# UNIVERSITY<sup>OF</sup> BIRMINGHAM University of Birmingham Research at Birmingham

# The Moses-Littenberg meta-analytical method generates systematic differences in test accuracy compared to hierarchical meta-analytical models

Dinnes, Jacqueline; Mallett, Susan; Hopewell, Sally; Roderick, Paul; Deeks, Jonathan

DOI: 10.1016/j.jclinepi.2016.07.011

*License:* Creative Commons: Attribution-NonCommercial-NoDerivs (CC BY-NC-ND)

Document Version Peer reviewed version

Citation for published version (Harvard):

Dinnes, J, Mallett, S, Hopewell, S, Roderick, P & Deeks, J 2016, 'The Moses-Littenberg meta-analytical method generates systematic differences in test accuracy compared to hierarchical meta-analytical models', *Journal of Clinical Epidemiology*, vol. 80, pp. 77-87. https://doi.org/10.1016/j.jclinepi.2016.07.011

Link to publication on Research at Birmingham portal

Publisher Rights Statement: Checked 1/8/2016

#### **General rights**

Unless a licence is specified above, all rights (including copyright and moral rights) in this document are retained by the authors and/or the copyright holders. The express permission of the copyright holder must be obtained for any use of this material other than for purposes permitted by law.

•Users may freely distribute the URL that is used to identify this publication.

•Users may download and/or print one copy of the publication from the University of Birmingham research portal for the purpose of private study or non-commercial research.

•User may use extracts from the document in line with the concept of 'fair dealing' under the Copyright, Designs and Patents Act 1988 (?) •Users may not further distribute the material nor use it for the purposes of commercial gain.

Where a licence is displayed above, please note the terms and conditions of the licence govern your use of this document.

When citing, please reference the published version.

#### Take down policy

While the University of Birmingham exercises care and attention in making items available there are rare occasions when an item has been uploaded in error or has been deemed to be commercially or otherwise sensitive.

If you believe that this is the case for this document, please contact UBIRA@lists.bham.ac.uk providing details and we will remove access to the work immediately and investigate.

# Accepted Manuscript

The Moses-Littenberg meta-analytical method generates systematic differences in test accuracy compared to hierarchical meta-analytical models

Jacqueline Dinnes, Susan Mallett, Sally Hopewell, Paul J. Roderick, Jonathan J. Deeks

PII: S0895-4356(16)30258-X

DOI: 10.1016/j.jclinepi.2016.07.011

Reference: JCE 9216

To appear in: Journal of Clinical Epidemiology

Received Date: 26 February 2016

Revised Date: 18 June 2016

Accepted Date: 23 July 2016

Please cite this article as: Dinnes J, Mallett S, Hopewell S, Roderick PJ, Deeks JJ, The Moses-Littenberg meta-analytical method generates systematic differences in test accuracy compared to hierarchical meta-analytical models, *Journal of Clinical Epidemiology* (2016), doi: 10.1016/ j.jclinepi.2016.07.011.

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.



### The Moses-Littenberg meta-analytical method generates systematic differences in test

accuracy compared to hierarchical meta-analytical models

Jacqueline Dinnes<sup>a</sup>,

Susan Mallett<sup>a</sup>,

Sally Hopewell<sup>b</sup>,

Paul J. Roderick <sup>c</sup>,

Jonathan J. Deeks a\*

a. Institute for Applied Health Research, University of Birmingham, Edgbaston, Birmingham, UK B15 2TT.

b. Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford / Botnar Research Centre, Windmill Road, Oxford, UK OX3 7LD.

c. Department of Public Health Sciences and Medical Statistics, University of Southampton, Southampton General Hospital, Tremona Road, Southampton, UK, SO16 6YD.

\*Corresponding author. Phone: + 44 (0) 121-414-5328 E-mail: j.deeks@bham.ac.uk

#### Abstract

*Objective*: To compare meta-analyses of diagnostic test accuracy using the Moses-Littenberg summary ROC (SROC) approach with those of the hierarchical SROC (HSROC) model.

*Study design and setting*: Twenty-six datasets from existing test accuracy systematic reviews were reanalysed with the Moses-Littenberg model, using equal weighting ('E-ML') and weighting by the inverse variance of the logDOR ('W-ML'), and with the HSROC model. The diagnostic odds ratios (DOR) were estimated and covariates added to both models to estimate relative DORs (RDORs) between subgroups. Models were compared by calculating the ratio of DORs, the ratio of RDORs and P-values for detecting asymmetry and effects of covariates on DOR.

