Guidance for deriving and presenting percentage study weights in meta-analysis of test accuracy studies

Danielle L. Burke,[1*] Joie Ensor,[1] Kym IE. Snell,[1] Danielle van der Windt,[1] Richard D. Riley[1]

Contact details:

* corresponding author:

Email: d.burke@keele.ac.uk;

Tel: +44 (0) 1782 734894

[1] Research Institute for Primary Care and Health Sciences, Keele University, Staffordshire, ST5 5BG

Acknowledgements

We would like to thank two anonymous reviewers for their constructive feedback on how to improve the article.

Funding

Danielle Burke is funded by an NIHR School for Primary Care Research Post-Doctoral Fellowship. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/jrsm.1283

ABSTRACT

Percentage study weights in meta-analysis reveal the contribution of each study toward the overall summary results, and are especially important when some studies are considered outliers or at high risk of bias. In meta-analyses of test accuracy reviews, such as a bivariate meta-analysis of sensitivity and specificity, the percentage study weights are not currently derived. Rather the focus is on representing the precision of study estimates on ROC plots by scaling the points relative to the study sample size or to their standard error. In this article, we recommend that researchers should also provide the percentage study weights directly, and we propose a method to derive them based on a decomposition of Fisher's information matrix. This method also generalises to a bivariate meta-regression, so that percentage study weights can also be derived for estimates of study-level modifiers of test accuracy. Application is made to two meta-analyses examining test accuracy: one of ear temperature for diagnosis of fever in children; and the other of positron emission tomography for diagnosis of Alzheimer's disease. These highlight that the percentage study weights provide important information that is otherwise hidden if the presentation only focuses on precision based on sample size or standard errors. Software code is provided for Stata, and we suggest that our proposed percentage weights should be routinely added on forest and ROC plots for sensitivity and specificity, to provide transparency of the contribution of each study towards the results. This has implications for the PRISMA-DTA guidelines that are currently being produced.

KEY WORDS:

Percentage study weight; bivariate meta-analysis, diagnostic test accuracy; Fisher's information.

RUNNING TITLE:

Percentage study weight in test accuracy meta-analysis

WORD COUNT:

4471

1 Introduction

Meta-analysis is the statistical synthesis of evidence from multiple studies to produce overall pooled results that can aid decision making. The PRISMA statement for reporting the results of a meta-analysis says that one must report the summary meta-analysis result and forest plot, and that, "…it is preferable also to include, for each study … the percentage weight" (Moher et al. 2009). Percentage study weights quantify the relative contribution of each study to the pooled meta-analysis result.

Study weights in standard univariate meta-analysis (e.g. of treatment effects) are well-known and reflect precision of the study estimates included in the meta-analysis: in a fixed effect model, they are inversely proportional to the within-study variances, and in a random effects model, they are inversely proportional to the sum of the within-study variance and the estimated between-study variance. In two recent articles, the study weights in more complex scenarios have been derived. Jackson et al. (2015) derived study weights in two-stage multivariate meta-analyses, then Riley et al. 2016 generalised these to any meta-analysis that involves a multi-parameter model, such as a one-stage meta-analysis model that uses individual participant data (IPD), or a meta-regression model. However, neither paper illustrated the importance of deriving the percentage study weights in meta-analysis of diagnostic test accuracy (DTA) studies, and we believe this deserves special attention in advance of the PRISMA guidelines being extended to DTA studies.

A meta-analysis of DTA studies often involves a bivariate meta-analysis of sensitivity and specificity (Rutter and Gatsonis 2001, Reitsma et al. 2005, Chu and Cole 2006, Harbord et al. 2007), and may be extended to a meta-regression to compare the accuracy of different tests. In such models, the derivation of percentage study weights is not immediately obvious, but is needed to reveal the contribution of each study toward the meta-analysis results of interest. This is especially important when some studies are considered to be outliers, at high risk of bias (Whiting et al. 2011), or in some sense less relevant (e.g. due to their more heterogeneous inclusion criteria). Currently, for example in Cochrane reviews, percentage study weights are not routinely displayed following a bivariate meta-analysis. Thus, the contribution of each study cannot be immediately ascertained by the reader.

Sometimes, a plot of points in the receiver operating characteristic (ROC) space is used to display the results of meta-analyses of DTA studies as an alternative to (or in addition to) forest plots. The visual display of study estimates in ROC space in RevMan (The Cochrane

Collaboration 2014) and Stata (StataCorp 2015) is to scale them relative to the standard error of the study-specific estimates of logit sensitivity and specificity, or relative to the sample size, in order to reflect precision of study estimates. However, relative precision of study estimates and percentage study weights are not the same thing, and may differ considerably. For example, Irwig et al. explained in the context of the diagnostic odds ratio that, "…weighting by the inverse of the estimated study-specific variance is inappropriate, as it is easily shown at equivalent sample sizes to give far more weight to studies which appear to show poorer accuracy" (Irwig et al. 1995). Therefore, in addition to providing a visualisation of precision of each study within a DTA meta-analysis (e.g. via study-specific confidence intervals on forest plots and circle size on summary ROC plots), it is important to report and display the percentage study weights.

In this article we propose how to derive and present percentage study weights in bivariate meta-analysis and meta-regression of DTA studies. We adopt a proposal for percentage study weights in multi-parameter meta-analysis models (Riley et al. 2017) based on a decomposition of Fisher's information matrix. This enables percentage study weights to be derived and presented appropriately for summary estimates of sensitivity and specificity, and for estimates of study-level modifiers of tests accuracy (e.g. differences between the accuracy of two tests) from bivariate meta-regression. Guidance on how to present the percentage study weights is also given. We do not demonstrate percentage study weights for the hierarchical summary receiver operating characteristic (HSROC) method (Harbord et al. 2007) since this method is the same as the bivariate normal model without covariates, and our focus is on the contribution towards the pooled sensitivity and specificity, rather than the HSROC curve.

