Sources of heterogeneity can be divided into two distinct groups: those that relate to characteristics of the patients included in a diagnostic study (e.g. gender, age, or severity of symptoms) and those that characterize the primary studies (e.g. whether all patients received the same reference standard or the year of publication). Exploring patient-level characteristics in a review brings additional challenges. In general, the power for examining patent-level characteristics is low unless the individual studies report separate two-by-two tables for the different categories of that factor. When such results are not available, the only approach left is to use study-level summary measures, such as mean age or percentage of males.

The use of study-level summary measures to investigate patient-level characteristics is problematic. For example, if there was a true difference in diagnostic accuracy between genders, this source of heterogeneity would go undetected if each study contained an equal number of males and females. Individual patient data meta-analyses are much more equipped for examining differences in accuracy that relate to patient-level characteristics (31). In general, review authors should be cautious when examining the relevance of patient-level characteristics in their review, unless primary studies report stratified data for that factor.

Methods of exploring sources of heterogeneity

Since it is not necessary to choose the same method of exploration for each potential source of heterogeneity, authors may consider the individual factors they are investigating before choosing an appropriate method. If the main interest is in a specific group of studies or one wants to study the robustness (i.e. thru a leave one out analysis) of the meta-analyses results, a sensitivity analysis can be considered. Visual inspection of the ROC plot can be used to detect outliers or overly influential studies, but it is subjective, especially when study size is not represented. A good reason to do a sensitivity analysis is to get an estimate from only the high quality studies by excluding the low-quality studies or studies with poor reporting. After all, the high quality studies are the ones upon which clinical inferences can best be drawn.

If the interest is in comparing summary estimates between groups (for example, seeing if the test performs differently in primary vs. secondary care), stratified analysis or meta-regression are logical choices. Stratified analysis is more focused on comparing the estimates between the subgroups while meta-regression is focused on whether the factor is associated with a difference in accuracy between the groups. However, results from meta-regression for categorical factors can also be presented in a stratified manner.

In meta-regression, it is possible to explore multiple factors simultaneously and to explore factors in their continuous form. Authors should report details on how factors were entered into the model and in what form. As authors performing meta-regression explored more factors than those doing stratified analysis or sensitivity analysis, it seems like meta-regression is being used more often for exploratory rather than confirmative analysis. However, the problem of multiple testing cannot be avoided by using multiple-regression as opposed to sensitivity analysis or stratified analysis.

Interpretation and presentation of results

Regardless of the chosen method for investigating sources of heterogeneity, it is important that authors define and report how they will conclude whether a factor is a significant source of heterogeneity. Researchers may consider presenting stratified results when a factor is detected to help convey the clinical or methodological relevance.

Although much work can go into investigating sources of heterogeneity, the effect of a factor may not always be identified. On the other hand, when a source of heterogeneity is identified statistically, caution should be exercised against jumping to the conclusion that that factor is actually causing the heterogeneity. If factors are closely related to each other, it is often impossible to determine which factor is causing the heterogeneity. Instead of just attributing heterogeneity to a factor, one should try to explain why that factor could be causing heterogeneity. When a identified sources of heterogeneity is a quality item, it is particularly important to explain what that item is as readers may not be familiar with them (12). Ultimately, the exploration of heterogeneity is performed with the hope that factors will be identified that are relevant for current clinical practice or future research.

Closing remarks

In this review, we found that methods for exploring sources of heterogeneity in meta-analyses of diagnostic studies vary widely between meta-analyses. Based on the variety in methods observed, we developed a list of items to consider and report on. While waiting for formal guidance to be developed, this list can be used by researchers in the meantime to improve the way that they explore sources of heterogeneity and report findings of this exploration in meta-analyses of diagnostic studies.

REFERENCE LIST

1. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002 Jun 15;21(11):1539-58.
2. Riley RD, Higgins JP, Deeks JJ. Interpretation of random effects meta-analyses. BMJ 2011 Feb 10;342:d549.
3. Deeks J, Bossuyt P, Gatsonis C. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0. The Cochrane Collaboration 2010Available from: URL: http://srdta.cochrane.org
4. Cochrane Handbook for Systematic Reviews of Interventions. Chichester: Wiley-Blackwell; 2008.
5. Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 7. Rating the quality of evidence--inconsistency. J Clin Epidemiol 2011 Dec;64(12):1294-302.
6. Schunemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ 2008 May 17;336(7653):1106-10.
7. Higgins JP, Green S. The Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0 ed. 2011.
8. Dahabreh IJ, Chung M, Kitsios GD, et al. Comprehensive Overview of Methods and Reporting of Meta-Analyses of Test Accuracy. Rockville (MD): Agency for Healthcare Research and Quality (US); 2012 Mar 1.
9. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011 Oct 18;155(8):529-36.
10. Porta M. A Dictionary of Epidemiology. Fifth ed. Oxford University Press; 2008.
11. Lilford RJ, Richardson A, Stevens A, Fitzpatrick R, Edwards S, Rock F, et al. Issues in methodological research: perspectives from researchers and commissioners. Health Technol Assess 2001;5(8):1-57.
12. Zhelev Z, Garside R, Hyde C. A qualitative study into the difficulties experienced by healthcare decision makers when reading a Cochrane diagnostic test accuracy review. Syst Rev 2013 May 16;2:32.
13. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003 Sep 6;327(7414):557-60.
14. Rucker G, Schwarzer G, Carpenter JR, Schumacher M. Undue reliance on I(2) in assessing heterogeneity may mislead. BMC Med Res Methodol 2008 Nov 27;8:79-8.

15. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol 2005 Oct;58(10):982-90.
16. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med 2001 Oct 15;20(19):2865-84.
17. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. Stat Med 1993 Jul 30;12(14):1293-316.
18. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol 2003 Nov 10;3:25.
19. Schueler S, Schuetz GM, Dewey M. The revised QUADAS-2 tool. Ann Intern Med 2012 Feb 21;156(4):323-4.
20. Smith TO, Drew BT, Toms AP. A meta-analysis of the diagnostic test accuracy of MRA and MRI for the detection of glenoid labral injury. [Review]. Archives of Orthopaedic & Trauma Surgery 132 (7):905 -19 , 2012 Jul.
21. Dinnes J, Deeks J, Kirby J, Roderick P. A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy. Health Technol Assess 2005 Mar;9(12):1-113, iii.
22. Ioannidis JP, Patsopoulos NA, Evangelou E. Uncertainty in heterogeneity estimates in meta-analyses. BMJ 2007 Nov 3;335(7626):914-6.
23. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002 Jun 15;21(11):1539-58.
24. Ioannidis JP. Interpretation of tests of heterogeneity and bias in meta-analysis. J Eval Clin Pract 2008 Oct;14(5):951-7.
25. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J Clin Epidemiol 1995 Dec;48(12):1503-10.
26. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996 Dec;49(12):1373-9.
27. Thompson SG, Higgins JP. How should meta-regression analyses be undertaken and interpreted? Stat Med 2002 Jun 15;21(11):1559-73.

28. Higgins JP, Thompson SG. Controlling the risk of spurious findings from meta-regression. Stat Med 2004 Jun 15;23(11):1663-82.

29. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med 2006 Jan 15.

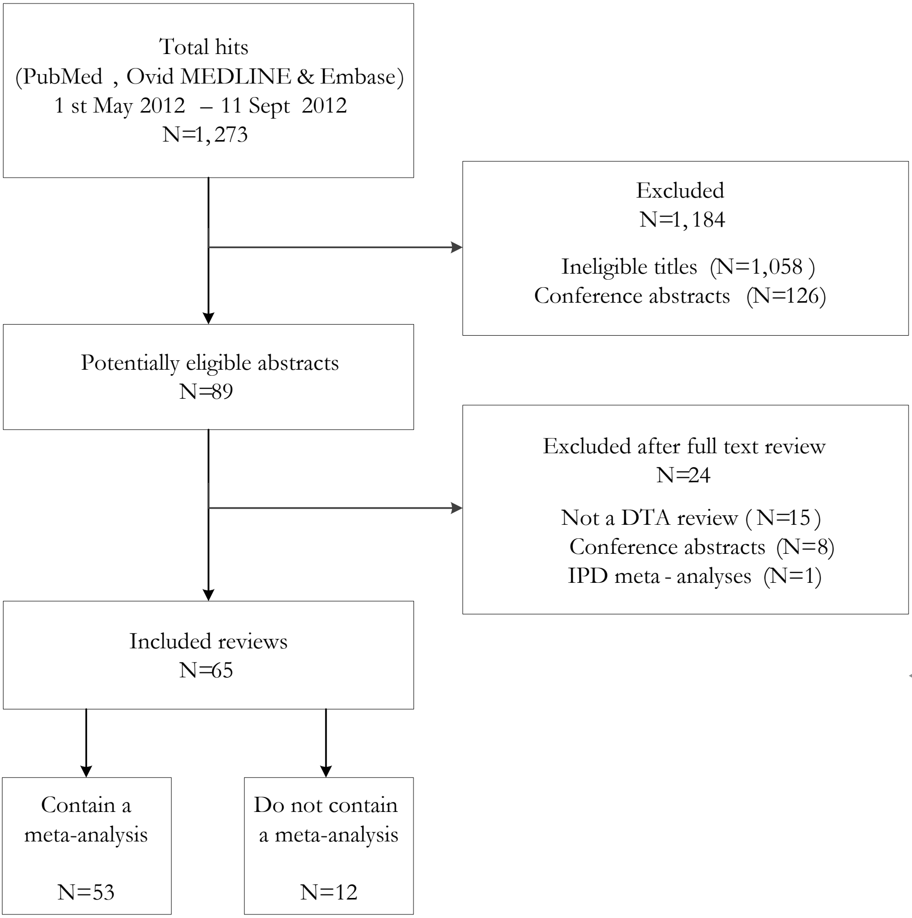
30. Whiting P, Harbord R, Kleijnen J. No role for quality scores in systematic reviews of diagnostic accuracy studies. BMC Med Res Methodol 2005 May 26;5:19.

31. ter RG, Bachmann LM, Kessels AG, Khan KS. Individual patient data meta-analysis of diagnostic studies: opportunities and challenges. Evid Based Med 2013 Oct;18(5):165-9.

APPENDIX 1. *Search strategy*

1. systematic.mp. [mp=ti, ab, sh, hw, tn, ot, dm, mf, dv, kw]
2. limit 1 to “reviews (best balance of sensitivity and specificity)”
3. predict*.ti,ab.
4. test.ti,ab.
5. tests.ti,ab.
6. 4 or 5
7. 2 and 3 and 6
8. screen*.mp. [mp=ti, ab, sh, hw, tn, ot, dm, mf, dv, kw]
9. 2 and 8
10. monitoring.mp. [mp=ti, ab, sh, hw, tn, ot, dm, mf, dv, kw]
11. 2 and 10
12. “multiple tests”.mp. [mp=ti, ab, sh, hw, tn, ot, dm, mf, dv, kw]
13. 2 and 12
14. “diagnostic test accuracy”.mp. [mp=ti, ab, sh, hw, tn, ot, dm, mf, dv, kw]
15. DTA.ti,ab.
16. exp “sensitivity and specificity”/
17. specificit*.tw.
18. “false negative”.tw.
19. accuracy.tw.
20. 14 or 15 or 16 or 17 or 18 or 19
21. 2 and 20
22. 7 or 9 or 11 or 13 or 21
23. limit 22 to (english language and yr=”2011 -2013”)

APPENDIX 2. *Study inclusion flow chart*


Total hits
(PubMed , Ovid MEDLINE & Embase)
1 st May 2012 – 11 Sept 2012
N=1, 273
Excluded
N=1, 184
Ineligible titles (N=1,058)
Conference abstracts (N=126)
Potentially eligible abstracts
N=89
Excluded after full text review
N=24
Not a DTA review (N=15)
Conference abstracts (N=8)
IPD meta - analyses (N=1)
Included reviews
N=65
Contain a meta-analysis
N=53
Do not contain a meta-analysis
N=12

APPENDIX 3. *List of articles included in the review*

Systematic reviews with a meta-analysis: (1-53)
Systematic reviews without a meta-analysis: (54-65)

1. Al-Sukhni E, Milot L, Fruitman M, Beyene J, Victor JC, Schmocker S, Brown G, McLeod R, Kennedy E. Diagnostic accuracy of MRI for assessment of T category, lymph node metastases, and circumferential resection margin involvement in patients with rectal cancer: a systematic review and meta-analysis. Annals of Surgical Oncology 2012 July;19(7):2212-23.
2. Alldred SK, Deeks JJ, Guo B, Neilson JP, Alfirevic Z. Second trimester serum tests for Down's Syndrome screening. Cochrane Database Syst Rev 2012;6:CD009925.
3. Banerjee A, Newman DR, Van den Bruel A, Heneghan C. Diagnostic accuracy of exercise stress testing for coronary artery disease: a systematic review and meta-analysis of prospective studies. [Review]. International Journal of Clinical Practice 2012 May;66(5):477-92.
4. Beynon R, Sterne JA, Wilcock G, Likeman M, Harbord RM, Astin MP, Burke M, Bessell A, Ben-Shlomo Y, Hawkins J, Hollingworth W, Whiting PF. Is MRI better than CT for detecting a vascular component to dementia? A systematic review and meta-analysis. BMC Neurol 2012 June 6;12(1):33.
5. Chang K, Lu W, Wang J, Zhang K, Jia S, Li F, Deng S, Chen M. Rapid and effective diagnosis of tuberculosis and rifampicin resistance with Xpert MTB/RIF assay: a meta-analysis. J Infect 2012 June;64(6):580-8.
6. Chen J, Yang R, Lu Y, Xia Y, Zhou H. Diagnostic accuracy of endoscopic ultrasound-guided fine-needle aspiration for solid pancreatic lesion: a systematic review. Journal of Cancer Research & Clinical Oncology 2012 September;138(9):1433-41.
7. Chen C, Yang Z, Li Z, Li L. Accuracy of several cervical screening strategies for early detection of cervical cancer: A meta-analysis. International Journal of Gynecological Cancer 2012 July;22(6):908-21.
8. Cheng X, Li Y, Liu B, Xu Z, Bao L, Wang J. 18F-FDG PET/CT and PET for evaluation of pathological response to neoadjuvant chemotherapy in breast cancer: a meta-analysis. Acta Radiologica 2012 July 1;53(6):615-27.
9. De Jong MC, Genders TSS, Van GRJ, Moelker A, Hunink MGM. Diagnostic performance of stress myocardial perfusion imaging for coronary artery disease: A systematic review and meta-analysis. European

Radiology 2012 September;22(9):1881-95.

10. Diel R, Loddenkemper R, Nienhaus A. Predictive value of interferon-release assays and tuberculin skin testing for progression from latent TB infection to disease state: A meta-analysis. Chest 2012 July;142(1):63-75.

11. Evangelista L, Cervino AR, Ghiotto C, Al-Nahhas A, Rubello D, Muzzio PC. Tumor marker-guided PET in breast cancer patients-a recipe for a perfect wedding: a systematic literature review and meta-analysis. [Review]. Clinical Nuclear Medicine 2012 May;37(5):467-74.

12. Fan L, Chen Z, Hao XH, Hu ZY, Xiao HP. Interferon-gamma release assays for the diagnosis of extrapulmonary tuberculosis: a systematic review and meta-analysis. FEMS Immunology & Medical Microbiology 2012 August;65(3):456-66.

13. Jaarsma C, Leiner T, Bekkers SC, Crijns HJ, Wildberger JE, Nagel E, Nelemans PJ, Schalla S. Diagnostic performance of noninvasive myocardial perfusion imaging using single-photon emission computed tomography, cardiac magnetic resonance, and positron emission tomography imaging for the detection of obstructive coronary artery disease: a meta-analysis. J Am Coll Cardiol 2012 May 8;59(19):1719-28.

14. Kiewiet JJ, Leeuwenburgh MM, Bipat S, Bossuyt PM, Stoker J, Boermeester MA. A systematic review and meta-analysis of diagnostic performance of imaging in acute cholecystitis. Radiology 2012 September;264(3):708-20.

15. Kim HP, Vance RB, Shaheen NJ, Dellon ES. The Prevalence and Diagnostic Utility of Endoscopic Features of Eosinophilic Esophagitis: A Meta-analysis. Clinical Gastroenterology & Hepatology 2012 September;10(9):988-96.

16. Kim HP, Vance RB, Shaheen NJ, Dellon ES. The Prevalence and Diagnostic Utility of Endoscopic Features of Eosinophilic Esophagitis: A Meta-analysis. Clinical Gastroenterology & Hepatology 2012 September;10(9):988-96.

17. Kocken M, Uijterwaal MH, de Vries AL, Berkhof J, Ket JC, Helmerhorst TJ, Meijer CJ. High-risk human papillomavirus testing versus cytology in predicting post-treatment disease in women treated for high-grade cervical disease: a systematic review and meta-analysis. [Review]. Gynecologic Oncology 2012 May;125(2):500-7.

18. Li C, Su N, Yang X, Yang X, Shi Z, Li L. Ultrasonography for detection of disc displacement of temporomandibular joint: a systematic review and meta-analysis. [Review]. Journal of Oral & Maxillofacial Surgery 2012 June;70(6):1300-9.

19. Lin CY, Chen JH, Liang JA, Lin CC, Jeng LB, Kao CH. 18F-FDG PET or PET/CT for detecting extrahepatic metastases or recurrent hepatocellular carcinoma: A systematic review and meta-analysis. European Journal of Radiology 2012 September;81(9):2417-22.

20. Lu YY, Chen JH, Lin WY, Liang JA, Wang HY, Tsai SC, Kao CH. FDG PET or PET/CT for Detecting Intramedullary and Extramedullary Lesions in Multiple Myeloma: A Systematic Review and Meta-analysis. Clinical Nuclear Medicine 2012 September;37(9):833-7.

21. Lu Y-Y, Chen J-H, Liang J-A, Wang H-Y, Lin C-C, Lin W-Y, Kao C-H. Clinical value of FDG PET or PET/CT in urinary bladder cancer: A systemic review and meta-analysis. European Journal of Radiology 2012 September;81(9):2411-6.

22. Mavromatis ID, Antonopoulos CN, Matsoukis IL, Frangos CC, Skalkidou A, Creatsas G, Petridou ET. Validity of intraoperative gross examination of myometrial invasion in patients with endometrial cancer: a meta-analysis. Acta Obstetricia et Gynecologica Scandinavica 2012 July;91(7):779-93.

23. Morris RK, Riley RD, Doug M, Deeks JJ, Kilby MD. Diagnostic accuracy of spot urinary protein and albumin to creatinine ratios for detection of significant proteinuria or adverse pregnancy outcome in patients with suspected pre-eclampsia: systematic review and meta-analysis. BMJ 2012;345:e4342.

24. Neto AS, Nassar AP, Jr., Cardoso SO, Manetta JA, Pereira VG, Esposito DC, Damasceno MC, Slooter AJ. Delirium screening in critically ill patients: a systematic review and meta-analysis. [Review]. Critical Care Medicine 2012 June;40(6):1946-51.

25. Pai NP, Balram B, Shivkumar S, Martinez-Cajas JL, Claessens C, Lambert G, Peeling RW, Joseph L. Head-to-head comparison of accuracy of a rapid point-of-care HIV test with oral versus whole-blood specimens: A systematic review and meta-analysis. The Lancet Infectious Diseases 2012 May;12(5):373-80.

26. Phillips B, Wade R, Westwood M, Riley R, Sutton AJ. Systematic review and meta-analysis of the value of clinical features to exclude radiographic pneumonia in febrile neutropenic episodes in children and young people. Journal of Paediatrics & Child Health 2012 August;48(8):641-8.

27. Qu X, Huang X, Yan W, Wu L, Dai K. A meta-analysis of (1)(8) FDG-PET-CT, (1)(8)FDG-PET, MRI and bone scintigraphy for diagnosis of bone metastases in patients with lung cancer. Eur J Radiol 2012

May;81(5):1007-15.

28. Romero J, Xue X, Gonzalez W, Garcia MJ. CMR imaging assessing viability in patients with chronic ventricular dysfunction due to coronary artery disease: a meta-analysis of prospective trials. Jacc: Cardiovascular Imaging 2012 May;5(5):494-508.

29. Sadeghi R, Gholami H, Zakavi SR, Kakhki VR, Horenblas S. Accuracy of 18F-FDG PET/CT for diagnosing inguinal lymph node involvement in penile squamous cell carcinoma: systematic review and meta-analysis of the literature. Clin Nucl Med 2012 May;37(5):436-41.