*Results*: Compared to the HSROC model, the Moses-Littenberg model DOR estimates were a median of 22% ('E-ML') and 47% ('W-ML') lower at Q\*, and 7% and 42% lower at the central point in the data. Instances of the ML-models giving estimates higher than the HSROC model also occurred. Investigations of heterogeneity also differed; the Moses-Littenberg models on average estimating smaller differences in RDOR.

*Conclusions*: Moses-Littenberg meta-analyses can generate lower estimates of test accuracy, and smaller differences in accuracy, compared to mathematically superior hierarchical models. This has implications for the usefulness of meta-analyses employing this approach. We recommend meta-analysis of DTA studies to be conducted using available hierarchical model based approaches.

Keywords: diagnostic test accuracy, meta-analysis, systematic review, hierarchical models, diagnostic odds ratio, summary ROC curves.

Running title: Comparison of meta-analytic approaches for diagnostic test accuracy data

Word count (excl Abstract, Tables and Figures): 3886

## What is new

- The Moses-Littenberg model for meta-analysing diagnostic test accuracy data on average generates lower estimates of test accuracy in comparison to the hierarchical summary ROC method for meta-analysis
- Substantial differences in results of investigations of heterogeneity between models are produced, both for estimates of the size of the effect and its statistical significance.
- Our findings support and extend those of a previous empirical comparison of methods, further raising concerns around the use of the results of Moses-Littenberg meta-analyses in clinical practice.
- Evidence for current recommendations for the preferential use of hierarchical models for diagnostic test meta-analyses is provided and should further encourage their uptake.

#### 1. Introduction

There is a considerable body of systematic reviews and meta-analyses of diagnostic test accuracy (DTA) in the public domain [1, 2]. Meta-analysis allows more precise estimation of test accuracy, can provide a stronger comparison of the accuracy of different tests compared to a single study, and allows the inevitable variability between studies to be quantified and formally investigated, as and when there is sufficient cause to suspect clinical or methodological variability [3, 4]. The statistical pooling of test accuracy studies presents an added level of complexity over and above that presented by trials of therapeutic interventions. Accuracy is usually quantified by two related statistics (sensitivity and specificity) rather than one, and meta-analysis must allow for the trade-off between the two. Approaches to DTA meta-analysis include separate pooling of sensitivity and specificity estimates, the linear regression approach to estimating summary ROC (SROC) curves developed by Moses and Littenberg (referred to here as Moses-Littenberg) [5, 6], and methods based on hierarchical models [7-11].

The theoretical limitations and advantages of the different meta-analytical approaches are well documented [12-14]. Whilst the Moses-Littenberg approach accounts for the bivariate nature of the data and the inverse correlation between sensitivity and specificity, it does not appropriately model within and between study variability to account for the uncertainty in the estimates, such that confidence intervals and hypotheses tests from the model are possibly invalid and summary operating points are hard to estimate. The hierarchical models preserve the bivariate nature of test accuracy data in terms of both model parameterisation and interpretation, accounting for the correlation between sensitivity and specificity; they appropriately weight studies to account for within study variability; and use a random effects approach to account for between-study variability. They also allow estimation of the diagnostic odds ratio (DOR) and an average operating point in terms of or sensitivity specificity pair, with associated confidence and prediction regions.

The hierarchical models have been recommended for DTA meta-analyses by the Cochrane Collaboration since 2007 [3]. However, the Moses-Littenberg model was used in 43% of 760 metaanalyses of diagnostic or predictive accuracy published between 1987 and 2009, and made up 86% of all meta-analyses that used ROC approaches to synthesis [15]. A survey of 100 DTA meta-

analyses published between September 2011 and January 2012 found that over half continued to pool studies using either simple linear regression-based SROC analysis or a univariate approach, and not the preferred hierarchical models [16]. Three-quarters of corresponding authors for 24 of the reviews using these 'traditional' approaches believed the approach to be "currently recommended" and 71% "believed that the method yielded precise estimates" [16]. Meta-analyses undertaken prior to this period are even more likely to have used the mathematically inferior models [1, 2].

Available empirical studies examining the impact of choice of meta-analytic model on overall results have reached conflicting conclusions, noting that hierarchical models produce different results to simpler models [14] or suggesting that conclusions may not differ [17, 18]. None of these studies considered possible effects on investigations of heterogeneity, such as looking at differences between tests or subgroups.

Given the widespread use of the Moses-Littenberg meta-analysis model there is a need to identify and quantify any potential differences in estimates of test accuracy and in investigations of sources of heterogeneity. The complexity of the hierarchical models and their omission from commonly used meta-analytical software programs [19] further compounds the need to determine the degree to which results might be affected by the choice of approach. Our objective was to empirically compare the Moses-Littenberg and hierarchical SROC (HSROC) approaches for the overall synthesis of DTA data and for the investigation of sources of heterogeneity using a large sample of systematic reviews.

