Strengthening methods of diagnostic accuracy studies

Ochodo, E.A.

Citation for published version (APA):

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

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: 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 (http://dare.uva.nl)
Chapter 6

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

Wynanda A. van Enst, Eleanor A. Ochodo, Rob Scholten, Lotty Hooft, Mariska M.G. Leeflang.

Accepted for publication by BMC Med Res Methodol
Abstract

Background: The validity of a meta-analysis can better be understood 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 clinical trials, leaving authors of diagnostic test accuracy (DTA) systematic reviews with limited guidance. The aim of this study was to investigate if and how publication bias is investigated in meta-analyses of DTA, and to compare the results of existing statistical methods to investigate 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. 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 test (n=4), and other (n=2). Our own comparison of the results of Begg, Egger and Deeks test for 91 meta-analyses of the included reviews indicated that up to 34% of the results did not correspond with each other.

Conclusions: 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. Authors could also refrain from testing as the principles for publication bias in DTA meta-analyses are not empirically investigated.
6.1 Background

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 contributes to the phenomenon of “small study-effects” [7]. This means that published trials 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].

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 for intervention reviews. A well-known graphical method is the funnel plot examination [9]. This tool is 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. 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 the test of Begg [10], and the test of Egger [11]. These methods are very well known and have been cited more than 2,500 (Begg) and 7,300 times (Egger) [12]. 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 a linear regression to assess the relation between the standardized effect estimates and the precision (Standard Error; SE). A significant result is for both of these tests an indication that the results might be affected by publication bias. These 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 studies 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 diseased and non-diseased patients are usually unequally divided depending on the setting of the study and the design of the study (cohort or case-control) which reduces the precision of a test accuracy estimate while in RCTs all participants are patients are divided into intervention or control by the research staff. Investigating whether meta-analyses of DTA studies have been influenced by publication bias in terms of small study-effects is debatable [13]. Moreover, the mechanisms that may induce publication bias in diagnostic studies or empirical evidence for their existence is scares, making it even more complex.

In 2002, Song and colleagues proposed that tests developed for intervention reviews, like the Begg and Egger method, could also be used to detect publication bias in DTA reviews [14]. They suggested to use the natural logarithm of the DOR (lnDOR) and plot it against its variance or SE and test for asymmetry [15]. 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. The draft Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy explicitly mentions not to use methods like the Begg or Egger test and argues that it is best to use the test proposed by Deeks [13]. This method 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 and non-diseased participants [16]. An unequal deviation between the numbers of diseased and non-diseased can lead to imprecision of the accuracy estimates. Using the ESS instead of sample size will reduce the unequal deviation and thereby enhance the precision of the accuracy estimates. The Cochrane Handbook, however, points out that this 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 [13].

The fact that little is known about publication bias in DTA studies and the various instructions with all pros and cons makes it hard for reviewers to know what to do best. One needs to choose if, and if so which test to use, but also the interpretation of the results and using them in formulating a conclusion can be challenging. To learn what is currently been done and if guidance is needed we aimed to review whether and how authors of DTA reviews investigated the possible threat of publication bias in their DTA reviews. We assessed which existing tests for publication bias were used and to what extent the results of these tests were incorporated in the conclusions of the review. Secondly, we applied existing methods for detection of publication bias in non-simulated data to assess if these methods provide similar results.

6.2 Methods

6.2.1 Study selection
MEDLINE was searched through the interface of PubMed for DTA 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 SPECIFICITY*[TW] OR "FALSE NEGATIVE"[TW] OR ACCURACY[TW]))) [17].

6.2.2 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 [18]. Therefore, the selection was further limited to reviews that included a meta-analysis. 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 study level.

6.2.3 Definitions of assessment of publication bias

In determining if authors would assess publication bias in their review, 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, or if they explicitly mentioned that they would assess 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 review when the authors described how publication bias might have affected the results of their review.

6.2.4 Data extraction

An online standardized data extraction form was used to extract data. We first piloted the form among all the team members. After all agreed on the data-extraction form, data 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).