30. Sadigh G, Carlos RC, Neal CH, Dwamena BA. Ultrasonographic differentiation of malignant from benign breast lesions: A meta-analytic comparison of elasticity and BIRADS scoring. Breast Cancer Research and Treatment 2012 May;133(1):23-35.

31. Sandroni C, Cavallaro F, Marano C, Falcone C, De SP, Antonelli M. Accuracy of plethysmographic indices as predictors of fluid responsiveness in mechanically ventilated adults: a systematic review and meta-analysis. Intensive Care Medicine 2012 September;38(9):1429-37.

32. Shang Y, Ju W, Kong Y, Schroder PM, Liang W, Ling X, Guo Z, He X. Performance of polymerase chain reaction techniques detecting perforin in the diagnosis of acute renal rejection: a meta-analysis. PLoS ONE [Electronic Resource] 2012;7(6):e39610.

33. Shen Y-C, Liu M-Q, Wan C, Chen L, Wang T, Wen F-Q. Diagnostic accuracy of vascular endothelial growth factor for malignant pleural effusion: A meta-analysis. Experimental and Therapeutic Medicine 2012 June;3(6):1072-6.

34. Siddiqui MR, Ashrafian H, Tozer P, Daulatzai N, Burling D, Hart A, Athanasiou T, Phillips RK. A diagnostic accuracy meta-analysis of endoanal ultrasound and MRI for perianal fistula assessment. [Review]. Diseases of the Colon & Rectum 2012 May;55(5):576-85.

35. Singh B, Parsaik AK, Agarwal D, Surana A, Mascarenhas SS, Chandra S. Diagnostic accuracy of pulmonary embolism rule-out criteria: a systematic review and meta-analysis. [Review]. Annals of Emergency Medicine 2012 June;59(6):517-20.

36. Smith TO, Lewis M, Song F, Toms AP, Donell ST, Hing CB. The diagnostic accuracy of anterior cruciate ligament rupture using magnetic resonance imaging: A meta-analysis. European Journal of Orthopaedic Surgery and Traumatology 2012 May;22(4):315-26.

37. Smith TO, Drew B, Toms AP, Jerosch-Herold C, Chojnowski AJ.

Diagnostic accuracy of magnetic resonance imaging and magnetic resonance arthrography for triangular fibrocartilaginous complex injury: a systematic review and meta-analysis. [Review]. Journal of Bone & Joint Surgery - American Volume 2012 May 2;94(9):824-32.

38. Smith TO, Drew BT, Toms AP. A meta-analysis of the diagnostic test accuracy of MRA and MRI for the detection of glenoid labral injury. Archives of Orthopaedic and Trauma Surgery 2012 July;132(7):905-19.

39. Tai T-W, Wu C-Y, Su F-C, Chern T-C, Jou I-M. Ultrasonography for Diagnosing Carpal Tunnel Syndrome: A Meta-Analysis of Diagnostic Test Accuracy. Ultrasound in Medicine and Biology 2012 July;38(7):1121-8.

40. Tashakkor AY, Nicolaou S, Leipsic J, Mancini GB. The Emerging Role of Cardiac Computed Tomography for the Assessment of Coronary Perfusion: A Systematic Review and Meta-analysis. Canadian Journal of Cardiology 2012 July;28(4):413-22.

41. Thaker NG, Turner JD, Cobb WS, Hussain I, Janjua N, He W, Gandhi CD, Prestigiacomo CJ. Computed tomographic angiography versus digital subtraction angiography for the postoperative detection of residual aneurysms: A single-institution series and meta-analysis. Journal of Neuro-Interventional Surgery 2012 May;4(3):219-25.

42. Thangaratinam S, Brown K, Zamora J, Khan KS, Ewer AK. Pulse oximetry screening for critical congenital heart defects in asymptomatic newborn babies: a systematic review and meta-analysis. Lancet 2012 May 1.

43. Treglia G, Castaldi P, Rindi G, Giordano A, Rufini V. Diagnostic performance of Gallium-68 somatostatin receptor PET and PET/CT in patients with thoracic and gastroenteropancreatic neuroendocrine tumours: A meta-analysis. Endocrine 2012 August;42(1):80-7.

44. Underwood M, Arbyn M, Redman C, Smith WP. Accuracy of colposcopic directed punch biopsies: A systematic review and meta-analysis. BJOG: An International Journal of Obstetrics and Gynaecology 2012 June;Conference(var.pagings):163.

45. van Teeffelen AS, Van Der Heijden J, Oei SG, Porath MM, Willekes C, Opmeer B, Mol BW. Accuracy of imaging parameters in the prediction of lethal pulmonary hypoplasia secondary to mid-trimester prelabor rupture of fetal membranes: a systematic review and meta-analysis. Ultrasound in Obstetrics & Gynecology 2012 May;39(5):495-9.

46. Wang Z, Dong ZY, Chen JQ, Liu JL. Diagnostic value of sentinel lymph node biopsy in gastric cancer: a meta-analysis. [Review]. Annals of Surgical Oncology 2012 May;19(5):1541-50.

47. Webb RC, Howard RS, Stojadinovic A, Gaitonde DY, Wallace MK, Ahmed J, Burch HB. The utility of serum thyroglobulin measurement at the time of remnant ablation for predicting disease-free status in patients with differentiated thyroid cancer: a meta-analysis involving 3947 patients. Journal of Clinical Endocrinology & Metabolism 2012 August;97(8):2754-63.

48. Wu L, Dai ZY, Qian YH, Shi Y, Liu FJ, Yang C. Diagnostic Value of Serum Human Epididymis Protein 4 (HE4) in Ovarian Carcinoma: A Systematic Review and Meta-Analysis. International Journal of Gynecological Cancer 2012 September;22(7):1106-12.

49. Wu L-M, Gu H-Y, Qu X-H, Zheng J, Zhang W, Yin Y, Xu J-R. The accuracy of ultrasonography in the preoperative diagnosis of cervical lymph node metastasis in patients with papillary thyroid carcinoma: A meta-analysis. European Journal of Radiology 2012 August;81(8):1798-805.

50. Wu L-M, Xu J-R, Ye Y-Q, Lu Q, Hu J-N. The clinical value of diffusion-weighted imaging in combination with T2-weighted imaging in diagnosing prostate carcinoma: A systematic review and meta-analysis. American Journal of Roentgenology 2012 July;199(1):103-10.

51. Wu L-M, Hu J-N, Hua J, Liu M-J, Chen J, Xu J-R. Diagnostic value of diffusion-weighted magnetic resonance imaging compared with fluorodeoxyglucose positron emission tomography/computed tomography for pancreatic malignancy: A meta-analysis using a hierarchical regression model. Journal of Gastroenterology and Hepatology 2012 June;27(6):1027-35.

52. Yuan Y, Gu ZX, Tao XF, Liu SY. Computer tomography, magnetic resonance imaging, and positron emission tomography or positron emission tomography/computer tomography for detection of metastatic lymph nodes in patients with ovarian cancer: a meta-analysis. Eur J Radiol 2012 May;81(5):1002-6.

53. Zhao L, He Z-Y, Zhong X-N, Cui M-L. 18FDG-PET/CT for detection of mediastinal nodal metastasis in non-small cell lung cancer: A meta-analysis. Surgical Oncology 2012 September;21(3):230-6.

54. Berger MY, Tabbers MM, Kurver MJ, Boluyt N, Benninga MA. Value of abdominal radiography, colonic transit time, and rectal ultrasound scanning in the diagnosis of idiopathic constipation in children: A systematic review. Journal of Pediatrics 2012 July;161(1):44-50.

55. Atluri S, Singh V, Datta S, Geffert S, Sehgal N, Falco FJ. Diagnostic

accuracy of thoracic facet joint nerve blocks: an update of the assessment of evidence. Pain Physician 2012 July;15(4):E483-E496.

56. Feitosa LA, Dornelas de AA, Reinaux CM, Britto MC. Diagnostic accuracy of exhaled nitric oxide in exercise-induced bronchospasm: Systematic review. Revista Portuguesa de Pneumologia 2012 July;18(4):198-204.

57. Mejare IA, Axelsson S, Davidson T, Frisk F, Hakeberg M, Kvist T, Norlund A, Petersson A, Portenier I, Sandberg H, Tranaeus S, Bergenholtz G. Diagnosis of the condition of the dental pulp: a systematic review. International Endodontic Journal 2012 July;45(7):597-613.

58. Crawford S, Evans J, Whitnall L, Robertson JA. A systematic review of the accuracy and clinical utility of the addenbrooke's cognitive examination and the addenbrooke's cognitive examination - Revised in the diagnosis of dementia. Brain Impairment 2011;Conference(var.pagings):4.

59. Tijssen M, van CR, Willemsen L, de VE. Diagnostics of femoroacetabular impingement and labral pathology of the hip: a systematic review of the accuracy and validity of physical tests. Arthroscopy 2012 June;28(6):860-71.

60. Quatman CE, Quatman-Yates CC, Schmitt LC, Paterno MV. The clinical utility and diagnostic performance of MRI for identification and classification of knee osteochondritis dissecans. [Review]. Journal of Bone & Joint Surgery - American Volume 2012 June 6;94(11):1036-44.

61. Cook C, Mabry L, Reiman MP, Hegedus EJ. Best tests/clinical findings for screening and diagnosis of patellofemoral pain syndrome: a systematic review. [Review]. Physiotherapy 2012 June;98(2):93-100.

62. Chagpar AB. Accuracy of sentinel lymph node biopsy in large and multifocal/multicentric breast carcinoma - A systematic review. Breast Diseases 2012;23(1):71-2.

63. Hellemons ME, Kerschbaum J, Bakker SJ, Neuwirt H, Mayer B, Mayer G, de ZD, Lambers Heerspink HJ, Rudnicki M. Validity of biomarkers predicting onset or progression of nephropathy in patients with Type 2 diabetes: a systematic review. [Review]. Diabetic Medicine 2012 May;29(5):567-77.

64. Quinn EM, Coveney AP, Redmond HP. Use of magnetic resonance imaging in detection of breast cancer recurrence: a systematic review. Annals of Surgical Oncology 2012 September;19(9):3035-41.

65. Leake PA, Cardoso R, Seevaratnam R, Lourenco L, Helyer L, Mahar A, Law C, Coburn NG. A systematic review of the accuracy and indications for diagnostic laparoscopy prior to curative-intent resection of gastric cancer. Gastric Cancer 2012 September;15 Suppl 1:S38-S47.

CHAPTER 7

Investigation of publication bias in meta-analyses of diagnostic test accuracy: a meta-epidemiological study

W. Annefloor van Enst
Eleanor Ochodo
Rob Scholten
Lotty Hooft
Mariska Leeflang

BMC Medical Research Methodology
2014;14(1):70

ABSTRACT

Background The validity of a meta-analysis can be understood better in light of the possible impact of publication bias. The majority of the methods to investigate publication bias in terms of small study-effects are developed for meta-analyses of intervention studies, leaving authors of diagnostic test accuracy (DTA) systematic reviews with limited guidance. The aim of this study was to evaluate if and how publication bias was assessed in meta-analyses of DTA, and to compare the results of various statistical methods used to assess publication bias.

Methods A systematic search was initiated to identify DTA reviews with a meta-analysis published between September 2011 and January 2012. We extracted all information about publication bias from the reviews and the two-by-two tables. Existing statistical methods for the detection of publication bias were applied on data from the included studies.

Results Out of 1,335 references, 114 reviews could be included. Publication bias was explicitly mentioned in 75 reviews (65.8%) and 47 of these had performed statistical methods to investigate publication bias in terms of small study-effects: 6 by drawing funnel plots, 16 by statistical testing and 25 by applying both methods. The applied tests were Egger's test (n=18), Deeks' test (n=12), Begg's test (n=5), both the Egger and Begg tests (n=4), and other tests (n=2). Our own comparison of the results of Begg's, Egger's and Deeks' test for 92 meta-analyses indicated that up to 34% of the results did not correspond with one another.

Discussion/Conclusion The majority of DTA review authors mention or investigate publication bias. They mainly use suboptimal methods like the Begg and Egger tests that are not developed for DTA meta-analyses. Our comparison of the Begg, Egger and Deeks tests indicated that these tests do give different results and thus are not interchangeable. Deeks' test is recommended for DTA meta-analyses and should be preferred.

INTRODUCTION

When the decision to publish the results of a study depends on the nature and direction of the results, publication bias arises. There are many forms and reasons for publication bias such as time-lag bias (due to delayed publication), duplicate or multiple publications, outcome reporting bias (selective reporting of positive outcomes) and language bias (1-6). These forms of biases tend to have more effect on small studies and contribute to the phenomenon of "small study-effects" (7). This means that published studies with small sample sizes tend to have larger and more favourable effects compared to studies with larger sample sizes. This is a threat to the validity of a systematic review and its meta-analyses (8).

For intervention reviews graphical and statistical methods have been developed to investigate if the results of the meta-analyses of the review might be affected by publication bias in terms of small study-effects. A well-known graphical method is the funnel plot examination (9). This method aims to construct a scatter plot of the study effect sizes on the horizontal axis against some measure of each study's size or precision on the vertical axis. The dots in this plot together look like an inverted funnel. An asymmetric funnel is an indication for publication bias. Since the plot gives a visual relationship between the effect and study size, its interpretation is subjective. This is not an issue when statistical tests are used to detect funnel plot asymmetry. There are eight tests available (10), but the test of Begg (11), and the test of Egger (12) are probably most common. They have been cited more than 2,500 (Begg) and 7,300 times (Egger) (13). The test of Begg assesses if there is a significant correlation between the ranks of the effect estimates and the ranks of their variances. The test of Egger uses linear regression to assess the relation between the standardized effect estimates and the standard error (SE). For both tests a significant result is an indication that the results might be affected by publication bias. These and other methods have been developed especially for systematic reviews of intervention studies and are not automatically suitable for reviews of diagnostic test accuracy (DTA) studies (9).

DTA meta-analyses have different characteristics making assessment of the potential for publication bias more complicated than for intervention reviews. The diagnostic odds ratio (DOR) usually takes high values, while intervention effects are usually quite small. Secondly, the SE of the DOR depends on the proportion of positive tests, but this proportion is influenced by the variation in threshold amongst different studies. Thirdly, the number of diseased and non-diseased patients are usually unequally divided, which reduces the precision of a test accuracy estimate while in RCTs equal numbers of participants are allocated to an intervention or control group. Investigating whether meta-analyses of DTA studies have been influenced by publication bias in terms of small study-effects is challenging (14). Even diagnostic meta-analyses free of publication bias might have an asymmetric funnel plot due to other reasons like the threshold effect. In addition, bivariate meta-analysis is recommended for DTA meta-analyses (14) but bivariate methods for the detection of publication bias are currently not available. Hence, the DOR is used as a univariate alternative to detect publication bias, but not for the final meta-analysis that assesses the accuracy.

Knowledge of the mechanisms that may induce publication bias in diagnostic studies or empirical evidence for the existence of publication bias is scarce. Selective publication of accuracy studies based on the magnitude of the sensitivity or specificity doesn't seem to be very plausible. In addition, what parameter is most important (and thus driving possible selective publication) depends also on the place of the test in the clinical pathway and it's role (15). Korevaar et al. compared prospective registered diagnostic studies to the publications. They concluded that failure to publish and selective publication were prevalent in diagnostic accuracy studies but the dataset was too small to draw firm conclusions (16). Brazelli and colleagues, however, tracked a cohort of conference abstracts and did not find evidence of publication bias in the process that occurs after abstract acceptance (17).

In 2002, Song and colleagues proposed that tests developed for intervention reviews, like Begg's and Egger's methods could also be used to detect publication bias in DTA reviews. They suggested to use the natural logarithm of the DOR (lnDOR) and plot it against its variance or SE and test for asymmetry (18). In 2005, however, Deeks and colleagues conducted a simulation study of tests for publication bias in DTA reviews. They concluded that existing tests that use the SE of the lnDOR can be seriously misleading and often have false positive results (19). The Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy explicitly mentions not to use methods like the Begg or Egger tests and argues that it is best to use the test proposed by Deeks (14). This test has been developed especially for test accuracy reviews and proposes plotting the lnDOR against 1/effective sample size (ESS)$^{1/2}$ and testing for asymmetry of this plot. The ESS is a function of the number of diseased (n1) and non-diseased (n2) participants: (4n1*n2)/(n1+n2). The ESS takes into account the fact that unequal numbers of diseased and non-diseased reduce the precision of the test accuracy estimates (19). Using the ESS instead of total sample size will reduce the unequal numbers of diseased and non-diseased and thereby enhance the precision of the accuracy estimates. The Cochrane Handbook, however, points out that even Deeks' test has low power to detect small study-effects when there is heterogeneity in the DOR. As heterogeneity in DTA reviews is the rule rather than the exception the Cochrane Handbook warns the authors against misinterpretation of this test (14).

Because little is known about the mechanisms behind and the existence of publication bias in DTA studies it is difficult for reviewers to select the correct method for addressing selective publication. In addition, the interpretation of the results of the various methods and incorporating those results in the formulation of the conclusions of the review is even more challenging. Different tests to identify publication bias in terms of small study-effects are expected to report different results. However, since all tests aim at assessing the same concept, publication bias, the differences should be minimal. A simulation study did show that differences in test outcomes are, however, quite substantial (19). This has not been confirmed in empirical data. To understand more about the assessment of publication bias in DTA reviews led us to following objectives.

The primary objective of this study was to assess which existing tests for publication bias have been used and to what extent the results of these tests have been incorporated in the review. A second objective was to compare the results of existing methods for the detection of publication bias in non-simulated data to assess if these various methods would provide similar results.

METHODS

Study selection

MEDLINE was searched through the interface of PubMed for DTA reviews published between September 2011 and January 2012. The search was performed in February 2012 by one author (EO) using a search filter for systematic reviews available from PubMed combined with a methodological filter for DTA studies: (systematic[sb] AND ((“diagnostic test accuracy” OR DTA[tiab] OR “SENSITIVITY AND SPECIFICITY”[MH] OR SPECIFICIT*[TW] OR “FALSE NEGATIVE”[TW] OR ACCURACY[TW]))) (20).

Eligibility criteria

Articles were eligible for inclusion if they systematically assessed the diagnostic accuracy of a test or biomarker and were published in English. Methods to investigate publication bias are developed to investigate publication bias in meta-analyses (14). Therefore, the selection was further limited to reviews that included a meta-analysis. Availability of the two-by-two tables of the included studies was not amongst the inclusion criteria to generate a representative cohort of reviews without possible selection on high level of reporting and perhaps review quality (21). Studies that assessed the accuracy by means of individual patient data were excluded as the methodology of such studies differs from those of meta-analyses on a study level.

Definitions of assessment of publication bias

In determining if authors would assess publication bias in their reviews, we scored if authors described a method how they would investigate publication bias like drawing a funnel plot or performing a test for publication bias. If the methods were lacking but the results of a publication bias assessment were described, it was also scored as an investigation of publication bias. We regarded the results of the assessments as being incorporated in the discussion of the reviews when the authors described how publication bias might have affected the results of their reviews.

Data extraction

An online standardized data extraction form was used to extract data. We first piloted the form among all team members. After everyone agreed on the data-extraction form, the actual extraction was then done by one reviewer (WE). An online randomization program selected a random sample of one third of the reviews that was checked by a second reviewer (ML, FW, RS). In case the number of differences between reviewers was <3%, no further data checking was done. Disagreements were resolved by discussion.

For the first objective, data was extracted on all reported matters concerning assessing publication bias: if the authors had planned to assess or assessed publication bias and the described methods, the number of studies that were included in the test, results of the test, and consideration of the test results with the interpretation of the pooled results. When authors had no intention to test for publication bias, the review was screened to find a reason for this and if the possible threat of publication bias was discussed or considered to formulate the conclusion. For the second objective, the two-by-two tables (true positives, false positives, false negatives, true negatives) were extracted when reported in the reviews or when they could be derived from other results (e.g. number of diseased and non-diseased combined with the sensitivity or specificity).

Comparison of tests for publication bias

The secondary objective of this study was to assess the concordance of publication bias test results in empirical data. We applied three univariate tests: the Begg test and Egger test because these are cited frequently, and Deeks' test because this test has been developed for DTA meta-analyses and is currently recommended in the Cochrane DTA Handbook (14). The tests were performed as follows:

- Begg's test: rank correlation of the lnDOR with the variance of the lnDOR (11);
- Egger's test: linear regression of lnDOR with the standard error of the lnDOR weighted by the inverse variance of the lnDOR (12);
- Deeks' test: linear regression of lnDOR with $1/\mathrm{ESS}^{1/2}$ weighted by the ESS (19).