#### 2. Methods

## 2.1. Identification, extraction and analysis of systematic review data

Existing test accuracy systematic reviews were identified from the Database of Abstracts of Reviews of Effects (DARE), the HTA database of the Cochrane Library and the MEDION database of systematic reviews. The cohort we analysed has been partially reported on previously [1], and includes reviews published between 2000 and 2005 that presented sufficient information to allow the construction of a 2x2 contingency table comparing a test to a reference test for at least five primary studies. Systematic reviews had to include at least five studies and report at least one study-level

covariate. One reviewer assessed whether reviews met these criteria. All subgroups in the dataset contained at least 3 studies.

Data on the experimental test and target disorder, the 2x2 contingency table data and data on any potential spectrum-related sources of heterogeneity per study were extracted independently by two reviewers. Disagreements were resolved by consensus or referral to a third reviewer.

#### 2.2 Meta-analytical methods

Three alternative meta-analytical methods were considered, two formulations of the Moses-Littenberg model with different study weighting, and the HSROC model.

The Moses-Littenberg model was fitted as originally described [5, 6]. The approach involves computing the logit-transforms of sensitivity and the complement of specificity and undertaking a linear regression of the difference between them (D, or the log of the DOR) on the sum (S, a measure of the proportion that are test positive, which is a proxy for test threshold) to estimate an intercept parameter  $\alpha$  and a slope parameter  $\beta$ . A summary ROC curve is derived from the fitted regression line. The test of the significance of the slope  $\beta$  is used to determine whether there is evidence that the DOR varies with the threshold parameter S, which implies that the summary ROC curve has an asymmetric shape. Differences between subgroups in the log DOR can be investigated by adding indicator variables as covariates. The parameter estimates for these are the log relative DOR (RDOR) comparing the DOR between the subgroups. Inclusion of interactions between each covariate and the S parameter allows for different shaped curves in each subgroup. We fitted models with and without the interaction term. The regression intercept term estimates the log of the DOR at the Q\* point, where sensitivity is equal to specificity. We used the DOR value at this point in accuracy comparisons, but as it may lie outside the results of the observed studies we also made comparisons of the DOR estimated at the mean value of S which will be central to the observed study results. We fitted the Moses-Littenberg model giving equal weight to all studies (denoted 'E-ML'), and weighting each study (denoted 'W-ML') by the inverse variance of the log DOR, in accordance with standard practice [12]. A standard zero cell correction (the addition of 0.5 to every cell of a 2x2 table that contains at least one zero) was used to avoid divide by zero errors.

Two versions of the hierarchical model are commonly used for meta-analysis, the HSROC and the bivariate models, which have been shown to be mathematically equivalent for single test metaanalyses but to estimate different parameters in investigations of heterogeneity [14]. We chose the HSROC model for comparison with the Moses-Littenberg models as both estimate RDOR in heterogeneity investigations. We fitted the HSROC model proposed by Rutter and Gatsonis [7, 8] which is based on a latent scale logistic regression model [20, 21]. The HSROC model assumes that there is an underlying ROC curve in each study with parameters  $\alpha$  and  $\beta$  which characterize the accuracy and asymmetry of the curve, in a similar way to the  $\alpha$  and  $\beta$  parameters in the linear regression method of Moses and Littenberg. The parameter  $\alpha$  estimates log DOR at the Q<sup>\*</sup> point and the significance of the shape term  $\beta$  can be used to assess whether a symmetric or asymmetric curve best fits to observed data. The model has a third parameter  $\theta$  related to the proportion test positive – its value is equivalent to the values of S/2 in the Moses-Littenberg model. An average value for  $\theta$  is estimated by the model, which allows identification of an average operating point on the summary ROC curve central to the observed data. The model is fitted at two levels. At the first level the proportions test positive in the reference standard positive and negative groups are modelled assuming they follow a binomial distribution, whereas at the second level variation between studies is modelled assuming normal distributions for log DOR ( $\alpha$ ) and the positivity threshold parameter ( $\theta$ ). To investigate heterogeneity using the HSROC model, we fitted parallel SROC curves by adding the covariate as a term to both the accuracy and threshold parts of the model (such that the shape of the SROC curves is determined using the whole set of studies). We also fitted models that included a covariate as a third term in the shape part of the model, to allow for variation in differences in log DOR with threshold (allowing the SROC curves to have different shapes; non-parallel).