For the first objective, data was extracted on all reported matters concerning testing for 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 involvement of test results in the conclusion. 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 somehow involved with the formulation of the discussion or 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).
6.2.5 Comparison of tests for publication bias

Different tests to identify publication bias in terms of small study-effects are expected to report different results. This has been shown in a simulation study [16], but not in real data. The secondary objective of this study was to assess if these methods provide similar results in real data. We applied three commonly used tests, Begg, Egger, and Deeks on the extracted two-by-two tables ourselves independently of the authors of the review had investigated publication bias. The use of Deeks test is what is currently recommended for DTA reviews by the Cochrane DTA methods group [18]. These test were performed accordingly:

- Begg: rank correlation of the lnDOR with variance of the lnDOR [10];
- Egger: linear regression of lnDOR with the standard error of lnDOR weighted by the inverse variance of lnDOR [11];
- Deeks: linear regression of lnDOR with 1/ESS^{1/2} weighted by the ESS [16].

Analyses were performed in the statistical program R [19]. We compared the p-values of the tests to each other. The level of significance was set at a p-value <0.05. We did not compare the results of our assessment of publication bias to the results reported in the reviews as they may have used different methods.

6.3 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.
Fifteen of these reviews (13.2%) mentioned why they did not study (62.7%) had performed methods to investigate publication bias in terms of small studies. 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.

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

---

**Figure 1:** Flow chart of the selection process and characters of the included studies
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 \( (n=1) \).

### 6.3.1 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 long cancer) and for different accuracy measures of a test \( (n=5) \) (e.g. sensitivity and specificity). In four reviews the authors made a comparisons of the accuracy of several clinical tests but used one single plot to investigate publication bias (two of these did however 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/root(ESS) \( (n=10) \), 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.

### 6.3.2 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 and Harbord test \( [20] \). One review did not specify which test was used. There number of studies included in the analyses for publication bias varied. Some review authors performed tests on
less than five studies [21-24] though two review authors mentioned a minimum of twenty homogeneous studies to perform a test [25,26].

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%)

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 each other, 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 [27] or imaging studies to detect osteomyelitis: Begg p=0.392 and p=Egger 0.063 [24]).

**Table 1.** Results of different tests for publication bias as reported in the reviews that applied tests (n=41).

<table>
<thead>
<tr>
<th>Type of Test</th>
<th>Publication bias</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Identified (%)</td>
<td>Not identified (%)</td>
<td>Total</td>
</tr>
<tr>
<td>Begg</td>
<td>3 (18.8)</td>
<td>13 (81.2)</td>
<td>16</td>
</tr>
<tr>
<td>Egger</td>
<td>16 (37.2)</td>
<td>27 (62.8)</td>
<td>43</td>
</tr>
<tr>
<td>Deeks</td>
<td>1 (6.7)</td>
<td>14 (93.3)</td>
<td>15</td>
</tr>
<tr>
<td>All tests</td>
<td>20 (27.0)</td>
<td>54 (73.0)</td>
<td>74</td>
</tr>
</tbody>
</table>

6.3.3 *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 improved 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. One review mentioned to have undertaken the review as the previous published review was likely to be influenced by publication bias (“This review was undertaken because in a previews review publication bias may have conflicted the results.”).

6.3.4 Comparison of tests to detect publication bias
We were able to obtain two by two tables for 52 reviews, including 92 different meta-analyses. There was moderate agreement between the significance of the different tests for publication bias. Reanalysis of the data to test for publication bias with the Begg, Egger and the Deeks test indicated an agreement of significant test results between 66% (Egger vs. Deeks test) and 87% (Begg vs. Egger test). Figures 2-4 present the (dis) agreement between the various tests. Begg’s test was significant in 22 meta-analyses (24%), Egger’s test in 23 meta-analyses (25%) and Deeks’ test in 14 meta-analyses (15%).
Figure 2. Comparison of the p-value produced by the Begg test (y-axis) compared to test by Deeks (x-axis) in 92 meta-analyses. The dotted lines indicate a p-value of 0.05. Thirteen tests have significant results with the Begg test but not with the Deeks test, and four tests have significant with the Deeks test but not with the Begg test. Agreement of 67% between tests.