Concordance between the results of tests defined as both having or not having a significant result (p-value <0.05) was presented as Cohen's weighted kappa, taking into account agreement due to chance. The simulation study of Deeks et al. indicated that tests would more frequently perform differently when the pooled DOR is 38 or higher (19). In addition tests need sufficient power to perform optimal which may be relevant for concordance. Therefore, we performed logistic regression to study whether concordance between tests was related to a pooled DOR >38, the number of primary studies, or the number of included patients. Analyses were performed in the statistical program R (22).

RESULTS

We identified 1,335 references of potential eligible studies, of which 152 were assessed on full text for eligibility. Finally, 114 DTA reviews were included for the current study. Details of the selection process are presented in Figure 1. There was optimal agreement (98.6%) when the second reviewer checked the data.

Publication bias was explicitly mentioned in 75 reviews (65.8%). Of these, 47 (62.7%) had performed methods to investigate publication bias in terms of small study-effects: 6 by investigating funnel plots, 16 by statistical testing for asymmetry and 25 by applying both methods. Table 1 gives details on how publication bias was investigated per review.

In 28 reviews (24.6%), publication bias was mentioned though it was not investigated. Fifteen of these reviews (13.2%) mentioned why they did not investigate publication bias. These reasons were: because the methods to investigate publication are lacking and can provide misleading results (n=7), lack of power to detect publication bias (n=6), too heterogeneous results to further investigate publication bias (n=1), and underlying principles of publication bias in DTA studies are not yet known and publication bias can therefore not be investigated (n=1).

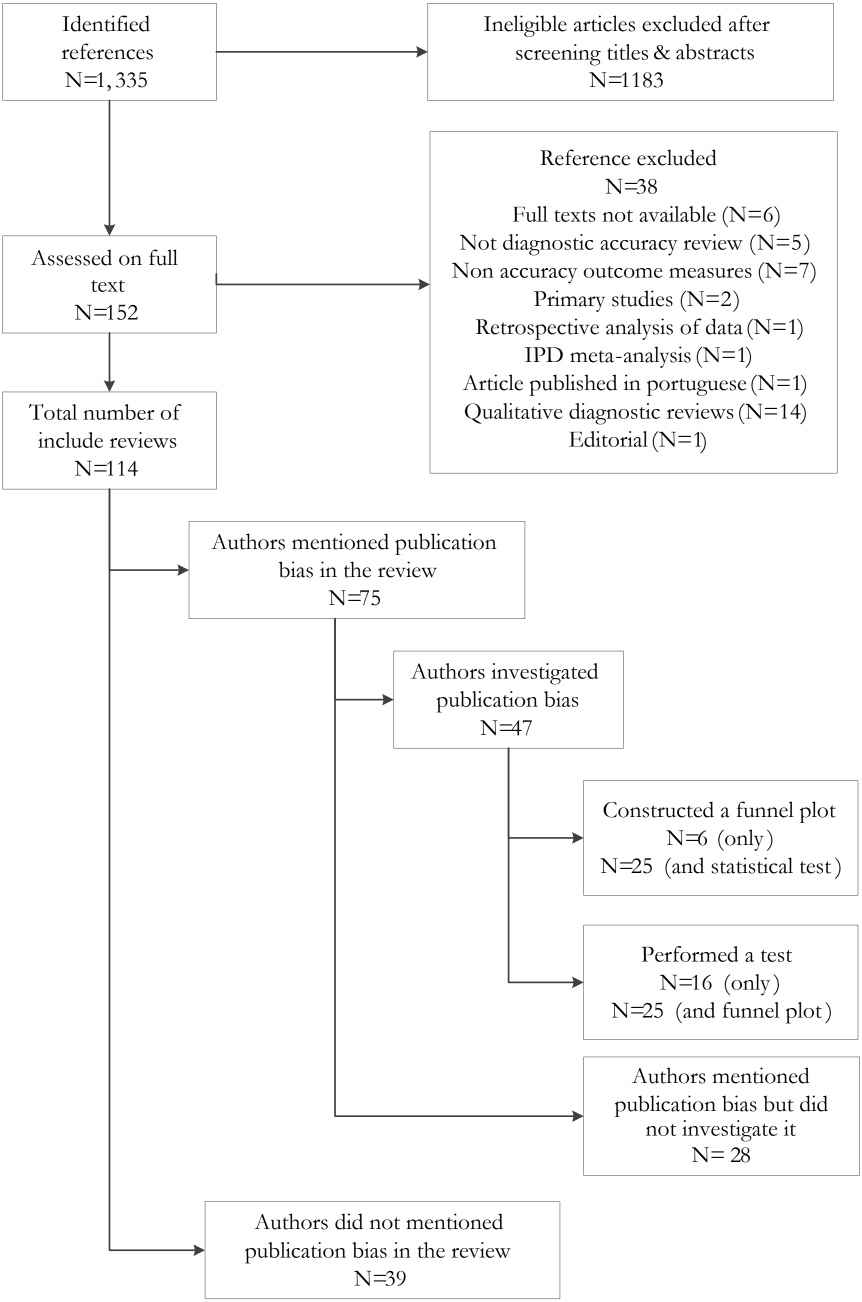

Identified references
N=1, 335
Ineligible articles excluded after screening titles & abstracts
N=1183
Reference excluded
N=38
Full texts not available (N=6)
Not diagnostic accuracy review (N=5)
Non accuracy outcome measures (N=7)
Primary studies (N=2)
Retrospective analysis of data (N=1)
IPD meta-analysis (N=1)
Article published in portuguese (N=1)
Qualitative diagnostic reviews (N=14)
Editorial (N=1)
Assessed on full text
N=152
Total number of include reviews
N=114
Authors mentioned publication bias in the review
N=75
Authors investigated publication bias
N=47
Constructed a funnel plot
N=6 (only)
N=25 (and statistical test)
Performed a test
N=16 (only)
N=25 (and funnel plot)
Authors mentioned publication bias but did not investigate it
N= 28
Authors did not mentioned publication bias in the review
N=39

Figure 1. *Flow chart of the selection process and characters of the included studies*

Table 1. *Overview of the applied methods to investigate publication bias*

Reference	Funnel plot		Results of the funnel plot	Test	Results of the test	Remarks
	x-axis	*y-axis*				
Chang 2011 (23)	-	-	-	Egger	3/7	
Chang 2012 (24)	Sensitivity Specificity	SE	Not considered	Begg Egger	1/2 1/2	
Cheng 2012 (25)	lnDOR	1/root(ESS)	No publication bias	Not specified	0/2	
Descatha 2012 (26)	lnDOR	1/root(ESS)	No publication bias	Deeks	0/2	
Dong 2011 (27)	-	-	-	Begg Egger	0/1 0/1	Results for a second diagnostic tool were not presented.
Dym 2011 (28)	Sensitivity Specificity	1/SE	Inconclusive 2/2	-	-	
Gao 2011 (29)	lnDOR	SE(lnDOR)	1/2	Begg	1/2	
Gargiulo 2011 (30)	lnDOR	1/root(ESS)	Not considered	Deeks	1/2	
Glasgow 2012 (31)	lnDOR	1/Var(lnDOR)	0/2	-	-	
Gong 2011 (32)	Sensitivity	Sample size	Inconclusive 2/2	-	-	Plots had too low power.
Hernaez 2011 (33)	-	-	-	Deeks	0/1	

Inaba 2012 (34)	lnDOR
	RR[1]
		SE(lnDOR)
		SE(RR)
			1/2	Egger[2]	1/2	Level of significance p-value <0.10

Kobayashi 2012 (35)	DOR	SE(DOR)	2/2	Begg	0/2	Both plots indicated publication though the tests were not significant.
Li 2011 (36)	-	-	-	Egger	1/1	Publication bias was detected for a subgroup by the test.
Li 2012 (37)	-	-	-	Egger	1/1	
Lu 2011 (38)	lnDOR	1/root(ESS)	Not considered	Deeks	0/1	
Lundstrom 2011 (39)	-	-	-	Egger	0/1	
Luo 2011 (40)	lnDOR	1/root (ESS)	Not considered	Egger	0/3	
Manea 2012 (41)	-	-	?	Begg	?	Results were not presented
Mao 2012 (42)	-	-	-	Egger	1/1	
Marton 2012 (43)	Not specified	Not specified	Not considered	Egger	1	One plot and test to investigate two diagnostic tools
Mathews 2011 (44)	AUC(ROC)[3]	SE(AUC(ROC))	0/2	Egger	0/2	
McInnes 2011 (45)	lnDOR	SE(lnDOR)	-	Egger	0/1	
Meader 2011 (46)	-	-	-		?	Results were not presented.
Mitchell 2011 (47)	-	-	-		?	Results were not presented.

[1] RR = *Relative Risk; It is unclear which estimates were used to calculate the RR*

[2] *The methods section specifies that the Egger test has been used though the text of the figures specified the Begg test*

[3] *AUC(ROC) = Area Under the Curve (AUC) of the Receiving Operating Characteristic (ROC)*

Onishi 2012 (48)	-	-	-	Egger	2/2	
Papathanasiou 2012 (49)	lnDOR	SE(lnDOR)	Not considered	Begg	1/1	
Plana 2012 (50)	lnDOR	1/root(ESS)	Not considered	Deeks	0/2	Not identified by tests Plots was not used to draw conclusions.
Qu 2011(51)	logDOR	Sample size	?/2	-	-	Results of funnel plots were inconclusive, too low power.

Sadeghi 2012 (52)	logDetection-Rate[4]
	logSensitivity
		SE(logDetect Rate)
		SE(logSens)
			0/2	Egger	0/2	

Sadigh 2011 (53)	-	-	-	Deeks	0/1	
Summah 2011 (54)	lnDOR	SE(lnDOR)	1/1	Egger	1/1	No publication bias was detected by the test.
Sun 2011 (55)	-	-	-	Deeks	0/1	Identified by plot though not by test.
Takakuwa 2011 (56)	lnDOR	1/root (ESS)	1/1	Deeks	0/1	Plots were not used to draw conclusion.
Thosani 2012 (57)	lnDOR	SE(lnDOR)	Not considered	Egger	2/2	Identified by plots though not by tests.
Tomasson 2012 (58)	Difference in arcsine[5]	Precision(Dif. in arcsine)	2/2	Egger	0/2	
Trallero-Araguas 2012 (59)	-	-	-	Deeks	0/1	
Wang 2011 (60)	-	-	-	Begg Egger	0/2 0/2	
Wang 2012 (61)	lnDOR	SE(lnDOR)	7/7	Egger	3/7	
Wang 2012 (62)	lnDOR	SE(lnDOR)	0/2	Begg Egger	0/2	
Wang 2012 (63)	lnDOR	SE(lnDOR)	0/2	-	-	
Wu 2012 (64)	lnDOR	1/root(ESS)	0/1	Deeks	0/1	
Xu 2011 (65)	-	-	-	Egger	0/1	

Xu 2011 (66)	lnDOR
	Standardized effect [6]
		SE(lnDOR)
		Precision(St. effect)
			0/2	Begg-Mazumdar
				Harbord-Egger
					0/2	

Ying 2011(67)	lnDOR	1/root(ESS)	0/2	Deeks	0/2	
Yu 2012 (68)	lnDOR	SE(lnDOR)	1/1	-	-	
Zhang 2011 (69)	lnDOR	1/root(ESS)	0/1	Deeks	0/1	

[4] *There was no definition for Detection Rate specified in the article*

[5] *Difference in arcsine = Transformed ratios of arcsine for those with rise in Anti-Neutrophil Cytoplasmic Antibody (ANCA) and persistent ANCA among subjects who had relapse and those who did not.*

[6] *Standardized effect was explained as differentiating benign and malignant lymph nodes.*

Funnel plots

In the 31 reviews that presented funnel plots, different concepts were plotted. Funnel plots were constructed per test under review (n=20), per target condition (n=2) (e.g. MRI to detect colon cancer or to detect lung cancer) and for different accuracy measures of a test (n=5) (e.g. sensitivity and specificity). In four reviews the authors made comparisons of the accuracy of several clinical tests but used one single plot to investigate publication bias (two of these, however, did construct different funnel plots for different accuracy measures).

The axes that were used to plot were diverse. On the horizontal axis the DOR (DOR or lnDOR) was most often used (n=24), but also other accuracy parameters like sensitivity or ROC area (n=5). Four reviews used other parameters (relative risk, detection rate, difference in the arcsine between two groups, and standardized effect). On the vertical axis we found a variety of precision measures: SE(lnDOR) (n=12), 1/variance(lnDOR) (n=1), 1/(ESS)$^{1/2}$ (n=10), and sample size (n=2). For two reviews the authors had constructed two plots per test: one plot with the sensitivity on the horizontal axis with 1/SE(sens) on the vertical axis and one plot of the specificity on the horizontal axis with 1/SE(spec) on the vertical axis.

Statistical tests

In 41 reviews a statistical test was performed to investigate publication bias. The applied tests were Egger's test (n=18), Deeks' test (n=12), Begg's test (n=5), both the Egger and Begg test (n=4), and both the Begg-Mazumdar and Harbord's test (70). One review did not specify which test was used. Two reviews used the trim and fill method to adjust for small study-effects. The median number of studies in the analyses was 13 (IQR 9-19) with a range from 4 to 118. Two review authors mentioned that a minimum of twenty homogeneous studies was required to perform a test (71;72).
Authors that had applied the Egger test most often reported significant results indicating the existence of publication bias (37.2%), while authors that applied the Deeks test least reported significant results in identifying publication bias (6.7%) (Table 2).

In 8 reviews the authors used more than one test to examine publication bias. The results of both tests in these reviews were in agreement with one another, though the p-values could be quite diverse (e.g. investigation of publication bias of FDG-PET studies to detect in breast cancer: Begg's p=0.462, Egger's p=0.052 (63) or imaging studies to detect osteomyelitis:

Begg's p=0.392 and Egger's p=0.063 (60)).

Table 2. *Reported results of different tests to assess small study in the included reviews (n=41)*

Type of test	Small study effects Identified (%)	Not identified (%)	Total
Begg	3 (18.8)	13 (81.2)	16
Egger	16 (37.2)	27 (62.8)	43
Deeks	1 (6.7)	14 (93.3)	15
Begg-Mazumdar	0	1 (100)	1
Harbord-Egger	0	1 (100)	1
All tests	20 (26.0)	56 (74.0)	76

Incorporation of results in the discussion

The results of investigation of publication bias were discussed in 25 out of 47 reviews that assessed publication bias. Six reviews based their conclusion about publication bias only on the plots, as they had not performed a test. One of these reviews concluded the existence of publication bias, two concluded no existence of publication and three were inconclusive about the influence of publication bias for their review. In reviews that had constructed a funnel plot and performed a test, the conclusions were based on the combination (funnel plot and test) or only on the test. In cases of disagreement between the results of a funnel plot and a test, all authors emphasized on the test results.

In fourteen reviews, the issue of publication bias was raised as a limitation to the results while five reviews concluded that there was no risk of publication bias. Two reviews discussed that the assessment had increased their confidence in the results of their review, though four reviews mentioned that it had affected the results and that these results should be considered cautiously.

Eleven reviews that did not assess publication bias mentioned that the possible existence of publication bias could be a limitation to the results of their review. In these reviews, authors stated that comprehensive searching, placing no limits on study quality or language could be used as precautions to prevent effects of publication bias. Two reviews also mentioned that excluding conference proceedings could have introduced publication bias.

Comparison of tests to detect publication bias

We were able to obtain two by two tables of 52 reviews, including 92 different meta-analyses. There was moderate concordance between the various tests for publication bias in terms of the presence or absence of significance (Figure

2, 3 and 4). Concordance of the Begg and Egger tests was significantly better depending on the number of included studies (OR 1.09; 95% CI 1.03 to 1.10). The number of included participants or a DOR >38 did not have a significant association with the concordance of tests (Table 3).

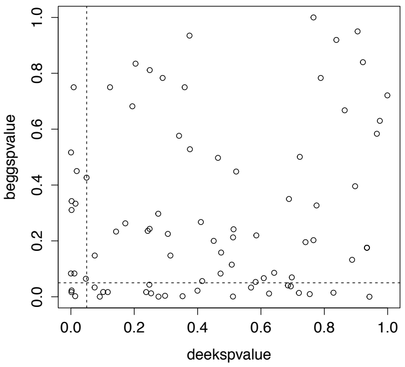


Figure 2. *Comparison of the p-values of the Begg test (y-axis) and Deeks' test (x-axis) in 92 meta-analyses. The dotted lines indicate a p-value of 0.05. Concordance between tests was 67% (k=-0.039; 95% CI -0.23 to 0.15).*

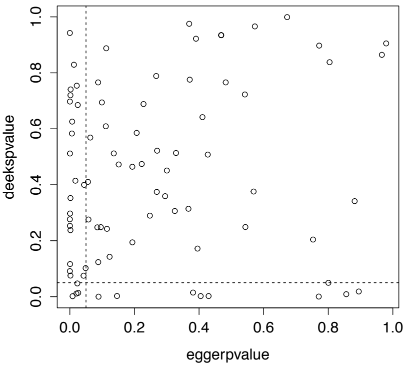


Figure 3. *Comparison of the p-values of the Egger test (y-axis) and Deeks' test (x-axis) in 92 meta-analyses. The dotted lines indicate a p-value of 0.05. Concordance between tests was 66% (k=-0.002; 95% CI -0.2 to 0.19).*

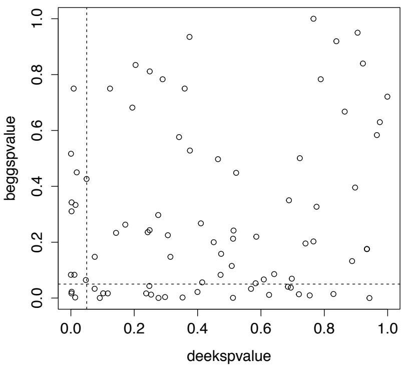


Figure 4. *Figure 4. Comparison of the p-values of the Begg test (y-axis) and the Egger test (x-axis) in 92 meta-analyses. The dotted lines indicate a p-value of 0.05. Concordance between tests was 87% between tests (k=0.68; 95% CI 0.51 to 0.86).*

Table 3. *Odd ratio's for the association between several factors and the concordance between tests*

Factor	Begg – Deeks OR (95% CI)	Egger –Deeks OR (95% CI)	Begg – Egger OR (95% CI)
Number of participants	1.00 (0.99 to 1.00)	1.00 (1.00 to 1.00)	1.00 (1.00 to 1.00)
Number of studies	0.96 (0.98 to 1.02)	1.00 (0.99 to 1.01)	1.09 (1.03 to 1.10)*
DOR > 38	1.02 (0.93 to 1.15)	0.955 (0.85 to 1.20)	0.999 (0.96 to 1.00)

**P-value <0.001*

DISCUSSION

Most authors of DTA reviews (65.8%) are concerned about publication bias. In 41.2% of the included reviews methods were applied to investigate publication bias. Funnel plots were constructed with a diversity of parameters on the axes and were sparsely used in isolation to formulate conclusions about the existence of publication bias. Forty-one reviews assessed publication bias with a statistical test. The Deeks test that is especially developed for reviews of diagnostic accuracy was only used in 12 reviews (10.5%). In 18 reviews (15.8%), the results of the publication bias assessment led to less confidence in the results. Our replication of three tests to detect publication bias (Begg, Egger and Deeks) using empirical data indicated that the results of the tests frequently conflict with one another. The study of Deeks et al. showed that a type 1 error is likely to occur in both the Begg and the Egger tests when the threshold for test positivity, the disease prevalence or the magnitude of the accuracy estimates varies between the included studies, especially when the DOR is high (DOR>38), which is present in almost every DTA review (19). Although, we cannot be sure in which reviews the test results were accurate and in which they were false, it seems likely that these two tests may have led to an overestimation of the presence of publication bias.