#### 2.3. Comparison of meta-analytic models

We compared the performance of each Moses-Littenberg regression model with the results of the HSROC model first for a simple meta-analysis by:

 i) comparing DOR estimates by computing the ratio of DOR from the Moses-Littenberg models compared to the HSROC model at the central point in the data (which is the equivalent of the average operating point) and at the Q\* value (the point where sensitivity=specificity) (sample SROC curves from each model shown in Box 1);  ii) comparing the detection of SROC curve asymmetry between the Moses-Littenberg and HSROC models by comparing P-values from Wald tests for the β term in both models;

and then in investigations of heterogeneity contrasting two subgroups

- iii) comparing estimates of the RDOR between the Moses-Littenberg models and the HSROC
  model again at the central point in the data and at Q\*; and
- iv) comparing the significance of the difference in DOR between the models estimated by comparing P-values from Wald tests for the log RDOR terms in both models.

For the Moses-Littenberg model we identified the central point in the data as being at the mean value of S, whilst for the HSROC model we identified it as being at the average operating point. For the comparison of subgroups, the RDOR can be less than or greater than one; summary statistics were standardised to code the subgroups such that the HSROC model always estimated an RDOR greater than one.



E-ML – equal weight Moses-Littenberg model; W-ML – Moses-Littenberg model weighted by inverse variance of D; HSROC – hierarchical SROC model; DOR – diagnostic odds ratio; Q\* - DOR plotted at point on SROC curve where sensitivity=specificity; average operating point – DOR point on the SROC curve near to the centre of the data using mean threshold.

We additionally investigated how differences in DOR between the models vary according to predefined aspects of a) magnitude of accuracy, b) prevalence of zero cells, c) variation in threshold (based on values for 'S' from the Moses-Littenberg model).

The Moses-Littenberg analyses were performed in STATA and the HSROC model was carried out using the PROC NLMIXED command in SAS.

#### 3. Results

#### 3.1. Search results

The search identified 97 systematic reviews of test accuracy that presented sufficient data to complete 2x2 contingency tables per study (Figure 1). Information on at least one covariate per study was presented in 29 reviews, however the HSROC model would not converge for three of the datasets (occasional non-convergence of the HSROC model, particularly when there are few studies or when all studies sit on one of the boundaries of SROC space is a recognised phenomenon [23]). The comparisons between models are therefore based on 26 datasets with a total of 55 spectrum-related covariate investigations (Supplementary Table 1). The HSROC model could not be completed for 9 covariate investigations (for one parallel curve SROC comparison, for four non-parallel curve SROC comparisons and for four covariates using both parallel and non-parallel SROC curves), either due to insufficient numbers of studies in at least one of the subsets (for 5 of the 9 covariates) or the studies exhibited exceptionally high specificities with varying sensitivities. The median number of studies per review was 16 (IQR 12, 26); median sample sizes of studies within each review ranged from 20 to 7575.

#### 3.2. Comparison of diagnostic odds ratios

Estimates of the DOR from the Moses-Littenberg methods were on average lower than those of the HSROC model. Evaluated at the Q\* point, the 'E-ML' model estimates of the DOR were a median of 22% lower (IQR 49% lower to 2% higher) than those from the HSROC model, whilst estimates from the 'W-ML' model were a median of 47% lower (IQR 76% lower to 28% lower). Differences between models were smaller at the central threshold, but still showed lower estimates for the Moses-

Littenberg models on average: 7% lower (IQR 32% lower to 4% higher) for 'E-ML' and 42% lower (IQR 64% lower to 22% lower) (Figure 2a).

We categorized the meta-analyses according to their DOR estimate from the HSROC model, the percentage of studies with zero cells in 2x2 tables, and the range of the threshold parameter S (Table 1). We noted greater discrepancy between the methods when DORs were high and with increasing percentages of zeros in 2x2 tables. There was no clear relationship with the range of the threshold parameter.

A wide range in results was observed for all model comparisons with both higher and lower estimates for Moses-Littenberg models compared to HSROC models (Figure 2b). The five reviews responsible for most of the extreme differences of DOR all had studies with very high or close to perfect sensitivity [24, 25] and/or specificity [26-28] (Supplementary Figure 1). The lack of data points near to Q\* for these reviews illustrates the unreliability of DORs estimated at this point. The three datasets that produced extreme values in all analyses were removed in a sensitivity analysis [24, 25, 28]; differences in DOR estimates between Moses-Littenberg and HSROC models remained for both the 'E-ML' and 'W-ML' model comparisons (the overall range in RORs being 0.48 to 1.20 and 0.24 to 1.13, respectively) (analysis not shown).

#### 3.3. Comparison of SROC curve shape

Figure 3 shows broad but imperfect agreement between the HSROC and Moses-Littenberg models in terms of tests of the asymmetry of the SROC curve. Taking P<0.20 as providing moderate to strong evidence of curve asymmetry, SROC curve asymmetry was identified for 14 of the 26 datasets using the HSROC model and 12 datasets for each of the Moses-Littenberg models, with 4 and 3 reviews showing disagreement for the 'E-ML' and 'W-ML' models respectively.