The paper is structured as follows. Section 2 briefly outlines the models for which percentage weights are illustrated and Section 3 details a summary of the derivation of percentage study weights for the meta-analysis of diagnostic test accuracy studies. Section 4 provides two examples, covering bivariate meta-analysis and meta-regression, and highlights that the percentage study weights are an important addition over and above representations of study precision. Section 5 concludes with discussion.

2 Meta-analysis models for diagnostic test accuracy

We begin by briefly outlining two well-known methods of meta-analysis for DTA studies: the bivariate normal random effects model (Reitsma et al. 2005) and the hierarchical logistic

regression model (Chu and Cole 2006). The bivariate normal random effects model is not recommended by Cochrane for meta-analyses of DTA studies when there is sparse data or zero cells in the 2 by 2 contingency tables. However, we include it here to highlight that study weights depend on the choice of model. We then extend the hierarchical logistic regression model to a meta-regression model. We focus on random effects models since heterogeneity is expected in DTA meta-analyses, and the derivation of weights in a fixed effect model context is a simplified case.

2.1 The bivariate normal random effects model

The meta-analysis of diagnostic test accuracy in a two-stage process involves the estimation of logit sensitivity and logit specificity with their corresponding standard errors for each study, followed by the calculation of a weighted average of these statistics across the studies (Reitsma et al. 2005). For study i (i=1, …, k), let $\hat{\mu}_{A,i}$ and $\hat{\mu}_{B,i}$ be the observed logit sensitivity and logit specificity, respectively, which we assume are normally distributed with true logit sensitivity ($\mu_{A,i}$) and true logit specificity ($\mu_{B,i}$) in each study, and corresponding variances $s^2_{A,i}$ and $s^2_{B,i}$, which are assumed known. If there are zero cells in the 2 by 2 contingency tables, then $\hat{\mu}_{A,i}$ and $\hat{\mu}_{B,i}$ are obtained by applying a continuity correction, usually +0.5 to each cell (Sweeting et al. 2004).

$$\begin{pmatrix}\hat{\mu}_{A,i}\\ \hat{\mu}_{B,i}\end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix}\mu_{A,i}\\ \mu_{B,i}\end{pmatrix}, \mathbf{C}_i\right) \text{ with } \mathbf{C}_i = \begin{pmatrix}s^2_{A,i} & 0\\ 0 & s^2_{B,i}\end{pmatrix} \qquad (1)$$

$$\begin{pmatrix}\mu_{A,i}\\ \mu_{B,i}\end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix}\mu_{A}\\ \mu_{B}\end{pmatrix}, \boldsymbol{\Sigma}\right) \text{ with } \boldsymbol{\Sigma} = \begin{pmatrix}\tau^2_{A} & \tau_{AB}\\ \tau_{AB} & \tau^2_{B}\end{pmatrix},$$

In this random effects model, the true logit sensitivities and logit specificities are assumed to be bivariate normally distributed with common mean values, μ_A and μ_B, and between-study variance-covariance matrix, $\boldsymbol{\Sigma}$, where τ^2_A and τ^2_B are the between-study variances and $\tau_{AB} = \rho \cdot \tau_A \cdot \tau_B$ is the covariance between the logit sensitivity and specificity across studies (ρ is their between-study correlation). This model can be fitted using, for example, the multivariate method of moments procedure (Jackson et al. 2013) or restricted maximum likelihood (REML) to give the summary estimates, $\hat{\mu}_A$ and $\hat{\mu}_B$, and the estimated between-study variance matrix, $\widehat{\boldsymbol{\Sigma}}$.

2.2 **Bivariate meta-analysis as a generalised linear mixed model**

A generalised linear mixed model (Chu and Cole 2006) is more often used in meta-analysis of diagnostic test accuracy, instead of the bivariate normal model, since it directly models the binomial distribution of the data (Macaskill et al. 2010). This approach is important when the within-study normality assumption is not appropriate, for example when there are sparse data and when there are zero cells in the 2 by 2 contingency tables, for which an arbitrary continuity correction is otherwise required (Hamza et al. 2008).

Let n_{11i}, n_{00i}, n_{01i}, and n_{10i} be the number of true positives, true negatives, false positives, and false negatives, respectively, in each study, i (i=1, …, K). Also, let N_{1i} be the number of diseased patients ($N_{1i} = n_{11i} + n_{10i}$) and N_{0i} be the number of non-diseased patients ($N_{0i} = n_{00i} + n_{01i}$). The bivariate generalised linear mixed model can be specified as follows:

$$
\begin{aligned}
n_{11i}|u_i &\sim Binomial(N_{1i}, Se_i) \\
n_{00i}|v_i &\sim Binomial(N_{0i}, Sp_i) \\
\text{logit}(Se_i) &= \mu_A + u_i \\
\text{logit}(Sp_i) &= \mu_B + v_i \\
\begin{pmatrix} u_i \\ v_i \end{pmatrix} &\sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_A^2 & \tau_{AB} \\ \tau_{AB} & \tau_B^2 \end{pmatrix}\right)
\end{aligned}
\qquad (2)
$$

where *Se* and *Sp* are sensitivity and specificity, respectively. As previously, the between-study variances of the logit sensitivity and logit specificity are denoted by τ_A^2 and τ_B^2, respectively, and the between-study covariance is given by $\tau_{AB} = \rho\tau_A\tau_B$. This model is commonly fitted using ordinary maximum likelihood (ML), for example using an integral approximation approach, such as Gauss-Hermite quadrature (Pinheiro JC and EC 2006). Model (2) is equivalent to the hierarchical summary receiver operating characteristic method that is also commonly adopted in meta-analysis of DTA studies (Rutter and Gatsonis 2001, Harbord et al. 2007).