The number of reviews investigating publication bias seems to have increased over time. In 2002, Song and colleagues investigated how authors assessed publication bias in a sample of 20 reviews including 28 DTA meta-analyses. They concluded that none of the included reviews had investigated publication bias and that only 4 out of 20 reviews had considered its likelihood in the discussion (18). Furthermore, in 2011, Parekh-Bhurke et al. conducted a review to examine the approaches that are used to deal with publication bias in different types of systematic reviews published in 2006. They reported that only 26% of all reviews used statistical methods to assess publication bias (73). Of the 50 diagnostic reviews that were included in this study, nine (18%) used funnel plot asymmetry to investigate publications bias and in three (6%) a statistical test. These numbers are remarkably lower than found in our study. This could be the result of the increased awareness of the possible threat of publication bias in DTA reviews.

The increased awareness of publication bias is a positive development, but the drawback here is that the majority of review authors use tests that are not fit for DTA meta-analyses. Our evaluation of 92 meta-analyses indicated that both the Begg and Egger tests give more significant results than Deeks' test. This result is in line with the expectation based on the simulation study by Deeks et al. (19). The trim and fill method was used in two reviews only. This method removes the most extreme small studies on the side of the desired outcome direction in the funnel plot, and recomputes the effect size at each iteration until the plot is symmetrical (17). A recent simulation study in DTA meta-analyses showed that the trim and fill method was more powerful than other tests like the Begg, Egger or Deeks test to detect possible publication bias (74). Therefore, this method may be used more frequently in future.

Our study is limited by the fact that we based our results on what is reported in the publications. It is possible that funnel plots were constructed for more reviews but were not included in the publication. This may have led to an underestimation of the actual number of reviews that constructed a funnel plot. Secondly, our own assessment of publication bias in the meta-analyses is based on the data reported in the reviews but it is, of course, not clear if any of the meta-analyses were actually biased by publication bias as a gold standard is currently absent (14).

As correctly mentioned in some of the reviews included in our study, little is known about the actual existence of selective publication of DTA studies (75). There is no evidence regarding the existence of biases like language bias or time lag bias in the DTA setting, nor if these biases affect the accuracy measures in the same way as they affect the effect of interventions. It could be argued that depending on the purpose of the test either the sensitivity or the specificity are more affected by selective publication than the DOR, and tests for publication bias should perhaps be directed to these two accuracy parameters. A special situation of selective publication may occur with non-inferiority designs for diagnostic test accuracy. This study design aims to compare the diagnostic accuracy of a new diagnostic test with a standard test and is based on the difference in paired partial area under the ROC curve. This difference can be tested with Bayesian methods that result in a p-value (76;77).

Because of this p-value, this design may be more susceptible to non-publishing negative findings and as such induces publication bias. However, as long as the mechanisms behind publication bias of diagnostic studies are not well understood, it is understandable that some reviewers decided not to formally investigate how publication bias may have affected their meta-analysis.

Prospective registration of intervention studies shown to be an effective measure to reduce selective publication or at least make it more transparent to investigators. At the moment, prospective registration is advocated for diagnostic accuracy studies but not a prerequisite like it is for intervention studies in order to be considered for publication in journals associated with the International Committee of Medical Journal Editors (ICMJE) (78). Empirical studies to assess and understand the mechanisms that may induce publication bias in DTA studies, however, are needed. A cohort of prospective diagnostic studies could be followed and the dissemination of study results may be compared to the study characteristics and results. Optimization could be achieved if prospective registration of diagnostic accuracy studies would be mandatory. This may, however, would not be beneficial for all types of diagnostic studies. For example diagnostic data are often collected as part of daily clinical care and retrospectively analysed. Still, prospective registration of at least the prospective diagnostic studies could improve the understanding of the process of selective publication of DTA studies and identify underlying mechanisms. This knowledge is needed for valid interpretation of results of meta-analyses of diagnostic studies.

CONCLUSIONS

We found that most DTA reviewers struggle how to deal with publication bias in their reviews. Suboptimal tests like Egger's and Begg's are frequently used, while the interpretation of the test results are frequently not linked to the pooled results. Deeks' tests should be preferred to assess publication bias in DTA meta-analyses and interpretation of a significant test result should be done within the perspective that we are unaware whether publication bias exists for DTA studies. We advise authors of DTA reviews to try to avoid the introduction of publication bias and apply thorough methods for identifying primary studies, alongside regular searches in electronic biomedical databases. This entails identifying grey literature, contacting experts and searching for conference proceedings. Prospective registration of diagnostic studies with a prospective design could be helpful in the perspective of selective reporting.

COMPETING INTERESTS

This research project has not been funded. We have no competing interest to report that could have affected the results of our study.

AUTHORS' CONTRIBUTIONS

WE has contributed to the protocol of the study, data extraction, data-analysis and wrote manuscript. EO contributed to the protocol and performed the search and selection process of studies. She helped with data-checking and contributed to the manuscript. LH has contributed to the protocol of the study and contributed to the manuscript. RS has contributed to the protocol of the study, helped with data-checking and contributed to the manuscript. ML has contributed to the protocol of the study, performed the selection process of studies, helped with data-checking and contributed to the manuscript.

ACKNOWLEDGEMENTS

We would like to thank Fleur van de Wetering (FW) for her help with data checking and John Deeks for his suggestions on the methods. Further, we are grateful to Aeilko Zwinderman for his help to perform the analyses.

REFERENCE LIST

1. Dickersin K. The existence of publication bias and risk factors for its occurrence. JAMA 1990 Mar 9;263(10):1385-9.
2. Egger M, Juni P, Bartlett C, Holenstein F, Sterne J. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. Health Technol Assess 2003;7(1):1-76.
3. Ioannidis JP, Cappelleri JC, Sacks HS, Lau J. The relationship between study design, results, and reporting of randomized clinical trials of HIV infection. Control Clin Trials 1997 Oct;18(5):431-44.
4. 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.
5. Moher D, Fortin P, Jadad AR, Juni P, Klassen T, Le LJ, et al. Completeness of reporting of trials published in languages other than English: implications for conduct and reporting of systematic reviews. Lancet 1996 Feb 10;347(8998):363-6.
6. Sampson M, Platt R, StJohn PD, Moher D, Klassen TP, Pham B, et al. Should meta-analysts search Embase in addition to Medline? J Clin Epidemiol 2003;56:943-55.
7. Sterne JA, Gavaghan D, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. J Clin Epidemiol 2000 Nov;53(11):1119-29.
8. Thornton A, Lee P. Publication bias in meta-analysis: its causes and consequences. J Clin Epidemiol 2000 Feb;53(2):207-16.
9. Sterne JA, Sutton AJ, Ioannidis JP, Terrin N, Jones DR, Lau J, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ 2011;343:d4002.
10. Sterne JA, Egger M, Moher D. Adressing reporting bias; detecting repoting bias. In: Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions. Wiley-Blackwell; 2009. p. 310-24.
11. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. Biometrics 1994 Dec;50(4):1088-101.
12. Egger M, Davey SG, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997 Sep 13;315(7109):629-34.
13. Web of knowledge. Thomson Reuters, editor. 2014. New York, USA, Thomson Reuters. 23-1-2014.

14. Macaskill P, Gatsonis C, Deeks JJ, Harbord RM, Takwoingi Y. Analysing and Presenting Results. In: Deeks JJ, Bossuyt PM, Gatsonis C, editors. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. The Cochrane Collaboration; 2010. p. 46-7.

15. Rifai N, Altman DG, Bossuyt PM. Reporting bias in diagnostic and prognostic studies: time for action. Clin Chem 2008 Jul;54(7):1101-3.

16. Korevaar DA, Ochodo EA, Bossuyt PM, Hooft L. Publication and Reporting of Test Accuracy Studies Registered in ClinicalTrials.gov. Clin Chem 2014 Apr;60(4):651-9.

17. Brazzelli M, Lewis SC, Deeks JJ, Sandercock PA. No evidence of bias in the process of publication of diagnostic accuracy studies in stroke submitted as abstracts. J Clin Epidemiol 2009 Apr;62(4):425-30.

18. Song F, Khan KS, Dinnes J, Sutton AJ. Asymmetric funnel plots and publication bias in meta-analyses of diagnostic accuracy. Int J Epidemiol 2002 Feb;31(1):88-95.

19. 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.

20. Deville WL, Bezemer PD, Bouter LM. Publications on diagnostic test evaluation in family medicine journals: an optimal search strategy. J Clin Epidemiol 2000 Jan;53(1):65-9.

21. Korevaar DA, van Enst WA, Spijker R, Bossuyt PM, Hooft L. Reporting quality of diagnostic accuracy studies: a systematic review and meta-analysis of investigations on adherence to STARD. Evid Based Med 2014 Apr;19(2):47-54.

22. R Core Team. R: A language and environment for statistical computing. 2013. R Foundation for Statistical Computing.

23. Chang KC, Yew WW, Zhang Y. Pyrazinamide susceptibility testing in Mycobacterium tuberculosis: a systematic review with meta-analyses. Antimicrob Agents Chemother 2011;55(10):4499-505.

24. Chang MC, Chen JH, Liang JA, Lin CC, Yang KT, Cheng KY, et al. Meta-analysis: comparison of F-18 fluorodeoxyglucose-positron emission tomography and bone scintigraphy in the detection of bone metastasis in patients with lung cancer. Acad Radiol 2012 Mar;19(3):349-57.

25. Cheng X, Li Y, Xu Z, Bao L, Li D, Wang J. Comparison of 18F-FDG PET/CT with bone scintigraphy for detection of bone metastasis: a meta-analysis. Acta Radiol 2011;52(7):779-87.

26. Descatha A, Huard L, Aubert F, Barbato B, Gorand O, Chastang JF.

Meta-analysis on the performance of sonography for the diagnosis of carpal tunnel syndrome. Semin Arthritis Rheum 2012 Jun;41(6):914-22.

27. Dong MJ, Zhao K, Liu ZF, Wang GL, Yang SY, Zhou GJ. A meta-analysis of the value of fluorodeoxyglucose-PET/PET-CT in the evaluation of fever of unknown origin. Eur J Radiol 2011 Dec;80(3):834-44.

28. Dym RJ, Burns J, Freeman K, Lipton ML. Is functional MR imaging assessment of hemispheric language dominance as good as the Wada test?: a meta-analysis. Radiology 2011 Nov;261(2):446-55.

29. Gao P, Li M, Tian QB, Liu DW. Diagnostic performance of des-gamma-carboxy prothrombin (DCP) for hepatocellular carcinoma: a bivariate meta-analysis. Neoplasma 2012;59(2):150-9.

30. Gargiulo P, Petretta M, Bruzzese D, Cuocolo A, Prastaro M, D'Amore C, et al. Myocardial perfusion scintigraphy and echocardiography for detecting coronary artery disease in hypertensive patients: a meta-analysis. Eur J Nucl Med Mol Imaging 2011 Nov;38(11):2040-9.

31. Glasgow SC, Bleier JI, Burgart LJ, Finne CO, Lowry AC. Meta-analysis of histopathological features of primary colorectal cancers that predict lymph node metastases. J Gastrointest Surg 2012 May;16(5):1019-28.

32. Gong X, Xu Q, Xu Z, Xiong P, Yan W, Chen Y. Real-time elastography for the differentiation of benign and malignant breast lesions: a meta-analysis. Breast Cancer Res Treat 2011 Nov;130(1):11-8.

33. Hernaez R, Lazo M, Bonekamp S, Kamel I, Brancati FL, Guallar E, et al. Diagnostic accuracy and reliability of ultrasonography for the detection of fatty liver: a meta-analysis. Hepatology 2011 Sep 2;54(3):1082-90.

34. Inaba Y, Chen JA, Bergmann SR. Carotid plaque, compared with carotid intima-media thickness, more accurately predicts coronary artery disease events: a meta-analysis. Atherosclerosis 2012 Jan;220(1):128-33.

35. Kobayashi Y, Hayashino Y, Jackson JL, Takagaki N, Hinotsu S, Kawakami K. Diagnostic performance of chromoendoscopy and narrow band imaging for colonic neoplasms: a meta-analysis. Colorectal Dis 2012 Jan;14(1):18-28.

36. Li BS, Wang XY, Ma FL, Jiang B, Song XX, Xu AG. Is high resolution melting analysis (HRMA) accurate for detection of human disease-associated mutations? A meta analysis. PLoS One 2011;6(12):e28078.

37. Li R, Liu J, Xue H, Huang G. Diagnostic value of fecal tumor M2-pyruvate kinase for CRC screening: a systematic review and meta-analysis. Int J Cancer 2012 Oct 15;131(8):1837-45.

38. Lu Y, Chen YQ, Guo YL, Qin SM, Wu C, Wang K. Diagnosis of invasive fungal disease using serum (1-->3)-beta-D-glucan: a bivariate meta-analysis. Intern Med 2011;50(22):2783-91.

39. Lundstrom LH, Vester-Andersen M, Moller AM, Charuluxananan S, L'hermite J, Wetterslev J. Poor prognostic value of the modified Mallampati score: a meta-analysis involving 177 088 patients. Br J Anaesth 2011 Nov;107(5):659-67.

40. Luo YX, Chen DK, Song SX, Wang L, Wang JP. Aberrant methylation of genes in stool samples as diagnostic biomarkers for colorectal cancer or adenomas: a meta-analysis. Int J Clin Pract 2011 Dec;65(12):1313-20.

41. Manea L, Gilbody S, McMillan D. Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis. CMAJ 2012 Feb 21;184(3):E191-E196.

42. Mao R, Xiao YL, Gao X, Chen BL, He Y, Yang L, et al. Fecal calprotectin in predicting relapse of inflammatory bowel diseases: a meta-analysis of prospective studies. Inflamm Bowel Dis 2012 Oct;18(10):1894-9.

43. Marton A, Xue X, Szilagyi A. Meta-analysis: the diagnostic accuracy of lactose breath hydrogen or lactose tolerance tests for predicting the North European lactase polymorphism C/T-13910. Aliment Pharmacol Ther 2012 Feb;35(4):429-40.

44. Mathews WC, Agmas W, Cachay E. Comparative accuracy of anal and cervical cytology in screening for moderate to severe dysplasia by magnification guided punch biopsy: a meta-analysis. PLoS One 2011;6(9):e24946.

45. McInnes MD, Kielar AZ, Macdonald DB. Percutaneous image-guided biopsy of the spleen: systematic review and meta-analysis of the complication rate and diagnostic accuracy. Radiology 2011 Sep;260(3):699-708.

46. Meader N, Mitchell AJ, Chew-Graham C, Goldberg D, Rizzo M, Bird V, et al. Case identification of depression in patients with chronic physical health problems: a diagnostic accuracy meta-analysis of 113 studies. Br J Gen Pract 2011 Dec;61(593):e808-e820.

47. Mitchell AJ, Meader N, Pentzek M. Clinical recognition of dementia and cognitive impairment in primary care: a meta-analysis of physician accuracy. Acta Psychiatr Scand 2011 Sep;124(3):165-83.

48. Onishi A, Sugiyama D, Kogata Y, Saegusa J, Sugimoto T, Kawano S, et al. Diagnostic accuracy of serum 1,3-beta-D-glucan for pneumocystis jiroveci pneumonia, invasive candidiasis, and invasive aspergillosis: systematic review and meta-analysis. J Clin Microbiol 2012 Jan;50(1):7-15.

49. Papathanasiou ND, Boutsiadis A, Dickson J, Bomanji JB. Diagnostic accuracy of (1)(2)(3)I-FP-CIT (DaTSCAN) in dementia with Lewy bodies: a meta-analysis of published studies. Parkinsonism Relat Disord 2012 Mar;18(3):225-9.

50. Plana MN, Carreira C, Muriel A, Chiva M, Abraira V, Emparanza JI, et al. Magnetic resonance imaging in the preoperative assessment of patients with primary breast cancer: systematic review of diagnostic accuracy and meta-analysis. Eur Radiol 2012 Jan;22(1):26-38.

51. Qu X, Huang X, Wu L, Huang G, Ping X, Yan W. Comparison of virtual cystoscopy and ultrasonography for bladder cancer detection: a meta-analysis. Eur J Radiol 2011 Nov;80(2):188-97.

52. Sadeghi R, Gholami H, Zakavi SR, Kakhki VR, Tabasi KT, Horenblas S. Accuracy of sentinel lymph node biopsy for inguinal lymph node staging of penile squamous cell carcinoma: systematic review and meta-analysis of the literature. J Urol 2012 Jan;187(1):25-31.

53. Sadigh G, Carlos RC, Neal CH, Dwamena BA. Ultrasonographic differentiation of malignant from benign breast lesions: a meta-analytic comparison of elasticity and BIRADS scoring. Breast Cancer Res Treat 2012 May;133(1):23-35.

54. Summah H, Tao LL, Zhu YG, Jiang HN, Qu JM. Pleural fluid soluble triggering receptor expressed on myeloid cells-1 as a marker of bacterial infection: a meta-analysis. BMC Infect Dis 2011;11:280.

55. Sun W, Wang K, Gao W, Su X, Qian Q, Lu X, et al. Evaluation of PCR on bronchoalveolar lavage fluid for diagnosis of invasive aspergillosis: a bivariate metaanalysis and systematic review. PLoS One 2011;6(12):e28467.

56. Takakuwa KM, Keith SW, Estepa AT, Shofer FS. A meta-analysis of 64-section coronary CT angiography findings for predicting 30-day major adverse cardiac events in patients presenting with symptoms suggestive of acute coronary syndrome. Acad Radiol 2011 Dec;18(12):1522-8.

57. Thosani N, Singh H, Kapadia A, Ochi N, Lee JH, Ajani J, et al. Diagnostic accuracy of EUS in differentiating mucosal versus submucosal invasion of superficial esophageal cancers: a systematic review and meta-analysis. Gastrointest Endosc 2012 Feb;75(2):242-53.

58. Tomasson G, Grayson PC, Mahr AD, Lavalley M, Merkel PA. Value of ANCA measurements during remission to predict a relapse of ANCA-associated vasculitis--a meta-analysis. Rheumatology (Oxford) 2012 Jan;51(1):100-9.

59. Trallero-Araguas E, Rodrigo-Pendas JA, Selva-O'Callaghan A, Martinez-Gomez X, Bosch X, Labrador-Horrillo M, et al. Usefulness of anti-p155 autoantibody for diagnosing cancer-associated dermatomyositis: a systematic review and meta-analysis. Arthritis Rheum 2012 Feb;64(2):523-32.

60. Wang GL, Zhao K, Liu ZF, Dong MJ, Yang SY. A meta-analysis of fluorodeoxyglucose-positron emission tomography versus scintigraphy in the evaluation of suspected osteomyelitis. Nucl Med Commun 2011 Dec;32(12):1134-42.

61. Wang QB, Zhu H, Liu HL, Zhang B. Performance of magnetic resonance elastography and diffusion-weighted imaging for the staging of hepatic fibrosis: A meta-analysis. Hepatology 2012 Jul;56(1):239-47.

62. Wang W, Li Y, Li H, Xing Y, Qu G, Dai J, et al. Immunodiagnostic efficacy of detection of Schistosoma japonicum human infections in China: a meta analysis. Asian Pac J Trop Med 2012 Jan;5(1):15-23.

63. Wang Y, Zhang C, Liu J, Huang G. Is 18F-FDG PET accurate to predict neoadjuvant therapy response in breast cancer? A meta-analysis. Breast Cancer Res Treat 2012 Jan;131(2):357-69.

64. Wu LM, Xu JR, Liu MJ, Zhang XF, Hua J, Zheng J, et al. Value of magnetic resonance imaging for nodal staging in patients with head and neck squamous cell carcinoma: a meta-analysis. Acad Radiol 2012 Mar;19(3):331-40.

65. Xu HB, Li L, Xu Q. Tc-99m sestamibi scintimammography for the diagnosis of breast cancer: meta-analysis and meta-regression. Nucl Med Commun 2011 Nov;32(11):980-8.

66. Xu W, Shi J, Zeng X, Li X, Xie WF, Guo J, et al. EUS elastography for the differentiation of benign and malignant lymph nodes: a meta-analysis. Gastrointest Endosc 2011 Nov;74(5):1001-9.

67. Ying L, Hou Y, Zheng HM, Lin X, Xie ZL, Hu YP. Real-time elastography for the differentiation of benign and malignant superficial lymph nodes: a meta-analysis. Eur J Radiol 2012 Oct;81(10):2576-84.

68. Yu YH, Wei W, Liu JL. Diagnostic value of fine-needle aspiration biopsy for breast mass: a systematic review and meta-analysis. BMC Cancer 2012;12:41.

69. Zhang L, Zong ZY, Liu YB, Ye H, Lv XJ. PCR versus serology for diagnosing Mycoplasma pneumoniae infection: a systematic review & meta-analysis. Indian J Med Res 2011 Sep;134:270-80.