#### 3.4. Comparison of estimates of RDOR in heterogeneity investigations

Estimates of RDOR from heterogeneity investigations were on average considerably lower when estimated by the Moses-Littenberg models compared to the HSROC model (Figure 4). When estimated assuming parallel SROC curves, RDOR values (median [IQR]) were 13% lower [42% lower

to 2% higher] for the 'E-ML' model and 20% lower [46% lower to 5% higher] for the 'W-ML' model. Including interaction terms to allow for non-parallel SROC curves substantially increased both the average difference and the range of differences when evaluated both at the Q\* point and the central point.

#### 3.4. Comparison of the statistical significance of sources of heterogeneity

There was poor agreement in the statistical significance of the covariate investigations between Moses-Littenberg and HSROC models with the P-value comparison points scattered widely around the plots (Figure 5). Disagreements appeared worse for the models with non-parallel SROC curves.

#### 4. Discussion

We found that the simpler Moses-Littenberg models produced lower estimates of DOR accuracy in comparison with the more complex HSROC model, with median differences of 22% and 47% for unweighted and weighted models at Q\*, but with wide inter-quartile and overall ranges for the differences. The greatest differences were observed when the Moses-Littenberg model was weighted by the inverse variance of InDOR, where the overall pooled DOR was over 100, and where there were a high proportion of zero cells; situations where the mathematical limitations of Moses-Littenberg method creates misleading estimates.

The weighted Moses-Littenberg model uses the approximate asymptotic standard error of the log DOR in allocating study weights. We have previously shown (in the context of funnel plots and tests for publication bias [29]) that the estimate of the standard error depends on the In DOR for values of DOR which are greater than one, leading funnel plot based tests such as the Begg and Egger tests to overestimate the frequency of sample size related effects. The same relationship will affect the estimates of standard errors for larger DOR, most often leading to overestimation of standard errors at higher DOR, which in an inverse variance weighted meta-analysis will lead to underestimation of overall accuracy.

Zero-cell corrections are used in the Moses-Littenberg model when sensitivities or specificities of 100% are observed, also when the DOR is typically high. Adding 0.5 to each cell of every 2x2 contingency table that contains zero false positives or false negatives always creates a downward

11

effect on the estimate of the DOR. The magnitude of effect will be greater in studies with smaller sample sizes in the disease positive and disease negative groups.

There was some disagreement in detection of SROC curve asymmetry between models, but the asymmetry detected was rarely statistically significant with either model. Many of our analyses contained few studies thus few of the meta-analyses will have had enough power to detect asymmetry.

Of importance, we observed substantial differences for investigations of heterogeneity, both in terms of the estimates of the size of the effect (RDOR) and their statistical significance. In line with their performance for simple test meta-analyses, Moses-Littenberg methods on average generated lower between group differences in accuracy compared to the HSROC model. The failure of the ML models to account for both within and between study variability compromises the models' ability to estimate standard errors used to assess statistical significance, leading to a notable lack of agreement in the statistical significance of relative diagnostic odds ratios. Inferences drawn from covariate investigations using the Moses-Littenberg models can therefore potentially be highly spurious. These findings have implications for meta-analyses that compare tests as well as those which investigate the performance of a single test in subgroups of studies.

Our results support and extend those of previous empirical comparisons (one using a very large dataset [15]) showing that SROC curves produced by the simpler Moses-Littenberg model can appear similar to those produced by more rigorous methods but can in some cases diverge [14, 15]; neither of these comparisons compared average operating points on the SROC curves however, and neither compared the results of heterogeneity investigations. Whilst our dataset has shown that although in some circumstances there may be little difference on average between methods, the variability observed around the average indicates the potential for large differences in test accuracy, both in the overall result of the meta-analysis and, perhaps more crucially, with the addition of covariates to the models.

Our comparison of models is subject to a number of limitations. Studies in our cohort were published between 2000 and 2005 with screening for inclusion conducted by only one reviewer. However, given that our purpose was to illustrate the potential effect across a number of datasets rather than to reach

12

a conclusion around the accuracy of a given test, the potential for reviewer bias is limited. Furthermore the currency of the review cohort has no impact on the statistical models used or external validity of our results as there is no reason to consider that the relationships between the methods would differ in more recent studies. In terms of the statistical analyses, heterogeneity was common and the number of studies per meta-analysis were typical but low, with a median of 16 (range 7 to 46). These factors will have contributed to some of the extreme results observed but many review authors will nevertheless make the decision to meta-analyse studies in these circumstances.