2.3 **Extension to bivariate meta-regression**

Bivariate meta-regression is an important extension to examine the impact of study-level covariates, such as type of test, on sensitivity and specificity across DTA studies. Models (1)

and (2) can easily be extended to include study-level covariates; for example, model (2) can be extended by including a set of covariates, $\boldsymbol{Z_i}$, relating to both *Se* and *Sp* (Chu and Cole 2006).

$$
\begin{aligned}
n_{11i}|u_i &\sim Binomial(N_{1i}, Se_i) \\
n_{00i}|v_i &\sim Binomial(N_{0i}, Sp_i) \\
\text{logit}(Se_i) &= \alpha_A + \boldsymbol{Z_i\beta_A} + u_i \\
\text{logit}(Sp_i) &= \alpha_B + \boldsymbol{Z_i\beta_B} + v_i \\
\begin{pmatrix} u_i \\ v_i \end{pmatrix} &\sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_A^2 & \tau_{AB} \\ \tau_{AB} & \tau_B^2 \end{pmatrix}\right)
\end{aligned}
\qquad (3)
$$

Where the vectors $\boldsymbol{\beta_A}$ and $\boldsymbol{\beta_B}$ contain the coefficients for the effect of each study-level covariate on the true logit sensitivity and logit specificity, respectively, and α_A and α_B denote the true logit sensitivity and logit specificity, respectively, when all covariates are zero.

3 Derivation of percentage study weights in a meta-analysis of sensitivity and specificity

We now review the current presentational approach for reporting results from a meta-analysis of DTA studies, which includes a representation of the precision of study-specific estimates, and then describe our proposal for additionally deriving and presenting percentage study weights.

3.1 Current approach to presenting study information in bivariate meta-analysis of DTA studies

After using one of the methods detailed above, or the HSROC method (Rutter and Gatsonis 2001), the graphical representation of the meta-analysis usually includes providing summary results (pooled sensitivity and specificity), alongside study estimates and a representation of their precision, either by a confidence interval (region) or by scaling the points based on the sample size (Van Houwelingen et al. 1993, Kontopantelis and Reeves 2013). For example, Harbord and Whiting developed a Stata module, 'metandi', which fits the hierarchical logistic regression model (2) for the meta-analysis of diagnostic test accuracy (Harbord and Whiting 2009). This module generates a plot that presents a graphical summary of the fitted model,

which includes, amongst other things, the summary point and its confidence region, and the individual study estimates. The plot scales the point of each study estimate by the total number in each study and presents this on the plot by an open circle. Similarly, in RevMan, the study weights are represented by circles in the ROC space, where the relative size of the circles are based on sample size or standard error of the study specific estimates of logit sensitivity and specificity (Review Manager (RevMan). 2014). Also, Phillips et al. propose a cross-hairs plot on ROC space, which shows the confidence interval width for each study point in both dimensions (Phillips et al. 2010).

Therefore, current methods focus on representing relative precision of study estimates based on sample size or confidence intervals standard errors. However, actual percentage study weights are not included, and indeed for bivariate meta-analysis or meta-regression models (2) and (3) no previous suggestions for deriving weights have been given.

3.2 An approach to deriving percentage study weights

We now suggest an approach to deriving actual percentage study weights in bivariate meta-analysis models of sensitivity and specificity, by adopting the approach of Riley et al. (Riley et al. 2017), which itself extends Jackson et al. (Jackson et al. 2015). Note we are not proposing a new weighting or estimation scheme for fitting DTA meta-analysis models: rather, we are proposing how to extract study weights that are inherent within existing DTA meta-analysis models (1) to (3) but otherwise hidden unless our method is used to extract them.

In a multi-parameter meta-analysis situation, such as models (1) to (3), we can decompose the variance matrix ($\mathbf{var}(\hat{\boldsymbol{\beta}})$) for a vector of main parameters, $\boldsymbol{\beta}$, (for example, $\boldsymbol{\beta}=(\mu_A, \mu_B)$ in models (1) and (2)) into the sum of independent weight matrices, $\boldsymbol{W}_i(\hat{\boldsymbol{\beta}})$ for each study, i. As described by Riley et al. (Riley et al. 2017), this is achieved by utilising a decomposition of Fisher's information matrix, where the total information matrix is defined as the inverse of $\mathbf{var}(\hat{\boldsymbol{\beta}})$, and where $\boldsymbol{I}_{total}(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{K} \boldsymbol{I}_i(\hat{\boldsymbol{\beta}})$ is the sum of the independent information matrices from each study, $i=1, \ldots, K$. This can be expressed as follows in equation (1):

$$\begin{aligned} \mathbf{var}(\hat{\boldsymbol{\beta}}) &= \mathbf{var}(\hat{\boldsymbol{\beta}}) \times \boldsymbol{I}_{total}(\hat{\boldsymbol{\beta}}) \times \mathbf{var}(\hat{\boldsymbol{\beta}}) \\ &= \mathbf{var}(\hat{\boldsymbol{\beta}}) \times \sum_{i=1}^{K} \boldsymbol{I}_i(\hat{\boldsymbol{\beta}}) \times \mathbf{var}(\hat{\boldsymbol{\beta}}) \end{aligned} \qquad (1)$$

$$= \sum_{i=1}^{K} \boldsymbol{W}_i(\hat{\boldsymbol{\beta}})$$

Equation (1) forms the basis of our derivation of percentage study weights in all meta-analysis models for DTA studies. It assumes independent studies, and provides study-specific weight matrices ($\boldsymbol{W}_i(\hat{\boldsymbol{\beta}})=\mathbf{var}(\hat{\boldsymbol{\beta}})\times\boldsymbol{I}_i(\hat{\boldsymbol{\beta}})\times\mathbf{var}(\hat{\boldsymbol{\beta}})$), which sum to give the total variance matrix for $\hat{\boldsymbol{\beta}}$.