70. Harbord RM, Egger M, Sterne JA. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. Stat Med 2006

Oct 30;25(20):3443-57.

71. Hazem A, Elamin MB, Malaga G, Bancos I, Prevost Y, Zeballos-Palacios C, et al. The accuracy of diagnostic tests for GH deficiency in adults: a systematic review and meta-analysis. Eur J Endocrinol 2011 Dec;165(6):841-9.

72. Singh B, Parsaik AK, Agarwal D, Surana A, Mascarenhas SS, Chandra S. Diagnostic accuracy of pulmonary embolism rule-out criteria: a systematic review and meta-analysis. Ann Emerg Med 2012 Jun;59(6):517-20.

73. Parekh-Bhurke S, Kwok CS, Pang C, Hooper L, Loke YK, Ryder JJ, et al. Uptake of methods to deal with publication bias in systematic reviews has increased over time, but there is still much scope for improvement. J Clin Epidemiol 2011 Apr;64(4):349-57.

74. Burkner PC, Doebler P. Testing for publication bias in diagnostic meta-analysis: a simulation study. Stat Med 2014 Apr 20.

75. de Vet HCW, Eisinga A, Riphagen II, Aertgeerts B, Pewsner D. Searching for Studies. In: The Cochrane Collaboration, editor. Cochrane Handbook for Systematic Reviews of Diagnosic Test Accuracy. 0.4 ed. 2008.

76. Li CR, Liao CT, Liu JP. A non-inferiority test for diagnostic accuracy based on the paired partial areas under ROC curves. Stat Med 2008 May 10;27(10):1762-76.

77. Liu JP, Ma MC, Wu CY, Tai JY. Tests of equivalence and non-inferiority for diagnostic accuracy based on the paired areas under ROC curves. Stat Med 2006 Apr 15;25(7):1219-38.

78. DeAngelis CD, Drazen JM, Frizelle FA, Haug C, Hoey J, Horton R, et al. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. Ann Intern Med 2004 Sep 21;141(6):477-8.

8

CHAPTER

Small study and time lag effects in diagnostic test accuracy

Wynanda A. van Enst
Christiana A. Naaktgeboren
Eleanor E. Ochodo
Joris A.H. de Groot
Mariska M. Leeflang
Johannes B. Reitsma
Karel G.M. Moons
Aeilko Zwinderman
Patrick M. Bossuyt
Lotty Hooft

Submitted to Systematic Reviews Journal, June 2014

ABSTRACT

Background Small study and time lag effects have been identified in meta-analyses of randomized trials. We evaluated whether these effects are also present in meta-analyses of diagnostic test accuracy studies.

Methods A systematic search identified test accuracy meta-analyses published between May and September 2012. Two-by-two accuracy tables from the primary studies and the publication year were extracted for included reviews. In each meta-analysis the strength of the associations between estimated accuracy of the test and sample size as well as between estimated accuracy and time since first publication within each meta-analysis were evaluated using weighted linear regression models. The regression coefficients over all meta-analyses were summarized using random effects meta-analysis.

Results Fifty meta-analyses and their corresponding primary studies (n=874) were included. There was a positive association between accuracy (Diagnostic odds ratio (DOR), sensitivity and specificity) and sample size, with larger studies reporting higher accuracy. A time effect was only observed for the DOR, which was significantly lower in the quartile of most recently published studies.

Discussion/Conclusion Small study and time lag effects do not seem to be as pronounced in meta-analyses of test accuracy studies as they are in meta-analyses of randomized trials.

INTRODUCTION

The validity and credibility of the results of a systematic review of diagnostic test accuracy studies depends on the methodological quality of the included studies, but also on the absence of selective reporting (1-3). Knowledge about the principles of selective reporting can help with the interpretation of the results of a meta-analysis.

A sample size effect in randomized trials has been described before. Published trials with smaller sample sizes tend to have larger and more favourable effects compared to studies with larger sample sizes (4;5). This phenomenon may occur for several reasons. 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 their estimated effect. This mechanism, which is called small study effect, can hamper the validity of a systematic review overestimating the "true" effect (3;6-8).

In addition to a small study effect, meta-analyses of randomized trials may also be influenced by the problems arising from a time lag effect. This effect can result from variability in the time it takes to complete and publish a study report, which may depend on the direction and strength of the trial results (9). Empirical studies have indicated that negative or null results take approximately two or three more years to be published compared to positive results (3;10). This time lag effect could influence the meta-analysis, especially when it includes a small number of studies. It therefore has implications for the timing of a review, inclusion of on-going studies, and updating the review.

Whereas these effects are well known and described for randomized trails, it is unclear whether phenomena such as small study or time lag effects translate to diagnostic studies (11-13). Publication of diagnostic studies may be influenced by a different set of factors than randomized trials. In general, test accuracy studies tend to rely less on statistical significance testing than randomized trials. Many studies do not report confidence intervals around estimates (14), and sample size calculations based on a desirable outcome are typically absent (15). However, there is some evidence of a failure to publish completed research projects. Korevaar et al. compared registered test accuracy studies to the reported publication and concluded failure to publish and selective reporting is also present in test accuracy studies (16). However, the mechanisms and possible explanations driving this publication bias of test accuracy studies are not known.

In this study we aimed to assess whether meta-analyses of diagnostic tests accuracy measures suffer from small study or time lag effects, using a set of recent meta-analyses of diagnostic test accuracy studies.

METHODS

Overarching project

This study was a part of a meta-epidemiologic project on systematic reviews of diagnostic studies. The goal of this project was to investigate several methodological topics such as small sample size effects, time lag bias, quality assessment, and how to interpret tests and measurements of heterogeneity.

Selection of reviews and meta-analyses

This study was part of a meta-epidemiological project on systematic reviews of diagnostic accuracy studies. On September 12th 2012, MEDLINE and EMBASE were searched for systematic reviews on test accuracy studies published between May 1st 2012 and September 11th 2012. For our analysis, we limited inclusion to reviews with a meta-analysis for which we were able to obtain two-by-two classification tables of the studies included in the meta-analysis. A meta-analysis was defined as an analysis producing a summary estimate for at least one accuracy statistic or, alternatively, producing a summary ROC curve (sROC). Reviews of tests in animals, prognostic tests, and of individual patient data were excluded, as there may be other effects related to publication in these types of studies. Only English language reviews were included. The search strategy is available in Appendix 1.

Data extraction

Data were extracted using an online structured data extraction form. An independent double data extraction pilot was performed for a subset of the reviews (30%) until all authors agreed on the items of the data-extraction form. After that, data were extracted by one reviewer (CN, EO or WvE) and checked by a second reviewer (CN, EO or WvE) for discrepancies. Disagreements were resolved during a consensus meeting.

For each eligible review, we classified the type of test under evaluation and the total number of studies included in the meta-analyses. Data were then collected on the primary study level for one meta-analysis for each included review. If there was more than one meta-analyses in the published review, we selected the one with the largest number of included primary studies. For each primary study in a meta-analysis we extracted the year of publication and data to populate the individual two-by-two accuracy table (i.e. number of true positives, false negatives, false positives, and true negatives).

Whenever information on the primary studies was not available to us directly from the published review, we contacted the authors of the review. When we were unable to reach the author after sending two reminders or when authors could not provide the data, data were extracted from the original primary study reports. A second author checked the results of the data extraction.

Data analysis

We evaluated the strength of the association between the estimated accuracy and sample size over all studies within each included meta-analysis separately. We performed similar analyses for the association between estimated accuracy and time since publication of the first study within each review.

The diagnostic odds ratio was chosen as the accuracy statistic of primary interest because it expresses accuracy as a single parameter (13;17;18). Secondary outcomes were the effects on sensitivity and specificity. To facilitate analyses, we used the natural logarithm of the DOR (lnDOR) and evaluated sensitivity and specificity on the logit scale. We added 0.05 to all the cells in the two-by-two tables to facilitate the analysis.

A weighted linear regression model was fitted to the studies in each meta-analysis, with the lnDOR of a study as the dependent variable and the study sample size as the independent variable. We selected the empirical Bayes model proposed for multiple linear regression for its ability to fit smaller samples (19). A similar model was built using time between the date of publication of each study and the date of the oldest publication in the meta-analysis as the independent variable.

The association between sample size and the lnDOR was also studied using the inverse of the effective sample size (ESS) as the independent variable. The ESS is a function of the number of diseased (n1) and non-diseased (n2) participants and can be calculated using the following formula: (4n1*n2)/(n1+n2). The ESS takes into account the fact that unequal numbers of diseased and non-diseased reduce the precision of test accuracy estimates for the total sample (17).

In evaluating associations between sample size and sensitivity and specificity estimates, we took the number of diseased and the number of non-diseased as the respective independent variables. In addition, we classified studies in each meta-analysis into four groups using quartiles of sample size and quartiles of time elapsed since the first publication in years, respectively, and used the quartile as an ordinal variable in the regression.

After fitting a regression equation for each included meta-analysis, the resulting regression coefficients and their precision were combined using DerSimonian and Laird's random effects model to estimate the overall association (20). This two-step approach was chosen to accommodate differences in accuracy between meta-analyses related to differences in tests and fields. All analyses were conducted in the statistical package R (21).

Subgroup analysis

Separate analyses were carried out for imaging tests and for laboratory tests. Our rationale for this subgroup analysis was based on the observation that imaging studies generally have an implicit threshold.

The reported accuracy in studies with an implicit threshold can be affected by the number of diseased patients and is more likely to change over time (22-24). In addition, with imaging, gradual improvements in techniques may also induce time trends. We therefore hypothesized that a small study or time effect might act differently in imaging studies than in laboratory studies.

RESULTS

Search results

The search identified 1,273 references. After screening the titles and abstracts 89 references were found potentially eligible and were read as full text articles. Attempts were made to obtain the two-by-two tables of 53 eligible reviews. In three reviews attempts were unsuccessful resulting in 50 reviews that were eventually included (see flow chart in Figure 1 and Additional file for references). The 50 meta-analyses combined contained a total of 874 primary studies.

Characteristics of the included reviews and meta-analyses

Fifteen reviews investigated a laboratory test, twenty-nine an imaging test and six addressed clinical examinations. The selected meta-analyses had a median of ten studies (interquartile range (IRQ) 5 - 21). The median prevalence of the target condition in the studies was 48% (IQR: 24% – 69%). More characteristics of the primary studies are presented in Table 1.

Table 1. *Characteristics of primary studies (N=874) in the included meta-analyses (N=50)*

Number of Sample Size	Median	Median Interquartile range	Range
Sample size	87	45 – 183	3 – 50,008
Effective Sample size†	56	31 – 110	0 – 3,040
Number of diseased	32	16 – 63	0 – 1,358
Number of non-diseased	36	18 – 100	0 – 49,973
Time lag (years) ‡	6	3 – 10	0 – 42

† *Effective sample size: (4n1*n2)/(n1+n2)*
‡*Time lag: time since the first publication within a meta-analysis*

Sample size

The median sample size of the included studies (n=874) was 87 participants (IQR 45 – 183), ranging from extremely small to very large (range: 3 to 50,008). In total, there were 52,178 diseased participants and 526,627 non-diseased. This skewed distribution was mainly caused by a small set of studies on screening tests with very large samples but very few diseased compared to non-diseased.

The summary regression coefficient for the association between sample size and DOR was 1.01 (95% CI 1.00 to 1.03). This indicates that, on average, larger studies produced significantly larger estimates of test accuracy. One meta-analysis was excluded from the analyses because it only included three primary studies, and fitting a regression model for this meta-analysis was not considered meaningful.

Enlarging the contrast between small and large sample sized studies by comparing quartiles indicated that studies in the fourth quartile (25% of studies with largest sample size) on average had a 1.45 higher DOR than studies with a sample size in the first quartile (95% CI 0.91 to 2.18). For the analysis with quartiles the model had to fit 4 variables to allow for different effects per quartile, meaning that 5 primary studies needed to be present to fit the analysis. This was possible for 42 meta-analyses.

When associations with sample size were studied using effective sample size as the independent variable, the regression coefficient of the DOR was 1.01 (95% CI 0.78 to 1.30). A comparison of the fourth quartile to the first quartile indicated that studies with an ESS in the fourth quartile had on average a 1.36 higher DOR compared to the studies in the first quartile (95% CI 0.85 to 2.17). The analysis for sensitivity and specificity revealed a similar pattern: studies with a higher number of evaluated study participants tended to report higher accuracy estimates for both sensitivity and specificity (Table 2).

Publication date

The primary studies included in the meta-analyses were published between 1969 and 2010. Within meta-analyses, the median time interval since the first included publication was 6 years (IQR: 3 – 10). There was no association between the sample size (or the ESS) and the time since first publication (change over time). The DOR of the studies in the quartile with the most recent published studies was significantly lower than for studies in the earliest studies (0.73; 95% CI 0.58 to 0.92). There were no other significant associations between time since first publication and the DOR, sensitivity or specificity (Table 2).

Subgroup analysis

None of the associations were significantly different between the subgroups. The regression coefficients for the associations had similar directions except for specificity. The OR for specificity decreased for imaging tools over time, while it seemed to improve for laboratory tests, but this difference was not significant (Table 3).

Table 2. *Small study effect and time lag effects assessed continuous and per quartile*[a]

	Accuracy measure	Relative increase per 100 participants	Q4 vs. Q3 (95% CI)	Q4 vs. Q2 (95% CI)	Q4 vs. Q1 (95% CI)
Sample Size	DOR§	1.01* (1.00 to 1.03)	1.23 (0.95 to 1.60)	1.15 (0.86 to 1.54)	1.45 (0.99 to 2.1)
Effective Sample Size†	DOR§	1.01 (0.78 to 1.30)	1.16 (0.88 to 1.53)	1.12 (0.81 to 1.55)	1.36 (0.85 to 2.17)
Number of diseased	Sensitivity	1.23* (1.09 to 1.39)	1.25* (1.02 to 1.53)	1.33* (1.07 to 1.66)	1.62* (1.30 to 2.04)
Number of non-diseased	Specificity	1.01 (0.99 to 1.05)	1.12 (0.92 to 1.37)	1.15 (0.91 to 1.46)	1.47* (1.10 to 1.96)
Time lag (years) ‡	DOR§	0.99 (0.95 to 1.03)	0.87 (0.56 to 1.35)	0.85 (0.64 to 1.12)	0.73* (0.58 to 0.92)
Time lag (years) ‡	Sensitivity	0.99 (0.96 to 1.02)	0.84 (0.66 to 1.08)	0.92 (0.76 to 1.12)	0.81 (0.62 to 1.06)
Time lag (years) ‡	Specificity	1.01 (0.98 to 1.04)	0.98 (0.76 to 1.26)	0.86 (0.68 to 1.10)	0.90 (0.74 to 1.10)

[a] *To facilitate analyses, we analysed the natural logarithm of the DOR (lnDOR) and evaluated sensitivity and specificity on the logit scale.*

§ *DOR: Diagnostic odds ratio*

† *Effective sample size: $(4n_1.n_2)/(n_1+n_2)$*

‡*Time lag: time since the first publication within a meta-analysis*

* *p-value < 0.05*

Table 3. *Small study and time lag effects assessed in subgroups of imaging and laboratory tests* [α]

	Accuracy measure	Imaging test *Q4 vs. Q1* *(95% CI)*	Laboratory test *Q4 vs. Q1* *(95% CI)*
Sample size	DOR[§]	1.46 (0.87 to 2.45)	1.61 (0.73 to 3.58)
Effective sample size[†]	DOR[§]	1.72 (0.85 to 3.49)	1.61 (0.73 to 3.58)
Diseased	Sensitivity	1.96[*] (1.56 to 2.47)	1.36 (0.86 to 2.16)
Non-diseased	Specificity	1.32 (0.90 to 1.91)	1.51 (0.86 to 2.64)
Time lag (years) [‡]	DOR[§]	0.70[*] (0.52 to 0.93)	0.83 (0.53 to 1.29)
Time lag (years) [‡]	Sensitivity	0.93 (0.66 to 1.30)	0.82 (0.48 to 1.40)
Time lag (years) [‡]	Specificity	0.91 (0.73 to 1.13)	1.07 (0.73 to 1.57)

[α] *To facilitate analyses, we analysed the natural logarithm of the DOR (lnDOR) and evaluated sensitivity and specificity on the logit scale.*

[§] *DOR: Diagnostic odds ratio*

[†] *Effective sample size: (4n1*n2)/(n1+n2)*

[‡] *Time lag: time since the first publication within a meta-analysis*

[*] *p-value < 0.05*

Sensitivity analysis

We observed some very small absolute numbers of diseased participants in the included studies: 118 studies had ten or less diseased participants and 121 studies had ten or less non-diseased participants. In very small studies, the possible values for the estimated accuracy are small. Small studies may easily underestimate the true accuracy when sensitivity and specificity are very high or low. For example, when accuracy is acquired from four diseased patients, the sensitivity could only be estimated as 0%, 25%, 50%, 75%, or 100%. When the true sensitivity would be 95%, sensitivity may easily be underestimated. This phenomenon in itself might be responsible for a small study effect (25). We

Table 4. *Sensitivity analysis of small study effects excluding all studies with n < 10 diseased or non-diseased* [α]

	Accuracy measure	Relative increase per 100 participants (95% CI)	Q4 vs Q1 (95% CI)
Sample Size	DOR[§]	1.01 (0.99 to 1.03)	1.29 (0.90 to 1.8)
Effective Sample Size[†]	DOR[§]	1.00 (0.93 to 1.08)	1.08 (0.76 to 1.55)
Number of diseased	Sensitivity	1.21[*] (1.09 to 1.33)	1.70[*] (1.26 to 2.31)
Number of non-diseased	Specificity	1.01 (0.99 to 1.03)	1.19 (0.92 to 1.57)

[α] *To facilitate analyses, we analysed the natural logarithm of the DOR (lnDOR) and evaluated sensitivity and specificity on the logit scale. Q4 was the quartile with 25% of studies with largest sample size or with 25% most recent published studies within each meta-analysis*

[§] *DOR: Diagnostic odds ratio*

[†] *Effective sample size: (4n1*n2)/(n1+n2)*

[*] *p-value < 0.05*

therefore decided to run additional sensitivity analysis, excluding all studies with less than ten diseased and those with less than ten non-diseased participants. In this sensitivity analysis regression coefficients were typically smaller, and no significant study effect could be observed, except for sensitivity.

DISCUSSION

We assessed the existence of small study and time lag effects in test accuracy meta-analyses using a meta-epidemiological analysis of a series of published systematic reviews. Opposite to what was expected, we observed that accuracy estimates of diagnostic studies with a small sample size tended to be lower than for studies with a larger sample size. The association was significant for various accuracy measures, but some of this may be an artefact of the very small studies, i.e. those with less than ten diseased patients or less than ten non-diseased. Furthermore, we found limited evidence for the existence of time lag effects. The association was only significant for the DOR when we compared the most extreme contrast of the 25% most recently published studies to the 25% first published studies. We did not observe different effects between imaging and laboratory tests.

The findings of this study are in contrast with earlier findings for meta-analyses of randomized trials, where higher treatment effect sizes of RCTs are strongly associated with small sample sizes (7;8;26). Nüesch et al. studied 13 meta-analyses with continuous outcomes and found on average 0.21 (95% CI 0.08 to 0.34) higher effect sizes in small trials than in large trials (7). Dechartres et al. included 93 meta-analyses with binary outcomes. They concluded that the quartile of smallest trials had 32% (95% CI 18% to 43%) larger treatment effects than the quartile that included the largest trials (8). The differences in effect size between small and large trials can be the result of the publication process, which elects positive and significant results over negative or null results (27). According to the review of Hopewell et al. the odds to find positive, significant results in a publication are four times higher than to find negative or null results(28).

We consider it very unlikely that the sample size effect we have found for DTA meta-analyses is the result of an actual preference to publish small studies with low accuracy measures rather than small studies with higher accuracy measures. The sensitivity analysis showed that an artefact caused by the very small studies might explain the sample size effect. The choice to exclude studies with less than ten diseased or non-diseased participants was arbitrary. In the sensitivity analysis, the positive relation between the number of diseased and sensitivity remained statistically significant. This might be an indication that our cut-off point of ten diseased in the sensitivity analysis was too conservative. Another factor that could have led to the large study effect is variability of methodological quality. For diagnostic research, large sampled studies often come from routine care data. Such data often suffer from verification problems, resulting in higher accuracy (29). So, first, presence of the small study effect calls for caution when including studies with a very small number of diseased or non-diseased participants in a meta-analysis. It would be worthwhile to investigate the minimal number of needed diseased or non-diseased patients. Second, further evaluation is needed if methodological quality if related to sample size.