Furthermore, we focused on comparing DORs rather than sensitivity and specificity estimates as the DOR is the natural parameter of both the Moses-Littenberg and HSROC models. This facilitated the comparison between models to be illustrated with a single parameter, however the DOR is not as intuitive as other measures of test accuracy. When we looked at summary estimates of sensitivity and specificity at the average operating point and at Q\*, the ML model produced at least one estimate in the order of 5% to 9% lower than the HSROC model in 12 of the 26 reviews, and 10% to 23% lower than the HSROC model for 8 of the 26 reviews. Only one review produced an estimate that was higher than that observed from the HSROC model (for the E-ML estimate of sensitivity). Although the differential effects on sensitivity and specificity vary from case to case, the directional 'bias' that we observed for the ML model in the DOR comparisons was supported by the comparisons of sensitivities and specificities, with differences large enough to mislead.

The strength of this study was the number of datasets available for reanalysis. Further investigation, for example using simulated data, might be able to identify circumstances under which the Moses-Littenberg methods more closely approximate those of the HSROC method. Whilst there are more limited software options for fitting hierarchical models than the Moses-Littenberg models, requiring programs that fit mixed linear (or non-linear for the HSROC) logistic models, support for fitting these models in SAS, STATA, R and WinBUGS, is now available such that they should be the preferred approach [30, 31]. Whilst convergence problems with the HSROC and bivariate model are encountered when there are few studies, fitting HSROC models with fewer parameters (e.g. setting variance parameter estimates to zero, or assuming symmetrical SROC curves) has been found to be a preferable approach .

13

We have provided empirical evidence that supports current recommendations advocating the use of hierarchical approaches to the statistical synthesis of test accuracy data [30, 31]. Ongoing efforts to extend the scope of the models [32-34] and to make them more accessible [35-37] should help to increase their uptake in future systematic reviews.

These findings raise concerns about the clinical use of existing reviews that have employed the Moses-Littenberg approach. We have demonstrated that not only might the pooled results be misleading, but conclusions drawn from models where covariates have been added may also be unreliable. This finding applies not only to investigations of heterogeneity but also to comparisons of the accuracy of two or more tests where test covariates are added to the models in the same way as for sources of heterogeneity. The problem may not be considered as serious where investigations are considered exploratory, but is of real concern where comparisons between tests are made with the aim of informing the selection of tests for use in clinical practice. Our results also potentially call into question the conclusions from seminal meta-epidemiology papers that have employed the Moses-Littenberg model to provide the basis for our understanding of the biases in operation in test accuracy studies [2, 38, 39].

There are a huge number of systematic reviews of diagnostic tests now in the public domain, many of which have employed less than optimal approaches to the synthesis of test accuracy data. We have demonstrated the potential for drawing misleading inferences from meta-analyses adopting the Moses-Littenberg model and recommend that caution is used when considering their use. Our findings support and encourage the future uptake of hierarchical models for diagnostic test meta-analyses.

Acknowledgements: Dr Dinnes was supported by an NIHR - Research Scientist in Evidence Synthesis Award. Professor Deeks is supported by an NIHR Senior Investigator Award.

## Reference list

1. Dinnes J, Deeks JJ, Kirby J, Roderick P. A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy. Health Technol Assess. 2005;9.

2. 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.

3. Cochrane Handbook for Systematic Reviews of Diagnostic Tests. The Cochrane Library. Issue 1 ed. Chichester, UK: John Wiley & Sons, Ltd.; 2008.

4. Deeks JJ, Higgins JP, Altman DG, on behalf of the Cochrane Statistical Methods Group. Chapter 9: Analysing data and undertaking meta-analyses. In: Higgins JP, Green S, editors. Cochrane Handbook for

Systematic Reviews of Interventions. Version 5.1.0 [updated March 2011]. ed. Available from <u>www.cochrane-handbook.org</u>. : The Cochrane Collaboration; 2011.

5. Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. Med Decis Making. 1993;13:313-21.

6. 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;12:1293-316.

7. Rutter CM, Gatsonis CA. Regression methods for meta-analysis of diagnostic test data. Acad Radiol. 1995;2:s48-s56.

8. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med. 2001;20:2865-84.

9. van Houwelingen HC, Zwinderman KH, Stijnen T. A Bivariate Approach to Metaanalysis. Stat Med. 1993;12:2273-84.

10. van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. Stat Med. 2002;21:589-624.

11. Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA. A unification of models for meta-analysis of diagnostic accuracy studies. Biostatistics. 2007;8:239-51.

12. Deeks JJ, Altman DG, Bradburn MJ. Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis. In: Egger M, Davey Smith G, Altman D, editors. Systematic reviews in health care: Meta analysis in context. 2nd ed. London: BMJ Books; 2001. p. 285-312.