The matrix $\mathbf{var}(\hat{\boldsymbol{\beta}})$ is immediately available post-estimation of the chosen meta-analysis model. However, the user must also obtain the study-specific information matrices, $\boldsymbol{I}_i(\hat{\boldsymbol{\beta}})$. To do this, Riley et al. (Riley et al. 2017) suggest using the generalised least squares solution for $\boldsymbol{I}_i(\hat{\boldsymbol{\beta}})$ obtained by

$$\boldsymbol{I}_i(\hat{\boldsymbol{\beta}})=(\mathbf{X}_i^{\mathrm{T}}\mathbf{V}_i^{-1}\mathbf{X}_i) \qquad (2)$$

where $\mathbf{X}_i^{\mathrm{T}}$ is the reduced design matrix for the fixed-effect parameters in the model, now just containing rows for participants in study *i*, and $\boldsymbol{V_i}$ is the reduced variance matrix (corresponding to just the binomial data for study *i*) with entries forced to be the same as those estimated for study *i* in the full analysis. As with general linear mixed model notation where $\mathbf{Y}=\mathbf{X\beta}+\mathbf{Zu}+\mathbf{e}$, $\mathbf{V}$ denotes the variance of $\mathbf{Y}$ (the response values) conditional on $\mathbf{X}$ (the design matrix) (Brown and Prescott 2015), and so $\mathbf{V}_\mathrm{i}$ includes both sampling variation in the ith study and between-study variance. The solution in (2) follows the common estimation method for general linear mixed models, such as (1), but is based on a pseudo-likelihood estimation approach for generalised linear mixed models, such as (2) and (3). We refer to the Supplementary Information for more technical detail.

Once $\boldsymbol{W}_i(\hat{\boldsymbol{\beta}})=\mathbf{var}(\hat{\boldsymbol{\beta}})\times\boldsymbol{I}_i(\hat{\boldsymbol{\beta}})\times\mathbf{var}(\hat{\boldsymbol{\beta}})$ are computed for each study, the percentage study weights can be obtained. For each parameter estimate within $\hat{\boldsymbol{\beta}}$, percentage study weights are derived by comparing the corresponding diagonal entries of $\boldsymbol{W}_i(\hat{\boldsymbol{\beta}})$ and $\mathbf{var}(\hat{\boldsymbol{\beta}})$. So, if the parameter corresponding to row *r* of $\hat{\boldsymbol{\beta}}$ is of interest, then the percentage weight of study *i* is given by

$$\% \text{ weight study } i = 100 * \frac{\boldsymbol{W}_i(\hat{\boldsymbol{\beta}})_{r,r}}{\sum_{i=1}^{K} \boldsymbol{W}_i(\hat{\boldsymbol{\beta}})_{r,r}} = 100 * \frac{\boldsymbol{W}_i(\hat{\boldsymbol{\beta}})_{r,r}}{\mathbf{var}(\hat{\boldsymbol{\beta}})_{r,r}} \quad (3)$$

where the '*r*,*r*' notation refers to the element (*r*,*r*) of the corresponding matrix. For example, in model (2) there are two parameters, μ_A and μ_B, which give the summary logit sensitivity and logit specificity, and so $\boldsymbol{W}_i(\hat{\boldsymbol{\beta}})$ is a 2 by 2 matrix for each study. Therefore, the $\boldsymbol{W}_i(\hat{\boldsymbol{\beta}})_{1,1}$ and $\mathbf{var}(\hat{\boldsymbol{\beta}})_{1,1}$, and the $\boldsymbol{W}_i(\hat{\boldsymbol{\beta}})_{2,2}$ and $\mathbf{var}(\hat{\boldsymbol{\beta}})_{2,2}$, are needed to derive the percentage study weights toward the summary logit sensitivity and logit specificity estimates, respectively. It is important to note that the percentage weight of study *i* may be different for each parameter, and so each should be reported.

We provide a step-by-step guide to deriving percentage study weights in Box 1.

To obtain the percentage study weights for any meta-analysis or meta-regression model for DTA studies, one needs to:

1. Fit the chosen meta-analysis / meta-regression model, and obtain the variance matrix (var($\hat{\boldsymbol{\beta}}$)) and its inverse (Fisher's total information matrix, $\boldsymbol{I}_{total}(\hat{\boldsymbol{\beta}})$) for the parameter estimates $(\hat{\boldsymbol{\beta}})$. For generalised linear mixed models (e.g. logistic regression with random effects such as models (2) or (3)) $\mathbf{var}(\hat{\boldsymbol{\beta}})$ and $\boldsymbol{I}_{total}(\hat{\boldsymbol{\beta}})$) should be calculated based on the pseudo-likelihood estimation solution (see Supplementary Information for more detail).
2. Derive $\boldsymbol{I}_i(\hat{\boldsymbol{\beta}})$ for each study based on equation (2).
3. Obtain a weight matrix $\boldsymbol{W}_i(\hat{\boldsymbol{\beta}})$ for each study using $\boldsymbol{W}_i(\hat{\boldsymbol{\beta}})$=var$(\hat{\boldsymbol{\beta}})$*$\boldsymbol{I}_i(\hat{\boldsymbol{\beta}})$*var$(\hat{\boldsymbol{\beta}})$; these weight matrices sum to give the variance matrix (i.e. var$(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{K} \boldsymbol{W}_i(\hat{\boldsymbol{\beta}})$), as shown in equation (1).
4. Use equation (3) to derive a study's percentage weight toward a particular parameter by comparing the corresponding diagonal elements of $\boldsymbol{W}_i(\hat{\boldsymbol{\beta}})$ and var$(\hat{\boldsymbol{\beta}})$. So, if the parameter corresponding to row *r* of $\hat{\boldsymbol{\beta}}$ is of interest, we would derive

$$\% \text{ weight of study } i = 100 * \frac{\boldsymbol{W}_i(\hat{\boldsymbol{\beta}})_{r,r}}{\mathbf{var}(\hat{\boldsymbol{\beta}})_{r,r}}$$

Box 1: A step-by-step guide on how to derive the percentage study weights.

3.3 Estimation and software

The percentage weights can easily be obtained for the parameters in the bivariate normal model (1) using the 'wt' option within the *mvmeta* package in Stata (White 2009). This also provides percentage study weights from a bivariate meta-regression extension to model (1).