Our findings on sample size effect were confirmed by the study of Haines and colleagues. They found a similar relation between sample size and the Youden's Index, a test statistic that captures test performance (30). Studies with larger samples had a higher Youden's Index. They claimed that this relationship was attributable to prematurely ceasing studies with poorer outcomes at smaller sample sizes. It will be challenging to assess if this hypothesis is valid because power calculations that specify the desired power at baseline of a study, are rarely reported in DTA-studies(31).

The time lag effect observed in our DTA meta-analyses was much smaller than identified for randomized trials (3;9;32). For example, the systematic review of Hopewell et al. indicated that the median time to publish significant results was 4.7 years, while this was 8.0 years for studies with negative or null results (10). Our evaluation does not indicate such a strong relationship between the time to publish and the outcome of the studies. None of the trends were significant over time, except for the DOR comparing the 25% most recent published studies to the 25% first published studies. The direction of the trend was similar to the trend of randomized trials, with lower DORs in later studies. Similar to our results, the study of Sonnad et al. found that earlier published studies had higher accuracy, but the relation was not significant (33).

Even in the absence of an overall effect, it is still possible that a time lag effect exist for specific tests. For example, the design of studies may change over time, from explorative case control type studies to prospective studies in consecutive patients (34). In addition, the setting and targeted patients may change over time, with better understanding of the most useful application of a diagnostic test (35). It would be worthwhile to study if specific study characteristics, such as study setting or patient spectrum, change over time in a large cohort of primary diagnostic accuracy studies.

Both small study and time lag effects are, among other reasons, consequences of publication bias. The meticulous follow-up of a cohort of diagnostic accuracy studies could be a way of documenting the actual mechanisms in the reporting and publication processes of such studies, and allows to analyse to what extent non-random publication bias exists (9;36).

Factors that influence the decision to submit or to accept a research article can also be studied from trial registers and present more direct information on publication bias. In 2006 the International Committee of Journal Editors (ICMJE) established prospective registration of trials, defined as "any research project that prospectively assigns human subjects to intervention and comparison groups to study the cause-and-effect relationship between a medical intervention and a health outcome" (37). At present, this definition does not seem to capture all test accuracy studies, and recent analyses have shown that only a small subset of such studies is currently registered before enrolment of the first patient (38).

CONCLUSION

Awaiting further evidence, our study results leads us to conclude that some of the typical mechanisms associated with publication bias which are well documented in the literature for randomized clinical trials are less prominent in test accuracy research. Delays in the reporting of studies with disappointing results and failures to report such studies at all if they are small, might not be as common as in randomized clinical trials of pharmaceuticals and other interventions. Confirmation of the findings of our study may provide reassurance to those relying on the published literature for evidence of the performance of medical tests.

AUTHORS' CONTRIBUTIONS

All authors have contributed to development of the protocol. CN, EO and WvE have performed study selection and data extraction. AZ and WvE developed and performed the analyses. In addition, PB and WvE performed the sensitivity analysis. All authors have contributed to the manuscript.

REFERENCE LIST

1. Dickersin K. The existence of publication bias and risk factors for its occurrence. JAMA 1990 Mar 9;263(10):1385-9.

2. Simes RJ. Publication bias: the case for an international registry of clinical trials. J Clin Oncol 1986 Oct;4(10):1529-41.

3. Song F, Parekh S, Hooper L, Loke YK, Ryder J, Sutton AJ, et al. Dissemination and publication of research findings: an updated review of related biases. Health Technol Assess 2010 Feb;14(8):iii, ix-iii,193.

4. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. Biometrics 1994 Dec;50(4):1088-101.

5. Egger M, Davey SG, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997 Sep 13;315(7109):629-34.

6. Kjaergard LL, Villumsen J, Gluud C. Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. Ann Intern Med 2001 Dec 4;135(11):982-9.

7. Nuesch E, Trelle S, Reichenbach S, Rutjes AW, Tschannen B, Altman DG, et al. Small study effects in meta-analyses of osteoarthritis trials: meta-epidemiological study. BMJ 2010;341:c3515.

8. Dechartres A, Trinquart L, Boutron I, Ravaud P. Influence of trial sample size on treatment effect estimates: meta-epidemiological study. BMJ 2013;346:f2304.

9. 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.

10. Hopewell S, Clarke M, Stewart L, Tierney J. Time to publication for results of clinical trials. Cochrane Database Syst Rev 2007;(2):MR000011.

11. Centre for Reviews and Dissemination. Clinical Tests. Systematic Reviews: CRD's guidance for undertaking reviews in health care. 2013.

12. Deeks JJ. Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests. BMJ 2001 Jul 21;323(7305):157-62.

13. Macaskill P, Gatsonis C, Deeks JJ, Harbord RM, Takwoingi Y. Analysing and Presenting Results. In: Deeks JJ, Bossuyt PM, Gatsonis C, editors. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. The Cochrane Collaboration; 2010. p. 46-7.

14 Korevaar DA, van Enst WA, Spijker R, Bossuyt PM, Hooft L. Reporting quality of diagnostic accuracy studies: a systematic review and meta-analysis of investigations on adherence to STARD. Evid Based Med 2013 Dec 24.

15. Bachmann LM, Puhan MA, ter RG, Bossuyt PM. Sample sizes of studies on diagnostic accuracy: literature survey. BMJ 2006 May 13;332(7550):1127-9.
16. Korevaar DA, Ochodo EA, Bossuyt PM, Hooft L. Publication and Reporting of Test Accuracy Studies Registered in ClinicalTrials.gov. Clin Chem 2013 Dec 31.
17. 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.
18. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM. The diagnostic odds ratio: a single indicator of test performance. J Clin Epidemiol 2003 Nov;56(11):1129-35.
19. Morris CN. Parametric empirical bayes inference - theory and applications. Journal of the American Statistical Association 1983;78:47-55.
20. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986 Sep;7(3):177-88.
21. R Core Team. R: A language and environment for statistical computing. 2013. R Foundation for Statistical Computing.
22. Willis BH. Empirical evidence that disease prevalence may affect the performance of diagnostic tests with an implicit threshold: a cross-sectional study. BMJ Open 2012;2(1):e000746.
23. Dias-Silva D, Pimentel-Nunes P, Magalhaes J, Magalhaes R, Veloso N, Ferreira C, et al. The learning curve for narrow-band imaging in the diagnosis of precancerous gastric lesions by using Web-based video. Gastrointest Endosc 2013 Nov 25.
24. Kheir F, Alokla K, Myers L, Palomino J. Endobronchial Ultrasound-Transbronchial Needle Aspiration of Mediastinal and Hilar Lymphadenopathy Learning Curve. Am J Ther 2014 Mar 10.
25. Nemes S, Jonasson JM, Genell A, Steineck G. Bias in odds ratios by logistic regression modelling and sample size. BMC Med Res Methodol 2009;9:56.
26. Shang A, Huwiler-Muntener K, Nartey L, Juni P, Dorig S, Sterne JA, et al. Are the clinical effects of homoeopathy placebo effects? Comparative study of placebo-controlled trials of homoeopathy and allopathy. Lancet 2005 Aug 27;366(9487):726-32.
27. Dickersin K, Chalmers I. Recognising, investigating and dealing with incomplete and biased reporting of clinical research: from Francis Bacon

to the World Health Organisation. James Lind Library 2010 [cited 2014 Mar 12];Available from: URL: www.jameslindlibrary.org

28. Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. Cochrane Database Syst Rev 2009;(1):MR000006.

29. Naaktgeboren CA, de Groot JA, van SM, Moons KG, Reitsma JB. Evaluating diagnostic accuracy in the face of multiple reference standards. Ann Intern Med 2013 Aug 6;159(3):195-202.

30. Haines TP, Hill K, Walsh W, Osborne R. Design-related bias in hospital fall risk screening tool predictive accuracy evaluations: systematic review and meta-analysis. J Gerontol A Biol Sci Med Sci 2007 Jun;62(6):664-72.

31. Carley S, Dosman S, Jones SR, Harrison M. Simple nomograms to calculate sample size in diagnostic studies. Emerg Med J 2005 Mar;22(3):180-1.

32. Sterne JA, Egger M, Moher D. Adressing reporting bias; detecting repoting bias. In: Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions. Wiley-Blackwell; 2009. p. 310-24.

33. Sonnad SS, Langlotz CP, Schwartz JS. Accuracy of MR imaging for staging prostate cancer: a meta-analysis to examine the effect of technologic change. Acad Radiol 2001 Feb;8(2):149-57.

34. Lijmer JG, Leeflang M, Bossuyt PM. Proposals for a phased evaluation of medical tests. Med Decis Making 2009 Sep;29(5):E13-E21.

35. Irwig L, Bossuyt P, Glasziou P, Gatsonis C, Lijmer J. Designing studies to ensure that estimates of test accuracy are transferable. BMJ 2002 Mar 16;324(7338):669-71.

36. Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. BMJ 1997 Sep 13;315(7109):640-5.

37. DeAngelis CD, Drazen JM, Frizelle FA, Haug C, Hoey J, Horton R, et al. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. JAMA 2004 Sep 15;292(11):1363-4.

38. Korevaar DA, Bossuyt PM, Hooft L. Infrequent and incomplete registration of test accuracy studies: analysis of recent study reports. BMJ Open 2014;4(1):e004596.

APPENDIX 1. *Search strategy*

1. systematic.mp. [mp=ti, ab, sh, hw, tn, ot, dm, mf, dv, kw]
2. limit 1 to “reviews (best balance of sensitivity and specificity)”
3. predict*.ti,ab.
4. test.ti,ab.
5. tests.ti,ab
6. 4 or 5
7. 2 and 3 and 6
8. screen*.mp. [mp=ti, ab, sh, hw, tn, ot, dm, mf, dv, kw]
9. 2 and 8
10. monitoring.mp. [mp=ti, ab, sh, hw, tn, ot, dm, mf, dv, kw]
11. 2 and 10
12. “multiple tests”.mp. [mp=ti, ab, sh, hw, tn, ot, dm, mf, dv, kw]
13. 2 and 12
14. “diagnostic test accuracy”.mp. [mp=ti, ab, sh, hw, tn, ot, dm, mf, dv, kw]
15. DTA.ti,ab.
16. exp “sensitivity and specificity”/
17. specificit*.tw.
18. “false negative”.tw.
19. accuracy.tw.
20. 14 or 15 or 16 or 17 or 18 or 19
21. 2 and 20
22. 7 or 9 or 11 or 13 or 21
23. limit 22 to (english language and yr=”2011 -2013”)

APPENDIX 2. *Search strategy*

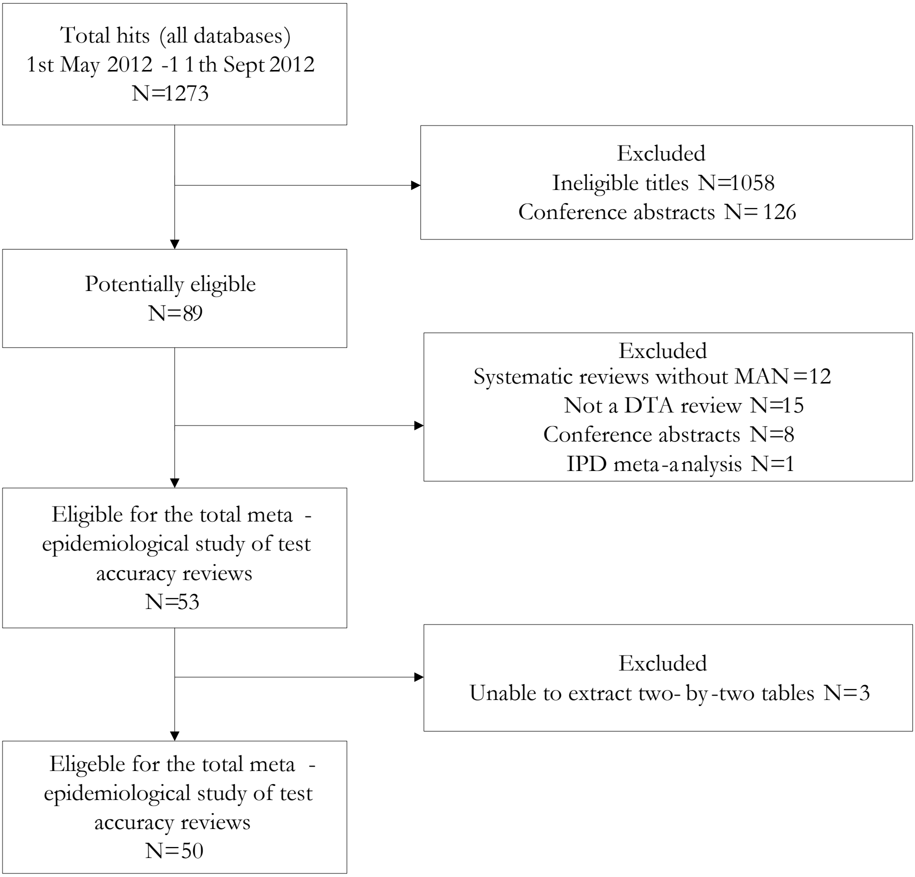

Total hits (all databases)
1st May 2012 -1 1th Sept 2012
N=1273
Excluded
Ineligible titles N=1058
Conference abstracts N= 126
Potentially eligible
N=89
Excluded
Systematic reviews without MAN =12
Not a DTA review N=15
Conference abstracts N=8
IPD meta-analysis N=1
Eligible for the total meta - epidemiological study of test accuracy reviews
N=53
Excluded
Unable to extract two- by-two tables N=3
Eligeble for the total meta - epidemiological study of test accuracy reviews
N=50

Figure 1_appendix. *Flow chart of selection process of the included reviews and meta-anlayses*

APPENDIX 3. *List of articles included in the review*

1. Al-Sukhni E, Milot L, Fruitman M, Beyene J, Victor JC, Schmocker S, Brown G, McLeod R, Kennedy E. Diagnostic accuracy of MRI for assessment of T category, lymph node metastases, and circumferential resection margin involvement in patients with rectal cancer: a systematic review and meta-analysis. Annals of Surgical Oncology 2012 July;19(7):2212-23.

2. Alldred SK, Deeks JJ, Guo B, Neilson JP, Alfirevic Z. Second trimester serum tests for Down's Syndrome screening. Cochrane Database Syst Rev 2012;6:CD009925.

3. Banerjee A, Newman DR, Van den Bruel A, Heneghan C. Diagnostic accuracy of exercise stress testing for coronary artery disease: a systematic review and meta-analysis of prospective studies. [Review]. International Journal of Clinical Practice 2012 May;66(5):477-92.

4. Beynon R, Sterne JA, Wilcock G, Likeman M, Harbord RM, Astin MP, Burke M, Bessell A, Ben-Shlomo Y, Hawkins J, Hollingworth W, Whiting PF. Is MRI better than CT for detecting a vascular component to dementia? A systematic review and meta-analysis. BMC Neurol 2012 June 6;12(1):33.

5. Chang K, Lu W, Wang J, Zhang K, Jia S, Li F, Deng S, Chen M. Rapid and effective diagnosis of tuberculosis and rifampicin resistance with Xpert MTB/RIF assay: a meta-analysis. J Infect 2012 June;64(6):580-8.

6. Chen J, Yang R, Lu Y, Xia Y, Zhou H. Diagnostic accuracy of endoscopic ultrasound-guided fine-needle aspiration for solid pancreatic lesion: a systematic review. Journal of Cancer Research & Clinical Oncology 2012 September;138(9):1433-41.

7. Chen C, Yang Z, Li Z, Li L. Accuracy of several cervical screening strategies for early detection of cervical cancer: A meta-analysis. International Journal of Gynecological Cancer 2012 July;22(6):908-21.

8. Cheng X, Li Y, Liu B, Xu Z, Bao L, Wang J. 18F-FDG PET/CT and PET for evaluation of pathological response to neoadjuvant chemotherapy in breast cancer: a meta-analysis. Acta Radiologica 2012 July 1;53(6):615-27.

9. De Jong MC, Genders TSS, Van GRJ, Moelker A, Hunink MGM. Diagnostic performance of stress myocardial perfusion imaging for coronary artery disease: A systematic review and meta-analysis. European Radiology 2012 September;22(9):1881-95.

10. Diel R, Loddenkemper R, Nienhaus A. Predictive value of interferon-release assays and tuberculin skin testing for progression from latent TB infection to disease state: A meta-analysis. Chest 2012 July;142(1):63-75.

11. Evangelista L, Cervino AR, Ghiotto C, Al-Nahhas A, Rubello D, Muzzio PC. Tumor marker-guided PET in breast cancer patients-a recipe for a perfect wedding: a systematic literature review and meta-analysis. [Review]. Clinical Nuclear Medicine 2012 May;37(5):467-74.

12. Fan L, Chen Z, Hao XH, Hu ZY, Xiao HP. Interferon-gamma release assays for the diagnosis of extrapulmonary tuberculosis: a systematic review and meta-analysis. FEMS Immunology & Medical Microbiology 2012 August;65(3):456-66.

13. Jaarsma C, Leiner T, Bekkers SC, Crijns HJ, Wildberger JE, Nagel E, Nelemans PJ, Schalla S. Diagnostic performance of noninvasive myocardial perfusion imaging using single-photon emission computed tomography, cardiac magnetic resonance, and positron emission tomography imaging for the detection of obstructive coronary artery disease: a meta-analysis. J Am Coll Cardiol 2012 May 8;59(19):1719-28.

14. Kiewiet JJ, Leeuwenburgh MM, Bipat S, Bossuyt PM, Stoker J, Boermeester MA. A systematic review and meta-analysis of diagnostic performance of imaging in acute cholecystitis. Radiology 2012 September;264(3):708-20.

15. Kim HP, Vance RB, Shaheen NJ, Dellon ES. The Prevalence and Diagnostic Utility of Endoscopic Features of Eosinophilic Esophagitis: A Meta-analysis. Clinical Gastroenterology & Hepatology 2012 September;10(9):988-96.

16. Kim HP, Vance RB, Shaheen NJ, Dellon ES. The Prevalence and Diagnostic Utility of Endoscopic Features of Eosinophilic Esophagitis: A Meta-analysis. Clinical Gastroenterology & Hepatology 2012 September;10(9):988-96.

17. Kocken M, Uijterwaal MH, de Vries AL, Berkhof J, Ket JC, Helmerhorst TJ, Meijer CJ. High-risk human papillomavirus testing versus cytology in predicting post-treatment disease in women treated for high-grade cervical disease: a systematic review and meta-analysis. [Review]. Gynecologic Oncology 2012 May;125(2):500-7.

18. Li C, Su N, Yang X, Yang X, Shi Z, Li L. Ultrasonography for detection of disc displacement of temporomandibular joint: a systematic review and meta-analysis. [Review]. Journal of Oral & Maxillofacial Surgery 2012 June;70(6):1300-9.

19. Lin CY, Chen JH, Liang JA, Lin CC, Jeng LB, Kao CH. 18F-FDG PET or PET/CT for detecting extrahepatic metastases or recurrent hepatocellular

19. carcinoma: A systematic review and meta-analysis. European Journal of Radiology 2012 September;81(9):2417-22.

20. Lu YY, Chen JH, Lin WY, Liang JA, Wang HY, Tsai SC, Kao CH. FDG PET or PET/CT for Detecting Intramedullary and Extramedullary Lesions in Multiple Myeloma: A Systematic Review and Meta-analysis. Clinical Nuclear Medicine 2012 September;37(9):833-7.

21. Lu Y-Y, Chen J-H, Liang J-A, Wang H-Y, Lin C-C, Lin W-Y, Kao C-H. Clinical value of FDG PET or PET/CT in urinary bladder cancer: A systemic review and meta-analysis. European Journal of Radiology 2012 September;81(9):2411-6.

22. Mavromatis ID, Antonopoulos CN, Matsoukis IL, Frangos CC, Skalkidou A, Creatsas G, Petridou ET. Validity of intraoperative gross examination of myometrial invasion in patients with endometrial cancer: a meta-analysis. Acta Obstetricia et Gynecologica Scandinavica 2012 July;91(7):779-93.

23. Morris RK, Riley RD, Doug M, Deeks JJ, Kilby MD. Diagnostic accuracy of spot urinary protein and albumin to creatinine ratios for detection of significant proteinuria or adverse pregnancy outcome in patients with suspected pre-eclampsia: systematic review and meta-analysis. BMJ 2012;345:e4342.