13. Zwinderman AH, Bossuyt PM. We should not pool diagnostic likelihood ratios in systematic reviews. Statistics in Medicine. 2008;27:687-97.

14. Harbord RM, Whiting P, Sterne JA, Egger M, Deeks JJ, Shang A, et al. An empirical comparison of methods for meta-analysis of diagnostic accuracy showed hierarchical models are necessary. J Clin Epidemiol. 2008;61:1095-103.

15. 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. Rockville (MD): Agency for Healthcare Research and Quality (US), 2012 Contract No.: 12-EHC044-EF.

16. Ochodo EA, Reitsma JB, Bossuyt PM, Leeflang MM. Survey revealed a lack of clarity about recommended methods for meta-analysis of diagnostic accuracy data. J Clin Epidemiol. 2013;66:1281-8.

17. Simel DL, Bossuyt PM. Differences between univariate and bivariate models for summarizing diagnostic accuracy may not be large. J Clin Epidemiol. 2009;62:1292-300.

18. Dahabreh IJ, Trikalinos TA, Lau J, Schmid C. An Empirical Assessment of Bivariate Methods for Meta-Analysis of Test Accuracy. Rockville (MD): Agency for Healthcare Research and Quality (US), 2012 Contract No.: No 12(13)-EHC136-EF.

19. Zamora J, Abraira V, Muriel A, Khan K, Coomarasamy A. Meta-DiSc: a software for meta-analysis of test accuracy data. BMC Medical Research Methodology. 2006;6:1-12.

20. McCullagh P. Regression models for ordinal data. Journal of the royal statistical society Series B (Methodological). 1980:109-42.

21. Tosteson ANA, Begg CB. A general regression methodology for ROC curve estimation. Med Decis Making. 1988;8:204-15.

22. Scheidler J, Hricak H, Yu KK, Subak L, Segal MR. Radiological evaluation of lymph node metastases in patients with cervical cancer. A meta-analysis. JAMA. 1997;278:1096-101.

23. Takwoingi Y, Guo B, Riley RD, Deeks JJ. Performance of methods for meta-analysis of diagnostic test accuracy with few studies or sparse data. Statistical Methods in Medical Research. 2015;(DOI: 10.1177/0962280215592269 PMID: 26116616).

24. Nallamothu BK, Saint S, Bielak LF, Sonnad SS, Peyser PA, Rubenfire M, et al. Electron-beam computed tomography in the diagnosis of coronary artery disease: a metaanalysis. Arch Intern Med. 2001;161:833-8.

25. Dijkhuizen FP, Mol BW, Brolmann HA, Heintz AP. The accuracy of endometrial sampling in the diagnosis of patients with endometrial carcinoma and hyperplasia: a meta-analysis. Cancer. 2000;89:1765-72.

26. Bricker L, Garcia J, Henderson J, Mugford M, Neilson J, Roberts T, et al. Ultrasound screening in pregnancy: a systematic review of the clinical effectiveness, cost-effectiveness and women's views. Health Technol Asses. 2000;4.

27. Eden K, Mahon S, Helfand M. Screening high-risk populations for thyroid cancer. Med Pediatr Oncol. 2001;36:583-91.

28. Medical Services Advisory Committee. Genetic test for fragile X syndrome. Assessment report. Australia: 2002 2002. Report No.: MSAC application 1035.

29. 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;58:882-93.

30. Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM. Systematic reviews of diagnostic test accuracy. Ann Intern Med. 2008;149:889-97.

31. Macaskill P, Gatsonis CA, Deeks JJ, Harbord RM, Takwoingi Y. Chapter 10: Analysing and Presenting Results. In: Deeks JJ, Bossuyt PM, Gatsonis CA, editors. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. Issue 1 ed. Available from: <u>http://srdta.cochrane.org/:</u> The Cochrane Collaboration; 2010.

32. Hamza TH, van Houwelingen HC, Heijenbrok-Kal MH, Stijnen T. Associating explanatory variables with summary receiver operating characteristic curves in diagnostic meta-analysis. J Clin Epidemiol. 2009;62:1284-91.

33. Leeflang MMG, Deeks JJ, Rutjes AWS, Reitsma JB, Bossuyt PMM. Bivariate metaanalysis of predictive values of diagnostic tests can be an alternative to bivariate metaanalysis of sensitivity and specificity. Journal of Clinical Epidemiology. 2012;65:1088-97.

34. Eusebi P, Reitsma JB, Vermunt JK. Latent class bivariate model for the meta-analysis of diagnostic test accuracy studies. BMC Med Res Methodol. 2014;14:88.

35. Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. Journal of Clinical Epidemiology. 2005;58:982-90.