The hierarchical logistic regression models (2) and (3) can be fitted in numerous different software packages, for example, 'meqrlogit' in Stata. Post-estimation, matrix algebra can be used to derive the percentage study weights and in the Appendix we provide a worked example of how to do this in Stata, which follows steps 1-4 in Box 1.

3.4 Presentation of percentage study weights

We recommend that the derived percentage study weights should be reported and presented as they provide important information for the reader of DTA meta-analysis results. Following model (1) or (2), the percentage study weights toward the summary sensitivity and specificity can be added most simply as columns within forest plots for each of sensitivity and specificity separately. This can be produced easily using packages such as Stata or RevMan. Although not customary in meta-analysis of DTA studies, the forest plots can also include scaled squares/circles for the study-specific point estimates, to reflect the weights. Stata code is provided in the Appendix to show how to use the Stata module 'metan' to force study weights to be presented on forest plots alongside the summary results and study-specific estimates and confidence intervals.

Additionally, the circles for each study estimate in the ROC space could be scaled according to the actual study weights, but this needs to be done in two directions using a rectangle or oval shape. The use of the oval shape to represent weight in two dimensions is already available in RevMan and in the metandi module in Stata, however, the ovals are currently scaled according to sample size or standard error of the study specific estimates of logit sensitivity and logit specificity. These software could be updated based on our proposal. For meta-regression model (3), percentage study weight toward a particular coefficient can also be presented in a table. Examples of all these suggestions are presented in Section 4.

4 Applied examples

We now illustrate the derivation of percentage study weights with two examples. We first introduce the datasets.

4.1 Datasets

Diagnosis of fever in children: The first example considers infrared ear thermometers for diagnosing fever in children, and consists of 23 studies and a total of 4100 children. Rectal temperature was used as the reference standard as it is a well-established method of measuring temperature in children, and most studies defined 38°C as the cut-off value for fever. More details of the original study can be found elsewhere (Craig et al. 2002, Dodd et al. 2006). A summary of the 2 by 2 tables of diagnostic accuracy is provided in Table 1. Where there are zero cells in a study, an arbitrary continuity correction of 0.5 is applied to obtain logit sensitivity and logit specificity estimates, and their variances, for model (1).

Diagnosis of Alzheimer's disease: The second example from Hamza et al. consists of nine studies that assess the test accuracy of positron emission tomography (PET) in the diagnosis of Alzheimer's disease (Hamza et al. 2008). There are small numbers of patients in each study (range: 19 to 50) with a total of 254 diseased and 210 non-diseased patients (Table 2). Once again, where there are zero cells, an arbitrary continuity correction of 0.5 is applied to enable model (1) to be applied.

4.2 Percentage study weights toward summary sensitivity and specificity

Models (1) and (2) were fitted to both datasets, and subsequently percentage weights derived for our proposal based on the step by step process outlined in Box 1. The results are shown in Table 3 and Table 4, alongside the relative scaling by sample size (shown in Figure 1) that is often currently used on the summary ROC plot. Figure 2 and Figure 3 show our suggestion for presenting percentage study weights within forest plots to aid interpretation of the results. We now discuss some important findings.

4.2.1 **Actual percentage study weights can be very different to the presentation of precision based on sample size**

The relative precision of each study estimate based on sample size is very different to the percentage study weights. Consider the Nypaver study in the fever data, for which the relative precision based on sample size of sensitivity and specificity were 24.9% and 18.9%, respectively. These values might infer that this study carried a large weight toward the pooled estimates. Intuitively, this seems sensible since the study is comparatively large and is represented by the largest circle on the ROC plot in Figure 1, where the circles are based on sample size.

However, the actual percentage study weights were much lower; for example, for the logistic regression model (2) the percentage weights for the Nypaver study are 5.3% and 5.4% for sensitivity and specificity, respectively. The reason is that the presentation based on sample size does not account for the magnitude of between-study heterogeneity or that the true sensitivity and specificity vary across studies. These modify the weight of each study because they both impact upon the variance of the study-specific binomial data (Riley et al. 2017).

4.2.2 **Percentage study weights for models (1) and (2) can differ**

In the fever data, given that the pooled estimates for sensitivity and specificity were similar for both models (Table 3), the percentage study weights were almost identical, with only slight differences due to the continuity correction in the bivariate normal model (1). This is not the case in the Alzheimer's dataset, as the percentage study weights differ depending on the choice of bivariate model. This is not unexpected since there are small patient numbers, and sensitivity and specificity estimates close, or equal, to one in several studies; in such situations the use of logistic regression model (2) is preferable over the normal model (1) (Hamza et al. 2008, Macaskill et al. 2010, Debray et al. 2013). For example, for study 3, the percentage weight for the pooled sensitivity was 14.4% for the logistic regression model (2), whereas the percentage weight for this parameter in the bivariate normal model (1) was 17.3% (Table 4).

4.2.3 **High between-study variability leads to similar percentage study weights for all studies**

Another important result highlighted in the fever dataset is that, largely due to relatively high between-study variability (τ_A = 0.91 and τ_B = 1.07, model (1)), the percentage study weights were very similar across all studies for both sensitivity and specificity (between 2.1% and 5.3% for sensitivity, and between 2.0% and 5.7% for specificity (model (1))), regardless of the sample size of each study (where total sample sizes range from 15 to 878). As a result, the pooled sensitivity and specificity estimates were similar to an unweighted average across all studies. This again highlights that the representation of the precision of study estimates of logit sensitivity and logit specificity based on sample size in the ROC space (or confidence intervals on a forest plot) are not sufficient, and we need to additionally report and present percentage study weights.