24. Neto AS, Nassar AP, Jr., Cardoso SO, Manetta JA, Pereira VG, Esposito DC, Damasceno MC, Slooter AJ. Delirium screening in critically ill patients: a systematic review and meta-analysis. [Review]. Critical Care Medicine 2012 June;40(6):1946-51.

25. Pai NP, Balram B, Shivkumar S, Martinez-Cajas JL, Claessens C, Lambert G, Peeling RW, Joseph L. Head-to-head comparison of accuracy of a rapid point-of-care HIV test with oral versus whole-blood specimens: A systematic review and meta-analysis. The Lancet Infectious Diseases 2012 May;12(5):373-80.

26. Phillips B, Wade R, Westwood M, Riley R, Sutton AJ. Systematic review and meta-analysis of the value of clinical features to exclude radiographic pneumonia in febrile neutropenic episodes in children and young people. Journal of Paediatrics & Child Health 2012 August;48(8):641-8.

27. Qu X, Huang X, Yan W, Wu L, Dai K. A meta-analysis of (1)(8) FDG-PET-CT, (1)(8)FDG-PET, MRI and bone scintigraphy for diagnosis of bone metastases in patients with lung cancer. Eur J Radiol 2012 May;81(5):1007-15.

28. Romero J, Xue X, Gonzalez W, Garcia MJ. CMR imaging assessing viability in patients with chronic ventricular dysfunction due to coronary artery disease: a meta-analysis of prospective trials. Jacc: Cardiovascular Imaging 2012 May;5(5):494-508.

29. Sadeghi R, Gholami H, Zakavi SR, Kakhki VR, Horenblas S. Accuracy of 18F-FDG PET/CT for diagnosing inguinal lymph node involvement in penile squamous cell carcinoma: systematic review and meta-analysis of the literature. Clin Nucl Med 2012 May;37(5):436-41.

30. Sadigh G, Carlos RC, Neal CH, Dwamena BA. Ultrasonographic differentiation of malignant from benign breast lesions: A meta-analytic comparison of elasticity and BIRADS scoring. Breast Cancer Research and Treatment 2012 May;133(1):23-35.

31. Sandroni C, Cavallaro F, Marano C, Falcone C, De SP, Antonelli M. Accuracy of plethysmographic indices as predictors of fluid responsiveness in mechanically ventilated adults: a systematic review and meta-analysis. Intensive Care Medicine 2012 September;38(9):1429-37.

32. Shang Y, Ju W, Kong Y, Schroder PM, Liang W, Ling X, Guo Z, He X. Performance of polymerase chain reaction techniques detecting perforin in the diagnosis of acute renal rejection: a meta-analysis. PLoS ONE [Electronic Resource] 2012;7(6):e39610.

33. Shen Y-C, Liu M-Q, Wan C, Chen L, Wang T, Wen F-Q. Diagnostic accuracy of vascular endothelial growth factor for malignant pleural effusion: A meta-analysis. Experimental and Therapeutic Medicine 2012 June;3(6):1072-6.

34. Siddiqui MR, Ashrafian H, Tozer P, Daulatzai N, Burling D, Hart A, Athanasiou T, Phillips RK. A diagnostic accuracy meta-analysis of endoanal ultrasound and MRI for perianal fistula assessment. [Review]. Diseases of the Colon & Rectum 2012 May;55(5):576-85.

35. Singh B, Parsaik AK, Agarwal D, Surana A, Mascarenhas SS, Chandra S. Diagnostic accuracy of pulmonary embolism rule-out criteria: a systematic review and meta-analysis. [Review]. Annals of Emergency Medicine 2012 June;59(6):517-20.

36. Smith TO, Lewis M, Song F, Toms AP, Donell ST, Hing CB. The diagnostic accuracy of anterior cruciate ligament rupture using magnetic resonance imaging: A meta-analysis. European Journal of Orthopaedic Surgery and Traumatology 2012 May;22(4):315-26.

37. Smith TO, Drew B, Toms AP, Jerosch-Herold C, Chojnowski AJ. Diagnostic accuracy of magnetic resonance imaging and magnetic resonance arthrography for triangular fibrocartilaginous complex injury:

a systematic review and meta-analysis. [Review]. Journal of Bone & Joint Surgery - American Volume 2012 May 2;94(9):824-32.

38. Smith TO, Drew BT, Toms AP. A meta-analysis of the diagnostic test accuracy of MRA and MRI for the detection of glenoid labral injury. Archives of Orthopaedic and Trauma Surgery 2012 July;132(7):905-19.

39. Tai T-W, Wu C-Y, Su F-C, Chern T-C, Jou I-M. Ultrasonography for Diagnosing Carpal Tunnel Syndrome: A Meta-Analysis of Diagnostic Test Accuracy. Ultrasound in Medicine and Biology 2012 July;38(7):1121-8.

40. Tashakkor AY, Nicolaou S, Leipsic J, Mancini GB. The Emerging Role of Cardiac Computed Tomography for the Assessment of Coronary Perfusion: A Systematic Review and Meta-analysis. Canadian Journal of Cardiology 2012 July;28(4):413-22.

41. Thangaratinam S, Brown K, Zamora J, Khan KS, Ewer AK. Pulse oximetry screening for critical congenital heart defects in asymptomatic newborn babies: a systematic review and meta-analysis. Lancet 2012 May 1.

42. Treglia G, Castaldi P, Rindi G, Giordano A, Rufini V. Diagnostic performance of Gallium-68 somatostatin receptor PET and PET/CT in patients with thoracic and gastroenteropancreatic neuroendocrine tumours: A meta-analysis. Endocrine 2012 August;42(1):80-7.

43. Underwood M, Arbyn M, Redman C, Smith WP. Accuracy of colposcopic directed punch biopsies: A systematic review and meta-analysis. BJOG: An International Journal of Obstetrics and Gynaecology 2012 June;Conference(var.pagings):163.

44. van Teeffelen AS, Van Der Heijden J, Oei SG, Porath MM, Willekes C, Opmeer B, Mol BW. Accuracy of imaging parameters in the prediction of lethal pulmonary hypoplasia secondary to mid-trimester prelabor rupture of fetal membranes: a systematic review and meta-analysis. Ultrasound in Obstetrics & Gynecology 2012 May;39(5):495-9.

45. Wang Z, Dong ZY, Chen JQ, Liu JL. Diagnostic value of sentinel lymph node biopsy in gastric cancer: a meta-analysis. [Review]. Annals of Surgical Oncology 2012 May;19(5):1541-50.

46. Webb RC, Howard RS, Stojadinovic A, Gaitonde DY, Wallace MK, Ahmed J, Burch HB. The utility of serum thyroglobulin measurement at the time of remnant ablation for predicting disease-free status in patients with differentiated thyroid cancer: a meta-analysis involving 3947 patients. Journal of Clinical Endocrinology & Metabolism 2012 August;97(8):2754-63.

47. Wu L, Dai ZY, Qian YH, Shi Y, Liu FJ, Yang C. Diagnostic Value of Serum

Human Epididymis Protein 4 (HE4) in Ovarian Carcinoma: A Systematic Review and Meta-Analysis. International Journal of Gynecological Cancer 2012 September;22(7):1106-12.

48. Wu L-M, Gu H-Y, Qu X-H, Zheng J, Zhang W, Yin Y, Xu J-R. The accuracy of ultrasonography in the preoperative diagnosis of cervical lymph node metastasis in patients with papillary thyroid carcinoma: A meta-analysis. European Journal of Radiology 2012 August;81(8):1798-805.

49. Wu L-M, Hu J-N, Hua J, Liu M-J, Chen J, Xu J-R. Diagnostic value of diffusion-weighted magnetic resonance imaging compared with fluorodeoxyglucose positron emission tomography/computed tomography for pancreatic malignancy: A meta-analysis using a hierarchical regression model. Journal of Gastroenterology and Hepatology 2012 June;27(6):1027-35.

50. Zhao L, He Z-Y, Zhong X-N, Cui M-L. 18FDG-PET/CT for detection of mediastinal nodal metastasis in non-small cell lung cancer: A meta-analysis. Surgical Oncology 2012 September;21(3):230-6.

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) and87% (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 frin 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.

REFERENCE LIST

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.
2. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. Biometrics 1994 Dec;50(4):1088-101.
3. Egger M, Davey SG, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997 Sep 13;315(7109):629-34.
4. Sampson M, Barrowman NJ, Moher D, Klassen TP, Pham B, Platt R, et al. Should meta-analysts search Embase in addition to Medline? J Clin Epidemiol 2003 Oct;56(10):943-55.
5. de Vet HCW, Eisinga A, Riphagen II, Aertgeerts B, Pewsner D. Chapter 7: Searching for Studies. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. Version 0.4 [updated September 2008] ed. The Cochrane Collaboration; 2008.
6. Lefebvre C, Manheimer E, Glanville J. Chapter 6. Searching for studies. In: Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions. 5.1.0 ed. Wiley; 2008.
7. Abaid LN, Grimes DA, Schulz KF. Reducing publication bias of prospective clinical trials through trial registration. Contraception 2007 Nov;76(5):339-41.
8. van Enst WA, Scholten RJ, Hooft L. Identification of additional trials in prospective trial registers for Cochrane systematic reviews. PLoS One 2012;7(8):e42812.
9. Wager E, Williams P. "Hardly worth the effort"? Medical journals' policies and their editors' and publishers' views on trial registration and publication bias: quantitative and qualitative study. BMJ 2013;347:f5248.

10 Dickersin K, Rennie D. Registering clinical trials. JAMA 2003 Jul 23;290(4):516-23.

11. Viergever RF, Karam G, Reis A, Ghersi D. The quality of registration of clinical trials: still a problem. PLoS One 2014;9(1):e84727.
12. All Trials Registered | All Results Reported. All Trials Campaign 2014 May 6Available from: URL: http://www.alltrials.net/
13. U.S.Government Information. Food and Drug Administration Amendments Act 801 Requirements. PUBLIC LAW 110–85. 2014. 121 STAT. 904.
14. Lefebvre C, Glanville J, Wieland LS, Coles B, Weightman AL.

Methodological developments in searching for studies for systematic reviews: past, present and future? Syst Rev 2013;2:78.

15. Cepeda MS, Lobanov V, Berlin JA. From ClinicalTrials.gov trial registry to an analysis-ready database of clinical trial results. Clin Trials 2013 Apr;10(2):347-8.

16. Holmstrom B, Johansson M, Bergh A, Stenman UH, Hallmans G, Stattin P. Prostate specific antigen for early detection of prostate cancer: longitudinal study. BMJ 2009;339:b3537.

17. Beynon R, Leeflang MM, McDonald S, Eisinga A, Mitchell RL, Whiting P, et al. Search strategies to identify diagnostic accuracy studies in MEDLINE and EMBASE. Cochrane Database Syst Rev 2013;9:MR000022.

18. Leeflang MM, Scholten RJ, Rutjes AW, Reitsma JB, Bossuyt PM. Use of methodological search filters to identify diagnostic accuracy studies can lead to the omission of relevant studies. J Clin Epidemiol 2006 Mar;59(3):234-40.

19. Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. Evidence-Based Medicine: how to practise and teach EBM. Second Edition ed. Edinburgh: Churchill Livingstone; 2000.

20. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008 Apr 26;336(7650):924-6.

21. Schunemann HJ, Oxman AD, Vist GE, Higgins JPT, Deeks JJ, Glasziou PP, et al. Chapter 12: Interpreting results and drawing conclusions. In: Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions. 5.1.0 ed. Wiley; 2008.

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

23. Korevaar DA, van Enst WA, Spijker R, Bossuyt PM, Hooft L. Reporting quality of diagnostic accuracy studies: a systematic review and meta-analysis of investigations on adherence to STARD. Evid Based Med 2014 Apr;19(2):47-54.

24. Ochodo EA, van Enst WA, Naaktgeboren CA, de Groot JA, Hooft L, Moons KG, et al. Incorporating quality assessments of primary studies in the conclusions of diagnostic accuracy reviews: a cross-sectional study. BMC Med Res Methodol 2014;14:33.

25. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol 2005 Oct;58(10):982-90.

26. Schunemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al. GRADE: Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ 2008 May 17;336(7653):1106-10.

27. van Enst WA, Ochodo E, Scholten RJ, Hooft L, Leeflang MM. Investigation of publication bias in meta-analyses of diagnostic test accuracy: a meta-epidemiological study. BMC Med Res Methodol 2014;14(1):70.

HOOFDSTUK 9

Samenvatting en conclusies

Het centrale thema van dit proefschrift is de methodologie van een systematische review, het hoogste niveau van bewijs voor het maken van klinische beslissingen. Dit proefschrift geeft verdere perspectieven hoe de validiteit van systematische reviews kan worden verbeterd. In het komende hoofdstuk zijn de resultaten van het proefschrift samengevat.

Een systematische review van hoge kwaliteit bestaat uit verschillende stappen die gezamenlijk leiden tot het verkrijgen van een valide antwoord op een onderzoeksvraag. In dit proefschrift zijn enkele stappen van het reviewproces onderzocht: identificeren van studies, het beoordelen van het risico op vertekening (bias), de kwaliteit van rapporteren in primaire studies, evalueren van heterogeniteit en evalueren van publicatie bias. Er werd een aanzienlijk aantal methodologische aandachtspunten geïdentificeerd. Het merendeel van ons onderzoek was gericht op vraagstukken bij het reviewproces en meta-analyses van diagnostische test accuratesse (DTA) studies. Dit is relatief een jong veld in het evidence-based onderzoek.

Hoofdstukken 2 en 3 zijn gericht op het identificeren van studies. **Hoofdstuk 2** bespreekt het gebruik van prospectieve trial registers voor het identificeren van interventiestudies (veelal gerandomiseerde gecontroleerde trials, RCT's). Deze methode wordt gebruikt in additie op de gewoonlijk gebruikte elektronische databases. De evaluatie toonde aan dat binnen een cohort van 210 Cochrane systematische interventie reviews, ongeveer een derde (38,1%) van de auteurs in prospectieve trial registers hadden gezocht. Zeventig procent van deze reviews gebruikte hierbij een zoekplatform waarin meerdere registers tegelijkertijd kunnen worden doorzocht. In 35% resulteerde de zoekactie in prospectieve trial registers voor het identificeren van een additionele relevante studie, waarvan uiteindelijk 14,3% van de studies daadwerkelijk kon worden geïncludeerd in de review. In 71,4% kon een studie worden benoemd als lopende studie, maar kon deze niet verder bijdragen aan de resultaten van de review. **In hoofdstuk 3** hebben we onderzocht of de gepoolde puntschatter in meta-analyses van diagnostische accuratesse studies wordt beïnvloed als een zoekactie wordt beperkt tot het zoeken in MEDLINE. De relatieve diagnostische odds ratio (RDOR) voor de vergelijking van enkel MEDLINE geïndexeerde studies ten opzichte van alle studies was 1,04 (95% CI 0,94 tot 1,15). Dit betekend dat een beperking tot MEDLINE studies de puntschatter licht zal overschatten, maar dit verschil was niet significant. De sensitiviteit zal 0,08% hoger zijn (95% CI -1% tot 1%) en de specificiteit 0,1% (95% CI -0,8% tot 1%). Een substantieel aantal reviews (57%), zal überhaupt geen verandering ondervinden omdat alle geïncludeerde studies waren geïndexeerd in MEDLINE. We konden concluderen dat gemiddeld genomen het uitsluiten van niet-MEDELINE-geïndexeerde studies de resultaten van diagnostische accuratesse meta-analyse niet zullen beïnvloeden. Echter is het resultaat voor een individuele review nog steeds onvoorspelbaar.

Hoofdstuk 4 beschrijft hoe in 65 recent gepubliceerde diagnostische accuratesse reviews wordt omgegaan met de methodische kwaliteit van primaire studies. De methodologische kwaliteit was door 92% van reviews geëvalueerd, waarvan 64% het Quality Assessment of Diagnostic Accuracy Studies I (QUADAS I) instrument gebruikten en 3% QUADAS II. In 72% van deze reviews werden de resultaten van de methodologische kwaliteitsbeoordeling bediscussieerd, maar enkel 9% van de reviews maakte daadwerkelijk een koppeling tussen de methodologische kwaliteitsbeoordeling en de conclusies. De helft van de reviews waarin de kwaliteitsbeoordeling werd meegewogen in de conclusies had geen meta-analyse uitgevoerd vanwege de aanwezigheid van aanzienlijke heterogeniteit of een hoog risico hadden op bias. De kwaliteitsbeoordeling was genoemd in 43% (n=28) van de samenvattingen van de reviews, waarbij slechts in vijf abstracts de conclusies werd gekoppeld aan de kwaliteit van onderliggende studies. We hebben geconcludeerd dat de kwaliteit van het rapporteren van kwaliteitsbeoordeling en studieresultaten in systematische reviews van diagnostische accuratesse studies moet verbeteren om zodoende de lezer een meer valide beeld te geven over de prestaties van de geëvalueerde test in de klinische praktijk.

Hoofdstuk 5 presenteert een overview van systematische reviews waarin de kwaliteit werd geëvalueerd van de rapportage in primaire diagnostische test accuratesse studies volgens de normen van de Standaard voor Rapportage in Diagnostische Accuratesse studies (STARD). Volledig en transparante rapportage van de primaire test accuratesse studie is een voorwaarde om de validiteit van het design, het uitvoeren van de studie en de analyses te evalueren en te interpreteren voor zowel reviewauteurs en gebruikers van reviews. Er waren en zestien reviews geïncludeerd waarin de kwaliteit van rapportage van in totaal 1496 geëvalueerd studies werden geëvalueerd. We concludeerde dat de kwaliteit van rapportage suboptimaal was. Van de in 25 STARD-items, werden er gemiddeld 9,1 tot 14,3 (mediaan 12,8) gerapporteerd. Het aantal items dat wordt gerapporteerd is licht gestegen sinds de introductie van STARD in 2003 (1,41 items; 95% BI: 0,65 tot 2,18). Een analyse voor de individuele items toonde aan dat zeven specifieke items bijzonder laag scoorden: item 10 (uitvoerder van de test), item 11 (blinderen van de testlezer), item 13 (methode voor het berekenen van de reproduceerbaarheid), item 16 (aantal geselecteerde studiedeelnemers die de test niet hebben ondergaan), item 20 (complicaties), item 22 (omgang met missende waarden) en item 24 (schatten van de reproduceerbaarheid). Dit is alarmerend aangezien enkele van deze items gerelateerd zijn aan het optreden van bias in de resultaten. In het algemeen kan gesteld worden dat hoewel er een kleine verbetering is in het rapporteren sinds de introductie van STARD, er nog veel verbetering kan worden behaald. Navolgen van STARD zou verder bekend gemaakt moeten worden en aangeraden worden onder onderzoekers, redacties, en peer reviewers vanaf het moment dat de studie wordt ontworpen en bij verdere ontwikkeling de studie.

Het bestuderen van heterogeniteit is in meta-analyses van diagnostische accuratesse studies moeilijker door het bivariate karakter van de data. **Hoofdstuk 6** is gericht op het beschouwen van heterogeniteit in 65 diagnostische accuratesse reviews. In twaalf van deze reviews was geen meta-analyse uitgevoerd waarbij de substantiële heterogeniteit vaak werd benoemd als rede. In 53 van de 65 reviews werden één of meerdere methoden toegepast om de heterogeniteit van de studiedata te onderzoeken. In 47,2% van de reviews werd een gestratificeerde analyse uitgevoerd, in 35,8% een sensitiviteitsanalyse en in 22,6% een meta-regressie. Het aantal bronnen van heterogeniteit dat werd onderzocht was hoog ten opzichte van het aantal studies in de meta-analyse (mediaan ratio 1:5). Op basis van de resultaten van de beschouwing hebben we suggesties gegeven wat review auteurs zouden kunnen overwegen en rapporteren wanneer ze heterogeniteit willen onderzoeken in diagnostische accuratesse reviews.