36. Macaskill P. Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. J Clin Epidemiol. 2004;57:925-32.

37. Arends LR, Hamza TH, van Houwelingen JC, Heijenbrok-Kal MH, Hunink MG, Stijnen T. Bivariate random effects meta-analysis of ROC curves. Med Decis Making. 2008;28:621-38.

38. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JHP, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA. 1999;282:1061-6.

39. Lijmer JG, Bossuyt PM, Heisterkamp SH. Exploring sources of heterogeneity in systematic reviews of diagnostci tests. Stat Med. 2002;21:1525-37.

Chillip Mark

	Comparison of DORs at mean threshold*	
Number of reviews	E-ML vs HSROC	W-ML vs HSROC
	median ROR [IQR]	median ROR [IQR]
Overall (n=26)	0.94 [0.68, 1.04]	0.59 [0.46, 0.78]
by size of $DOR^\dagger$		
DOR < 35, n=9	0.99 [0.88, 1.16]	0.78 [0.60, 0.95]
DOR 35-100, n= 10	1.00 [0.83, 1.05]	0.53 [0.42, 0.69]
DOR > 100, n= 7	0.62 [0.30, 0.75]	0.42 [0.16, 0.59]
by % zero cells <sup>‡</sup>		
<5%, n=10	1.01 [0.88, 1.05]	0.79 [0.53, 0.88]
5 to 10%, n=8	1.02 [0.94, 1.11]	0.63 [0.50, 0.74]
>10%, n=8	0.62 [0.30, 0.72]	0.42 [0.16, 0.63]
by range in 'S'∫		
range 3 to <6, n=7	0.82 [0.48, 1.03]	0.75 [0.37, 0.79]
range 6 to <8, n=14	0.90 [0.68, 1.00]	0.53 [0.49, 0.75]
range ≥8, n=5	1.05 [0.99, 1.16]	0.57 [0.42, 0.73]

#### Table 1 Stratified comparison of diagnostic odds ratio (DOR) estimates

DOR – diagnostic odds ratio; E-ML – equal weight Moses-Littenberg model; W-LM – weighted Moses-Littenberg model; ROR - ratio of DORs between models; median ROR – ROR at the median; IQR – inter quartile range in ROR, from 25th percentile to 75<sup>th</sup> percentile

\* Each Moses-Littenberg model is compared to the HSROC model (denominator)

<sup>†</sup> The stratification by DOR is based on the HSROC overall pooled estimate at mean threshold

<sup>‡</sup> number of zero false positive and false negative cells as a percentage of the total number of cells per analysis

 $\int$  based on values for 'S' from Moses-Littenberg model, where S = logit(sensitivity) + logit(1 - specificity)

## Figure 1 Flowchart of the review selection process



Figure 2 Comparison of diagnostic odds ratios: ML models versus HSROC model





b. Scatter plot of ratio of DORs between models (HSROC model estimate as reference)



E-ML – equal weight Moses-Littenberg model; W-ML – Moses-Littenberg model weighted by inverse variance of D; HSROC – hierarchical SROC model; DOR – diagnostic odds ratio; Q\* - point on SROC curve where sensitivity=specificity; mean threshold – point on the SROC curve near to the centre of the data; max – maximum ratio of DORs; p75 – ratio of DORs at the 75<sup>th</sup> percentile; p50 – ratio of DORs at the 50<sup>th</sup> percentile (median); p25 – ratio of DORs at the 25<sup>th</sup> percentile; min – minimum ratio of DORs; [] denotes reference number for five reviews responsible for most of the extreme differences of DOR









Figure 4 Ratio of relative DORs (RDORs) between models (HSROC model estimate as reference)

DOR – diagnostic odds ratio; RDOR – relative diagnostic odds ratio; E-ML – equal weight Moses-Littenberg model; W-ML – Moses-Littenberg model weighted by inverse variance of D; HSROC – hierarchical SROC modelparallel curves – R DORs between study subgroups estimated assuming SROC curves have same shape (parallel); non-parallel curves - R DORs between study subgroups estimated allowing SROC curves to have different shapes (non-parallel); Q\* - point on SROC curve where sensitivity=specificity; reference group threshold – RDOR estimated at the point on the SROC curve near to the centre of the data in the reference group; max – maximum ratio of DORs ; p75 – ratio of DORs at the 75<sup>th</sup> percentile; p50 – ratio of DORs at the 50<sup>th</sup> percentile (median); p25 – ratio of DORs at the 25<sup>th</sup> percentile; min – minimum ratio of DORs;



### Figure 5 Comparison of tests of statistical significance for difference in accuracy

E-ML – equal weight Moses-Littenberg model; W-ML –Moses-Littenberg model weighted by inverse variance of D; HSROC – hierarchical SROC model