Indeed, if weights are not presented, there is a danger that many readers will use the relative precision across studies to infer the study weights wrongly. For example, in the Nypayer study, there were 282 true positives out of 425 diseased, and 445 true negatives out of 453 non-diseased participants. The corresponding percentage weights were 5.3% and 5.4% for sensitivity and specificity, respectively, based on model (2). However, the scaling of study points based on relative sample size were 24.9% and 18.9% for sensitivity and specificity, respectively. For a trial with far fewer patients, such as the Bernardo trial where there were zero true positives out of three diseased, and 33 true negatives out of 35 non-diseased participants, the percentage weights based on model (2) were not too dissimilar to those for the Nypayer study, with values of 2.7% and 3.8% for sensitivity and specificity, respectively, due to the large between-study variance estimates. However, the relative size of the study points based on sample size were much smaller with values of 0.2% and 1.5% for sensitivity and specificity, respectively.

4.3 **Bivariate meta-analysis and meta-regression**

As previously explained in Section 2.3, meta-regression can be used to examine the impact of study-level covariates on test accuracy; for example, to compare different types of tests to diagnose the same disease, to compare different manufacturers of the same test, or to compare accuracy of the test in different populations (e.g. country). It is also possible to

derive percentage study weights for any parameter in a meta-regression model. We now illustrate this in the fever dataset using model (3). In this example, interest was in whether the accuracy of the test varied by the manufacturer of the device, which was either a product from the company, FirstTemp, or from a different company (denoted by FirstTemp=1 if FirstTemp, FirstTemp=0 if other device) (Craig et al. 2002). We note that each study only assessed one device; therefore, this represents an example of an indirect comparison of tests (Bossuyt et al. 2013, Takwoingi et al. 2013).

The percentage weights for each parameter are shown in Table 5 for bivariate model (3). Now there are percentage study weights for each of multiple parameters: the pooled sensitivity (and pooled specificity) when the covariate is zero (i.e. for a device other than FirstTemp), and the difference in the pooled sensitivity (and the difference in the pooled specificity) when the covariate is one. Theoretically, the trials for which FirstTemp=1 do not provide a contribution toward the pooled sensitivity when FirstTemp=0. However, both covariate values of zero and one contribute toward the estimate of the difference in the pooled sensitivity when the covariate equals one. This is also true for the two parameters for the pooled specificity.

The key parameter of interest is the difference in the pooled sensitivity (and pooled specificity), as it reveals whether there is a difference in test accuracy when using FirstTemp compared to the other devices. The results suggest that the sensitivity of the FirstTemp device was no different to that of the other devices. However, the specificity of the FirstTemp device was statistically significantly higher than that of the other devices (odds ratio estimate: 3.34, 95% CI: 1.17 to 9.53). Some studies contributed more weight toward this estimated difference than the others, such as Akinyinka, Hoffman 1999c, Loveys 1999b, and Wilshaw, for which the percentage study weights were 7.7, 7.4, 7.7 and 7.9%, respectively.

A study's weight toward the *differences* in sensitivity (or differences in specificity) between tests may be different than its weights toward the *overall* sensitivity or specificity. For example, the percentage weight for the overall pooled sensitivity for the Akinyinka study was 5.1% from model (2) (Table 3), whereas its percentage weight for the difference in the sensitivity between tests was 9.1% from model (3) (Table 5).

5 Discussion

In this article we have shown how to apply the framework of Riley et al. to derive percentage study weights in commonly used bivariate meta-analysis and meta-regression models for combining diagnostic test accuracy studies, and illustrated the method with two example datasets. Currently, for example in Cochrane reviews, percentage study weights are not routinely derived in DTA meta-analysis despite the fact that they are a suggested requirement in the reporting of traditional meta-analyses of randomised trials (Moher et al. 2009). We showed how to derive percentage study weights specifically for the bivariate random effects models, but the same methodology can be used to produce weights for models with a fixed effect assumption, or for trivariate models (Ma et al. 2016).

Our work leads to some important findings for meta-analyses of DTA studies. Firstly, in addition to presenting study-specific estimates and their precision, we recommend that the percentage study weights are routinely calculated and additionally reported and presented. These add important information and can be very different to comparisons of the relative precision of study estimates, which may otherwise be wrongly used by readers to infer the relative contribution of each study. For presentation of weights, further research of the best approach is needed, for example through communication with reviewers, patients and lay readers. For now, we recommend that forest plots provide summary results and percentage weights, and that further work is needed to enhance the ROC plot display to show the weights, in addition to sample size. For example, perhaps point estimates could be denoted by rectangles or ovals that are scaled in two dimensions to denote relative study weight for sensitivity and specificity for each study.

Our percentage study weights will often be similar for both bivariate models (1) and (2), unless there are small patient numbers and/or zero cells in the 2 by 2 contingency table (i.e. sensitivity or specificity equal to zero or one), and when one must use a continuity correction in the bivariate normal model (1). In this case, the hierarchical logistic regression model is preferred (Hamza et al. 2008), and this corresponds to differences in the weighting of each study in this model compared to the bivariate normal model (1). When there is high between-study variation relative to that within studies, the percentage study weights are likely to be very similar for all the studies, which supports the need to derive percentage study weights in such situations. Whilst high heterogeneity in DTA meta-analyses is common, it is not the rule, and sometimes a fixed effect meta-analysis may be plausible (e.g. if all studies used the

same threshold value and were done in the same setting). The change in the percentage study weights when using fixed effect or random effects models is also helpful, to reveal how the contribution of each study is affected by the choice of model and the magnitude of heterogeneity. In meta-regression models, it would be of interest to examine how study weights were affected by changes in the between-study variance assumptions (e.g. have different between study variances and correlation for each covariate in the model).

In situations where DTA systematic reviews include informative forest or ROC plots without a meta-analysis, the concepts of study weights and precision of study estimates are irrelevant. In this case no scaling of study points on ROC plots is appropriate. For example, currently in Cochrane reviews, study points on a forest plot of sensitivity and specificity are unweighted and summary points are not shown. Without summary results, the addition of study weights is not necessary. However, if summary results have been derived, we suggest a more complete forest plot would rather include the summary results and percentage study weights.