Hoofdstukken 7 en 8 gaan beide over selectieve publicatie. Het wordt vaak suggereert dat kleine studies vaker worden gepubliceerd wanneer ze positieve en significante resultaten hebben. Grotere studies hebben meer kans om gesubmit, geaccepteerd en gepubliceerd te worden ongeachte de resultaten ten opzichte van kleine studies. In **hoofdstuk 7** zijn de methoden beschreven die gebruikt worden om selectieve publicatie in DTA meta-analyses te onderzoeken. In een cohort van 114 reviews werd in 41,2% de mogelijkheid van publicatie bias onderzocht door het toepassen van een grafische of statistische methoden waarmee een zogeheten “kleine-studie effect” kan worden aangetoond. De methoden zijn ontwikkeld om te onderzoeken of er een relatie bestaat tussen de grootte van de studie en het resultaat van de studie. In 31 reviews werd een trechtergrafiek (funnelplot) geconstrueerd waarbij een grote variatie aan diagnostische parameters werd gebruikt. In 41 reviews werd een statistische test toegepast om de aanwezigheid van een klein-studie effect te onderzoeken. De statische test van Deeks (1) is specifiek ontwikkeld voor meta-analyses van diagnostische accuratesse studies, maar werd enkel gebruikt in 29,2% reviews. De meest gebruikte testen was de Egger test (43,9%) (2). Deze test heeft een verhoogde kans op een type-1 fout wanneer deze wordt gebruikt in meta-analyses van diagnostische studies. Behalve de evaluatie van de gebruikte methoden om selectieve publicatie in DTA reviews te identificeren, onderzochten wij ook de meeste gebruikte testen op de concordantie van de resultaten van de verschillende testen. Hiervoor gebruikte we de data van de geïncludeerde reviews en pasten hier de Begg, Egger en Deeks test op toe (1-3). Terwijl deze testen allemaal hetzelfde concept trachten te onderzoeken was de concordantie voor de resultaten van de verschillende testen 87% (Begg vs. Deeks) tot een teleurstellende 66% (Deeks vs. Egger). Wij raden reviewers aan om de Deeks test te gebruiken als primaire methode. Deze test kan worden toegepast onder de voorwaarde dat de resultaten met bedachtzaamheid zullen geïnterpreteerd aangezien de mechanisme die ten grondslag liggen aan het kleine-studie effect nog niet bekend zijn. In **hoofdstuk 8** hebben wij de aanwezigheid van het kleine-studie effect en tijdseffect onderzocht in meta-analyses van diagnostische studies. In plaats van het verwachtte kleine-studie effect te identificeren vonden we het tegenovergestelde: kleine studies hadden gemiddeld een lagere accuratesse dan grote studies. Mogelijk was dit resultaat een gevolg van een statistische artefact: odds ratio's worden overschat in kleine studies als gevolg van de eigenschappen van een logisch regressie model. Echter, dit is geen volledige verklaring voor het gevonden resultaat. We vonden wel een tijdeffect, maar enkel wanneer we het contrast vergoten door de 25% meest recente te vergelijken met de 25% oudste studies binnen een meta-analyse. We vonden een lichte daling van de Diagnostische Odds Ratio (DOR), maar deze was lang niet zo sterk als in meta-analyses van interventie studies. We concludeerde dat sommige mechanismen die worden geassocieerd met selectieve publicatie van RCT's minder sterk aanwezig zijn in studies naar test accuratesse.

REFERENCE LIST

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.

2. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. Biometrics 1994 Dec;50(4):1088-101.

3. Egger M, Davey SG, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997 Sep 13;315(7109):629-34.

APPENDIX I

Abbreviations

ABBREVIATIONS

ANCA: Anti-Neutrophil Cytoplasmic Antibody
AMSTAR: Assessment of Multiple Systematic Reviews
ANZCTR: Australian New Zealand Clinical Trials Registry
AUC: Area Under the Curve

CDSR: Cochrane Database of Systematic Reviews
ChiCTR: Chinese Clinical Trial Register
CLIB: Cochrane Library
CONSORT: Consolidated Standards of Reporting Trials

DOR: Diagnostic Odds Ratio
DTA: Diagnostic Test Accuracy

e.g: exempli gratia
ESS: Effective Sample Size

FDA: Food and Drug Administration

GRADE: Grading of Recommendations Assessment, Development and Evaluation

HSROC: Hierarchical Summary Receiving Operating Characteristic

IA: Invasive aspergillosis
ICMJE International Committee of Medical Journal Editors
ICTRP: International Clinical Trial Registry Platform
i.e: id est
iM: in MEDLINE
IQ: interquartile
IQR: interquartile range
ISRCTN: International Standard Randomised Controlled Trial Number Register

lnDOR: natural logarithm of the Odds Ratio
log: logarithm

MA: meta-analysis
MeSH: Medical Subject Headings

N: number
NiM: Not in MEDLINE
NNR: Number Needed to Read
NTR: Netherlands Trial Register

OR: Odds Ratio

PRISMA: preferred reporting items for systematic reviews and meta-analyses

QUADAS: Quality Assessment for Diagnostic Accuracy Studies tool

RCT: randomized controlled trials
RDOR: Relative Diagnostic Odds Ratio
ROC: Receiving Operating Characteristic
RR: Relative Risk

SE: Standard Error
Sens: sensitivity
Spec: specificity
SROC: Summary Receiving Operating Characteristic
STARD: Standards for Reporting of Diagnostic Accuracy Studies

Vs: versus

WHO: World Health Organisation

APPENDIX 2

PhD Portfolio

PHD PORTFOLIO

Name PhD student:	Wynanda Annefloor van Enst
PhD period:	March 2010 – September 2014
Name PhD supervisors:	Prof. dr. R.J.P.M. Scholten Dr. L. Hooft

PHD TRAINING

Courses	Institute	Year
Scientific Writing in English for Publication	AMC Graduate School	2010
Advanced Topics in Clinical Epidemiology-	AMC Graduate School	2010
Oral presentation in English	AMC Graduate School	2010
Systematic Reviews of Interventions	Dutch Cochrane Centre	2010
Evidence Based Searching	Dutch Cochrane Centre	2010
Expert Management of Medical Literature	Medical Library, AMC	2010
AMC World of Science	AMC Graduate School	2010
Evidence-based Development of Guidelines	CBO / Dutch Cochrane Centre	2011
Project Management	AMC Graduate School	2011
Entrepreneurship in Health and Life Sciences	AMC Graduate School	2012
Cochrane Systematic Reviews of Diagnostic Test Accuracy	Dutch Cochrane Centre	2012
Statistical Methods for Diagnostic Test Accuracy Reviews	Cochrane Systematic Reviews of DTA, University of Birmingham	2013
GRADE	CBO/Dutch Cochrane Centre	2013
Educational Skills Training	AMC Graduate School	2013
Practical Biostatistics	AMC Graduate School	2013
Citation Analysis and Impact Factors	AMC Graduate School	2013
Career Development	AMC Graduate School	2013

Seminars, workshops and master classes	Institute	Year
Weekly department seminars		2010-2014
Workshop GRADE for systematic reviews by prof. dr. Holger Schünemann	Keystone, USA	2010
Master Class "Who wrote my paper" by prof. dr. Drummond Rennie	AMC Graduate School	2013
Equator lectur: "Reporting and reproducible research" by prof. dr. John Ioannidis	Freiburg, Germany	2013

Oral presentations	Institute	Year
Extending the search to find ongoing and unpublished trials: A survey of methods and results of Cochrane reviews	The Cochrane Colloquium, Keystone, USA	2010
Could a search for a diagnostic test accuracy review be restricted to MEDLINE?	The Cochrane Colloquium, Madrid, Spain	2011
Exploring mechanisms of publication bias in systematic reviews of diagnostic test accuracy.	The Cochrane Colloquium, Québec, Canada	2013

Poster presentations	Institute	Year
Could a search for a diagnostic test accuracy review be restricted to MEDLINE?	WEON, IJmuiden, the Netherlands	2011
How is publication bias investigated in diagnostic test accuracy reviews?	Methods for Evaluating Medical Tests & Biomarkers, Birmingham, United Kingdom	2013
Is a search in MEDLINE sufficient for a diagnostic test accuracy review?	Methods for Evaluating Medical Tests & Biomarkers, Birmingham, United Kingdom	2013
How do authors investigate selective publication in diagnostic test accuracy reviews?	The Cochrane Colloquium, Québec, Canada	2013
Poor interpretation of quality assessment results in diagnostic accuracy reviews.	The Cochrane Colloquium, Québec, Canada	2013

Teaching and supervision	Institute	Year
Tutor for the course Evidence-Based	Medicine in clinical practice	2012-2014
Teaching systematic review methodology	the AMC Graduate School	2011-2014
Teaching Cochrane systematic review methodology	authors of Cochrane systematic reviews	2011-2014
Supporting review authors with therapeutic and diagnostic test accuracy systematic reviews		2011-2014

LIST OF INTERNATIONAL PUBLICATIONS

1. Korevaar DA, Wang J, van Enst WA, Leeflang MM, Hooft L, Smidt N, Bossuyt PMM. Reporting Diagnostic Accuracy Studies: Still Improving After Nine Years of STARD? Radiology (accepted June 2014).
2. Henschke N, van Enst WA, Froud R, Ostelo R. Responder analyses in randomised controlled trials for chronic low-back pain: an overview of currently used methods. European Spine Journal. 2014 Apr; 23(4):772-8.
3. Mallee WH, Hennt EP, van Dijk CN, Kamminga S, van Enst WA, Kloen P. Clinical evaluation in suspected scaphoid fractures: a systematic review and meta-analysis. Journal of Hand Surgery (accepted June 2014).
4. Van de Glind EM, van Enst WA, van Munster BC, Olde Rikkert MG, Scheltens P, Scholten RJ, Hooft L. Pharmacological treatment of dementia: a scoping review of systematic reviews. Dement Geriatr Cogn Disord. 2013;36(3-4):211-28.
5. Bouwmeester W, van Enst WA, van Tulder M. Quality of low back pain guidelines improved. Spine 2009;34: 2562-2567.

LIST OF NATIONAL PUBLICATIONS

1. Van Enst WA, Dekker F. Ibuprofen of paracetamol bij migraine. Nederlands Tijdschrift voor Geneeskunde, 2011;155:A3414.
2. Van Enst WA and P Mistiaen. Telemonitortechnologie bij chronisch hartfalen. Nederlands Tijdschrift voor Geneeskunde, 2011;155:A3250.
3. Elbers, GM, van Enst WA. Helmplicht voor fietsers. Nederlands Tijdschrift voor Geneeskunde, 2010;154:A2728.
4. Hooft L, van Enst WA, Heus P, Langendam MW, Limpens CEJM, van de Wetering FT, Scholten RJPM. Sleeve Gastrectomie: UPDATE. Een systematische review. Dutch Cochrane Centre, 2013.
5. Van Enst WA, Langendam MW, Limpens CEJM. Bariatrische chirurgie voor prediabetes en diabetes mellitus type 2: UPDATE. Een systematische review. Dutch Cochrane Centre, 2012.
6. Langendam MW, van Enst WA, Hooft L, Spijker R. Effectiviteit van interspinale implantaten: systematische review. Dutch Cochrane Centre, 2012.
7. Hooft L, van de Glind E, Langendam MW, Bexkens B, Heus P, van Enst WA. De onzichtbare kant van Parkinson. Dutch Cochrane Centre, 2012.
8. Langendam MW, Hooft L, Heus P, van de Glind E, van Enst WA, Elbers GMH, Spijker R. Medicamenteuze en psychosociale interventies voor

9. GMH, Spijker R. Medicamenteuze en psychosociale interventies voor patiënten met dementie: scoping review. Dutch Cochrane Centre, 2012.

10. Kramer SF, Elbers GMH, van Enst WA, Limpens CEJM, Langendam MW. Sleeve Gastrectomie, een systematische review. Dutch Cochrane Centre, 2011.

11. Langendam MW, van Enst WA, Limpens CEJM. Bariatrische chirurgie voor prediabetes en diabetes mellitus type 2. Een systematische review. Dutch Cochrane Centre, 2011.

12. Kramer SF, Elbers GMH, van Enst WA, Langendam MW, Hooft L, Spijker R, Scholten RJPM. Fysio- en oefentherapie bij chronische aandoeningen: osteoporose. Een overview van systematisch reviews. (Physiotherapy interventions for osteoporosis). Dutch Cochrane Centre, 2010.

APPENDIX 3
Authors' affiliations

AUTHORS' AFFILIATIONS

Academic Medical Center, Amsterdam
Department of Clinical Epidemiology, Biostatistics and Bioinformatics
Aeilko H. Zwinderman
Daniël A. Korevaar
Eleanor A. Ochodo
Mariska M. Leeflang
Patrick M. Bossuyt

Dutch Cochrane Centre
Lotty Hooft
Rene Spijker
Rob J.P.M. Scholten

University of Bristol / Kleijnen Systematic Reviews, York, United Kingdom
School of Social and Community Medicine
Penny Whiting

University Medical Center, Utrecht
Julius Center for Health Sciences and Primary Care
Christiana A. Naaktgeboren
Johannes B. Reitsma
Joris A. de Groot
Karel G.M. Moons

4

APPENDIX
Dankwoord

Dit boekje is tot stand gekomen met hulp en steun van vele collega's, vrienden en familie. Graag wil ik hier iedereen die een bijdrage heeft geleverd, van harte bedanken.

Prof. dr. R.J.P.M. Scholten, beste Rob, ik ben ontzettend vereerd dat ik mijn promotieonderzoek bij het Dutch Cochrane Centre mocht uitvoeren. Je gaf mij de vrijheid om te doen wat ik interessante thema's vond en ondersteunde me in het gehele proces. Jouw deskundige commentaar heeft mijn werk zeker verbeterd.
Dr. L. Hooft, beste Lotty. Naast jouw wetenschappelijke ondersteuning, hielp je mij ook een richting te kiezen als jonge onderzoeker. Je bent optimistisch, motiverend en altijd betrokken. De beste besprekingen hadden we tijdens het halen van onze cappuccino's en ik zal die moment ook missen in de toekomst. Bedankt voor je steun, motivatie en vertrouwen.
Prof. dr. A.H. Zwinderman, beste Koos, ik ken weinig mensen die zo onbaatzuchtig zijn als jij. Ik kon met al mijn vragen bij je terecht en je nam ruim de tijd om ze te beantwoorden. Ook toen het DCC uit het AMC vertrok, zorgde je ervoor dat ik niet een nieuwe plek bij de KEBB vond. Het is een genoegen om met je te mogen samenwerken.
Prof. dr. P.M.M. Bossuyt, beste Patrick, ondanks dat ik geen promovenda van je was, nodigde je mij toch uit bij de BiTE groep en liet je me zelfs naar een congres gaan en cursus volgen in Birmingham. Daarnaast heb je me talloze keren geholpen met mijn onderzoek. Ik bewonder en waardeer je didactische kwaliteiten. Ik heb veel van je geleerd tot en met het maken van een paddenstoelensoep.
Mijn promotiecommissie wil ik bedanken voor hun bereidheid plaats te nemen in deze commissie, mijn proefschrift te lezen en beoordelen.

Lieve Cochrane-collega's. Ik heb veel geleerd van jullie diversiteit in werkwijzen en kennis. Daarnaast heb ik genoten van alle gezelligheid tijdens onze taartmomenten, etentjes en congressen, maar ook van de projecten die we samen hebben uitgevoerd. We hebben erg hard gewerkt maar het werden altijd mooie systematische reviews waar we trots op waren. Jullie hebben me gesteund bij mijn promotieonderzoek en hebben al mijn presentaties bezocht. Veel dank voor jullie bijdrage.

Daarnaast wil ik mijn kamergenoten bedanken. Allereerst, Roy en Sharon, we hebben erg veel gelachen, geëvalueerd en gefilosofeerd. Jullie waren

geïnteresseerd in mijn onderzoek en ik kon mijn overwegingen altijd met jullie bespreken. Zelfs toen Sharon al in Australië woonde.
Daarna Esther, ik was heel blij dat jij bij kwam zitten. We hebben samen gewerkt aan een complexe review. Ik heb geleerd van je optimisme, goede gesprekken, en je tomeloze kracht. Ik wens je veel succes bij je verdediging op 4 september.
Finally, Jérémie, Mareen and Daniël. You were the perfect combination of science and fun. It was very nice to discuss topics on publication bias, quality of reporting en the future Dutch EQUATOR centre with you but I also loved to have a chat during lunches, borrels and diners. I really appreciated your interest and support. I hope to keep seeing you.

Dear BiTE group, I was really happy that I could join you. I learnt from the interesting discussions we had during our BiTE-lunches, about test evaluation, but also world history, politics, languages, cultural traditions and all other things.

Christiana and Eleanor, I would like to thank you for our fruitful cooperation on the database-project. It led to several papers fort his thesis and it was also a pleasure to work with both of you. Christiana, I was really amazed by your terrific organisation. Especially when I found out you even make Excel spread sheet to organize some of your personal life.

Lieve Damesch '07. Jullie zijn een bijzonder diverse groep vriendinnen en misschien daarom ook wel zo leuk. Ik ben erg veel met mijn proefschrift bezig geweest en heb jullie te weinig gezien. Ik hoop dat dit vanaf nu weer anders wordt en vaak van jullie gezelligheid kan genieten.
Lieve Maud, onze etentjes en fietsafspraken waren ontzettend fijn. Zeker in periode waarin ik veel te gelijk moest doen, waren jouw opmerkingen tomeloos relativerend en ging ik weer zorgeloos op huis aan. Bedankt voor je belangstelling en vriendschap.
Lieve Inge, ons werk heeft veel parallellen. Het heeft me geholpen om van jou te horen hoe nuchter jij om ging met het verdelen van je tijd tussen werk en promotie en later ook toen je aan je nieuwe baan begon. Daarnaast is schaatsen en wielrennen met jou ook erg prettig om even aan andere dingen te denken dan werk. Ik hoop dat we daar nog lang mee doorgaan.
Lieve Esther, ik moet je veel te veel missen. Daarom ben ik blij dat je tijdens mijn verdediging naast me zal staan als paranimf. In Engeland zal jij niet een dergelijke feestelijke verdediging hebben ook al verdien jij dat natuurlijk wel! Ik

hoop dat jij ook van deze dag gaat genieten.
Lieve Pieter en Liesbeth, bedankt dat jullie altijd voor me klaar staan, jullie interesse en steun. Het was fijn dat ik vanuit het AMC altijd even langs kon fietsen en bij jullie kon zijn.
Lieve Conny, jij bent als vriendin en familie. Ik ben je dankbaar voor jouw scherpe redigeren, maar ook voor jouw steun, motivatie en liefde. Ik ben extra blij dat je speciaal voor mij uit Canada komt.
Lieve Roselien, je bent er voor me als ik je nodig hebt, zoals vandaag. Ik ben blij dat jij mijn paranimf bent. Ik ben altijd trots om naast mijn grote zus te staan.
Lieve pappa en mamma. Bedankt voor jullie onvoorwaardelijke steun en mogelijkheden. Jullie hebben mij geleerd om altijd te doen waar je zelf in geloofd ongeacht wat anderen daar van vinden. Dat is lang niet altijd makkelijk maar het brengt mij waar ik wil zijn. Ik ben blij dat jullie bij me zijn.
Lieve Dion, met jou bij me lukt alles. Bedankt dat je bent zoals je bent.

APPENDIX 5

Curriculum Vitea

Wynanda Annefloor van Enst werd op 24 januari 1986 in Arnhem geboren, als jongste in een gezin met twee dochters. Na een basisschool in Doorwerth en Zwolle ging zij naar het Atheneum+ aan de Thorbecke Scholengemeenschap te Zwolle. Zij volgende het Natuur en Gezondheid-profiel in combinatie met filosofie. Voor haar profielwerkstuk deed ze onderzoek naar het gebruik van cholesterolverlagende margarine. In 2004 begon zij met haar studie Algemene Gezondheidswetenschappen gevolgd door een Master in Lifestyle and Chronic disorders aan de Vrije Universiteit te Amsterdam. Hiervoor deed ze enkele onderzoeken naar chronische lage rugklachten. In 2009 startte ze bij het Dutch Cochrane Centre in het AMC te Amsterdam. Zij werkte als junior onderzoeker mee aan talloze systematische reviews en aan het onderwijs op gebied van systematische reviews en evidence-based medicine. Daarnaast deed zij promotieonderzoek naar empirische methoden voor systematische reviews en evidence-based medicine.
Sinds april 2014 is Annefloor werkzaam bij het Kennisinstituut van Medisch Specialisten te Utrecht en ondersteunt evidence-based richtlijnontwikkeling.

FINANCIAL CONTRIBUTORS

THE COCHRANE
COLLABORATION

kebb
klinische epidemiologie, biostatistiek & bioinformatica

Kennisinstituut
van Medisch Specialisten

NOTES