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

Citation for published version (APA):
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 trials, 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)

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

† 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

<table>
<thead>
<tr>
<th>Accuracy measure</th>
<th>Relative increase per 100 participants</th>
<th>Q4 vs. Q3 (95% CI)</th>
<th>Q4 vs. Q2 (95% CI)</th>
<th>Q4 vs. Q1 (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Size</td>
<td>DOR§</td>
<td>1.01* (1.00 to 1.03)</td>
<td>1.23 (0.95 to 1.60)</td>
<td>1.15 (0.86 to 1.54)</td>
</tr>
<tr>
<td>Effective Sample Size‡</td>
<td>DOR‡</td>
<td>1.01 (0.78 to 1.30)</td>
<td>1.16 (0.88 to 1.53)</td>
<td>1.12 (0.81 to 1.55)</td>
</tr>
<tr>
<td>Number of diseased</td>
<td>Sensitivity</td>
<td>1.23* (1.09 to 1.39)</td>
<td>1.25* (1.02 to 1.53)</td>
<td>1.33* (1.07 to 1.66)</td>
</tr>
<tr>
<td>Number of non-diseased</td>
<td>Specificity</td>
<td>1.01 (0.99 to 1.05)</td>
<td>1.12 (0.92 to 1.37)</td>
<td>1.15 (0.91 to 1.46)</td>
</tr>
<tr>
<td>Time lag (years) ‡</td>
<td>DOR‡</td>
<td>0.99 (0.95 to 1.03)</td>
<td>0.87 (0.56 to 1.35)</td>
<td>0.85 (0.64 to 1.12)</td>
</tr>
<tr>
<td>Time lag (years) ‡</td>
<td>Sensitivity</td>
<td>0.99 (0.96 to 1.02)</td>
<td>0.84 (0.66 to 1.08)</td>
<td>0.92 (0.76 to 1.12)</td>
</tr>
<tr>
<td>Time lag (years) ‡</td>
<td>Specificity</td>
<td>1.01 (0.98 to 1.04)</td>
<td>0.98 (0.76 to 1.26)</td>
<td>0.86 (0.68 to 1.10)</td>
</tr>
</tbody>
</table>

* 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: \( \frac{4n_1 \cdot n_2}{n_1 + n_2} \)

‡ 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).

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

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

* 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

Imaging test

<table>
<thead>
<tr>
<th></th>
<th>Accuracy measure</th>
<th>Imaging test Q4 vs. Q1 (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size</td>
<td>DOR§</td>
<td>1.46 (0.87 to 2.45)</td>
</tr>
<tr>
<td>Effective sample size†</td>
<td>DOR§</td>
<td>1.72 (0.85 to 3.49)</td>
</tr>
<tr>
<td>Diseased</td>
<td>Sensitivity</td>
<td>1.96* (1.56 to 2.47)</td>
</tr>
<tr>
<td>Non-diseased</td>
<td>Specificity</td>
<td>1.32 (0.90 to 1.91)</td>
</tr>
<tr>
<td>Time lag (years)</td>
<td>DOR§</td>
<td>0.70* (0.52 to 0.93)</td>
</tr>
<tr>
<td>Time lag (years)</td>
<td>Sensitivity</td>
<td>0.93 (0.66 to 1.30)</td>
</tr>
<tr>
<td>Time lag (years)</td>
<td>Specificity</td>
<td>0.91 (0.73 to 1.13)</td>
</tr>
</tbody>
</table>

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
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

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.
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.


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.


27. Dickersin K, Chalmers I. Recognising, investigating and dealing with incomplete and biased reporting of clinical research: from Francis Bacon


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

**Figure 1**. Flow chart of selection process of the included reviews and meta-analyses.
APPENDIX 3. List of articles included in the review


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.


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


28. Romero J, Xue X, Gonzalez W, Garcia MJ. CMR imaging assessing viability


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:


44. van Teeffelen AS, Van Der Heijden J, Oei SG, Porath MM, Willekes C, Underwood M, Arbyn M, Redman C, Smith WP. Accuracy of colposcopic directed punch biopsies: A systematic review and meta-analysis.


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