The illustrative examples have been for the simplest case of meta-analysis of DTA studies where there was no missing sensitivity or specificity data, and a common threshold in all studies; however the framework extends naturally to more complex scenarios, for example involving multiple thresholds (Riley et al. 2014). Our work, and indeed the proposal of Riley et al., also extends to derive percentage study weights in meta-analysis models based on individual participant data (IPD) (Riley et al. 2008).

In summary, we have described and illustrated how to derive percentage study weights for meta-analysis models in DTA reviews. We hope that this encourages users conducting future meta-analyses to reveal the contribution of each study toward meta-analysis results to inform decision making, particularly in the presence of studies with high risk of bias. A sensitivity analysis is typically advocated that excludes studies which fail to meet some standard of quality criteria or level of evidence. Here, the percentage study weights would reveal whether those studies at high risk of bias are influential studies. We recommend that the PRISMA-DTA guidelines include an item that encourages percentage study weights to be reported for DTA meta-analyses. Similar advice should be incorporated in the Cochrane Handbook for DTA reviews. Updates to the current packages in programmes, such as Stata and RevMan, will be explored to make the derivation and representation of percentage study weights more accessible.

APPENDIX

Worked example using Stata to derive percentage study weights for model (2) using the fever dataset.

```
mkmat se sp, mat(X)  # create the fixed effect design matrix
mat XT = X'          # transpose the design matrix

mat Z = I(46)        # create the random effects design matrix - 23 studies, 2 random
parameters

gen invn=1/n         # Create A (diagonal matrix) using variable, n (number of diseased for
                       sensitivity and number of non-diseased for specificity) in dataset
mkmat invn, mat(colA)      # make a matrix with the variable 'invn'
mat A=diag(colA)           # create a diagonal matrix with the matrix, colA, above

meqrlogit y se sp, nocons || trialid: se sp, ///
nocons cov(un) binomial(n) refineopts(iterate(3)) intpoints(7)      # Fit model (2)
predict events, mu         # predict the number of events
gen p=events/n             # create variable containing the probability of true positive and
                           true negative
gen var=p*(1-p)            # create variance of Bernoulli distribution
mkmat var, mat(Bvec) # Create B (diagonal matrix) based on the predicted probability,
                           p̂·(1-p̂) after fitting the model
mat B=diag(Bvec)

                           # Creating the G matrix containing the variances of the random effects
estat recovariance         # Stores the random effects covariance matrix
mat G_one = r(Cov2) # G matrix for one trial
mat list G_one             # Obtain G matrix values
mata:
        G_one=(1.2485,-0.6779\-0.6779,1.1260)      # Insert values from G_one
        G_five = blockdiag(G_one, blockdiag(G_one, blockdiag(G_one,blockdiag(G_one,
        G_one))))
        G_twentyone = blockdiag(G_five, blockdiag(G_five, ///
        blockdiag(G_five,blockdiag(G_five, G_one))))
        G_twentythree = blockdiag(G_twentyone, blockdiag(G_one,G_one))
        st_matrix("G", G_twentythree)      # store G_twentythree to use later, call it G
end

mat V = (Z*G*Z')+(A*syminv(B))             # create variance matrix for the observations
mat invV = invsym(V)                       # invert the variance matrix, V
mat fish = XT*invV*X       # derive Fisher's Information matrix
mat varb = invsym(fish)    # invert Fisher's Information matrix to obtain var(β̂)
matlist varb               # display var(β̂)

forvalues i=1/23 {         # Loop over studies to obtain the study specific percentage
                           weights
        mat V`i'=V
```

```
        mat V`i'[(`i'*2)-1,(`i'*2)-1] = 1000000000    # Replace study i so that it has zero
                                                      information
        mat V`i'[(`i'*2)-1,`i'*2] = 0
        mat V`i'[`i'*2,(`i'*2)-1] = 0
        mat V`i'[`i'*2,`i'*2] = 1000000000
        mat invV`i' = invsym(V`i')              # recalculate matrices when study i removed
        mat fish`i' = XT*invV`i'*X
        mat fish`i'_`i' = fish - fish`i'
        mat weight`i' = varb*fish`i'_`i'*varb           Equation (1)

        mat pctwgt`i'sens = 100*(weight`i'[1,1]/varb[1,1])   # derive percentage weight for
study i for sensitivity
        mat pctwgt`i'spec = 100*(weight`i'[2,2]/varb[2,2])   # derive percentage weight for
study i for specificity
}

forvalues i=1/23 {              # Display the percentage weights and check they sum to 100%
        di pctwgt`i'sens[1,1]
}

forvalues i=1/23 {
        di pctwgt`i'spec[1,1]
}
```

Worked example using Stata to force percentage study weights and model results into the forest plot using 'metan' and the fever dataset.

```
metan sens lcisens ucisens, xtitle(Sensitivity, size(vsmall)) first(0.71 0.59 0.80 Model(2))
label(namevar=trial) wgt(Wtsensmod2) nowt effect(Sensitivity) lcols(trial tp fp fn tn
Wtsensmod2 Wtspecmod2) plotr(lc(none)) xlabel(0,0.2,0.4,0.6,0.8,1) force name(sens,
replace)

metan spec lcispec ucispec, xtitle(Specificity, size(vsmall)) first(0.96 0.93 0.98 Model(2))
label(namevar=trial) wgt(Wtspecmod2) nowt effect(Specificity) lcols(trial tp fp fn tn
Wtsensmod2 Wtspecmod2) plotr(lc(none)) xlabel(0,0.2,0.4,0.6,0.8,1) force name(spec,
replace)
```

References

Bossuyt, P. M., C. Davenport, J. Deeks, C. Hyde, M. M. Leeflang and R. J. Scholten (2013). Chapter 11: Interpreting results and drawing conclusions. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. J. Deeks, P. M. Bossuyt and C. A. Gatsonis, The Cochrane Collaboration. **0.9**.

Brown, H. and R. Prescott (2015). Applied Mixed Models in Medicine, 3rd Edition. Chicester, John Wiley & Sons.