Figure 3. Comparison of the p-value produced by the Egger test (y-axis) compared to the test by Deeks (x-axis) in 92 meta-analyses. The dotted lines indicate a p-value of 0.05. Fifteen tests have significant results with the Egger test but not with the Deeks test, and three tests have significant with the Deeks test but not with the Egger test. Agreement of 66% between tests.
Investigating publication bias in meta-analysis of test accuracy

Figure 4. Comparison of the p-value produced by the Begg test (y-axis) compared to the test by Egger (x-axis) in 92 meta-analyses. The dotted lines indicate a p-value of 0.05. Seventeen test results are in concordance. The Begg test has seven significant results that are not significant with the Egger test, and the Egger test has six tests significant results that are not significant for the Begg test. Agreement of 87% between tests.

6.4 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 twelve reviews (10.5%). In 18 reviews (15.8%), the results of the publication bias assessment lead to less confidence in the results. Our own evaluation of three tests to detect publication bias (Begg, Egger and Deeks) using real data indicated that the results of the tests could conflict with each other. Up to 34% of the test results were in disagreement (Egger's test compared to Deeks' test). The simulated data study of Deeks et al. showed that a type 1 error is likely to occur in both the Begg and the Egger test when there is a threshold, prevalence or accuracy effect, especially when the DOR is high.
(DOR>38), which is present in almost every DTA review [16]. Though, we cannot be sure in which reviews the test results were accurate and for which they were false, it seems like these two tests might have led to an overestimation of the effect of publication bias.

The number of reviews investigating publication bias seems to have increased over time. In 2002, Song and colleagues investigated how authors investigate 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 probability in the discussion [15]. 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 the reviews used statistical methods to assess publication bias [28]. Of the 50 diagnostic reviews that were included in this study, 9 (18%) used funnel plot asymmetry to investigate publications bias and in 3 (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.

The increased awareness of publication bias is a positive development, but the drawback here is that the majority of reviews use tests that are not fit for DTA meta-analyses. Our evaluation of 91 meta-analyses indicated that both the Begg and Egger test give more significant results than Deeks test. This result is in line with the expectation based on the simulation study by Deeks et al.

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.
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 [29]. There is no evidence regarding the existence of biases like language bias or time lag bias exist 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 either the sensitivity or the specificity are more affected by selective publication depending on the purpose of the test than the DOR, and tests for publication bias should perhaps be directed to these pertinent accuracy parameter. Either way, 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.

Empirical studies to assess and understand the mechanisms that may induce publication bias in DTA studies are needed. Prospective registration of intervention studies turned out to be an effective measure to reduce selective publication or at least make it more transparent to investigators. DTA research, however, is usually performed by observational studies and data are often collected as part of daily clinical care. For this type of study prospective registration is advocated 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) [30]. Mandatory prospective registration of diagnostic accuracy studies could help to understand better the process of selective publication of DTA studies and identify underlying mechanisms that can be used with interpretation of results of meta-analyses of diagnostic studies.

6.5 Conclusions
We advise authors to try to avoid introduction of publication bias by using thorough search methods to identify grey literature by contacting experts and search for conference proceedings, besides regular searches in electronic biomedical databases. The Begg test and Egger test developed to detect publication bias in intervention reviews, are not interchangeable with Deeks test to investigate publication bias in DTA reviews. These test give different results.
Chapter 6

The results of our evaluation of these tests correspond to the results of the simulation study of Deeks where publication bias could be modelled in to the analyses. This enlarges our confidence that Deeks methodology to construct funnel plots and to statistically test should be preferred in DTA meta-analyses. Interpretation of a significant test result should be done within the perspective that we are unaware whether publication bias exists for DTA studies. Authors could also refrain from testing as the principles for publication bias in DTA meta-analyses are not empirically investigated.

References