Chu, H. and S. R. Cole 2006. "Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach." *J Clin Epidemiol* **59**(12): 1331-1332; author reply 1332-1333 DOI: 10.1016/j.jclinepi.2006.06.011.

Craig, J. V., G. A. Lancaster, S. Taylor, P. R. Williamson and R. L. Smyth 2002. "Infrared ear thermometry compared with rectal thermometry in children: a systematic review." *Lancet* **360**(9333): 603-609 DOI: 10.1016/S0140-6736(02)09783-0.

Debray, T. P., K. G. Moons, G. M. Abo-Zaid, H. Koffijberg and R. D. Riley 2013. "Individual participant data meta-analysis for a binary outcome: one-stage or two-stage?" *PLoS One* **8**(4): e60650 DOI: 10.1371/journal.pone.0060650.

Dodd, S. R., G. A. Lancaster, J. V. Craig, R. L. Smyth and P. R. Williamson 2006. "In a systematic review, infrared ear thermometry for fever diagnosis in children finds poor sensitivity." *J Clin Epidemiol* **59**(4): 354-357 DOI: 10.1016/j.jclinepi.2005.10.004.

Hamza, T. H., H. C. van Houwelingen and T. Stijnen 2008. "The binomial distribution of meta-analysis was preferred to model within-study variability." *J.Clin.Epidemiol.* **61**(1): 41-51 DOI: S0895-4356(07)00150-3 [pii];10.1016/j.jclinepi.2007.03.016.

Harbord, R., J. Deeks, M. Egger, P. Whiting and J. A. Sterne 2007. "A unification of models for meta-analysis of diagnostic accuracy studies." *Biostatistics* **8**(2): 239-251 DOI: 10.1093/biostatistics/kxn023.

Harbord, R. M. and P. Whiting 2009. "metandi: Meta-analysis of diagnostic accuracy using hierarchical logistic regression." *Stata Journal* **9**(2): 211-229

Irwig, L., P. Macaskill, P. Glasziou and M. Fahey 1995. "Meta-analytic methods for diagnostic test accuracy." *Journal of Clinical Epidemiology* **48**(1): 119-130 DOI: Doi 10.1016/0895-4356(94)00099-C.

Jackson, D., I. R. White, M. Price, J. Copas and R. D. Riley 2015. "Borrowing of strength and study weights in multivariate and network meta-analysis." *Stat Methods Med Res* DOI: 10.1177/0962280215611702.

Jackson, D., I. R. White and R. D. Riley 2013. "A matrix-based method of moments for fitting the multivariate random effects model for meta-analysis and meta-regression." *Biom J* **55**(2): 231-245 DOI: 10.1002/bimj.201200152.

Kontopantelis, E. and D. Reeves 2013. "A short guide and a forest plot command (ipdforest) for one-stage meta-analysis." *The Stata Journal* **13**(3): 574-587

Ma, X., L. Nie, S. R. Cole and H. Chu 2016. "Statistical methods for multivariate meta-analysis of diagnostic tests: An overview and tutorial." *Stat Methods Med Res* **25**(4): 1596-1619 DOI: 10.1177/0962280213492588.

Macaskill, P., C. A. Gatsonis, J. Deeks, R. Harbord and Y. Takwoingi (2010). Chapter 10: Analysing and Presenting Results. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0. J. Deeks, P. M. Bossuyt and C. A. Gatsonis, The Cochrane Collaboration.

Moher, D., A. Liberati, J. Tetzlaff, D. G. Altman and P. Group 2009. "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement." *PLoS Med* **6**(7): e1000097 DOI: 10.1371/journal.pmed.1000097.

Phillips, B., L. A. Stewart and A. J. Sutton 2010. "'Cross hairs' plots for diagnostic meta-analysis." *Res Synth Methods* **1**(3-4): 308-315 DOI: 10.1002/jrsm.26.

Pinheiro JC and C. EC 2006. "Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models." *Journal of Computational and Graphical Statistics* **15**(1): 58-81

Reitsma, J. B., A. S. Glas, A. W. Rutjes, R. J. Scholten, P. M. Bossuyt and A. H. Zwinderman 2005. "Bivariate analysis of sensitivity and specificity produces informative

summary measures in diagnostic reviews." *J Clin Epidemiol* **58**(10): 982-990 DOI: 10.1016/j.jclinepi.2005.02.022.

Review Manager (RevMan). (2014). Copenhagen, The Nordic Cochrane Centre, The Cochrane Collaboration.

Riley, R., J. Ensor, D. Jackson and D. L. Burke 2017. "Deriving percentage study weights in multi-parameter meta-analysis models: with application to meta-regression, network meta-analysis, and one-stage individual participant data models." *Stat Methods Med Res*: 962280216688033 DOI: 10.1177/0962280216688033.

Riley, R., Y. Takwoingi, T. A. Trikalinos, A. Guha, A. Biswas, J. Ensor, K. Morris and J. Deeks 2014. "Meta-analysis of test accuracy studies with multiple and missing thresholds: a multivariate-normal model." *Biometrics & Biostatistics* **5**(3): 1000196

Riley, R. D., S. R. Dodd, J. V. Craig, J. R. Thompson and P. R. Williamson 2008. "Meta-analysis of diagnostic test studies using individual patient data and aggregate data." *Stat Med* **27**(29): 6111-6136 DOI: 10.1002/sim.3441.

Rutter, C. M. and C. A. Gatsonis 2001. "A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations." *Stat Med* **20**(19): 2865-2884

StataCorp (2015). Stata Statistical Software: Release 14. College Station, TX, StataCorp LP.

Sweeting, M. J., A. J. Sutton and P. C. Lambert 2004. "What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data." *Statistics in Medicine* **23**: 1351-1375

Takwoingi, Y., M. M. Leeflang and J. J. Deeks 2013. "Empirical evidence of the importance of comparative studies of diagnostic test accuracy." *Ann Intern Med* **158**(7): 544-554 DOI: 10.7326/0003-4819-158-7-201304020-00006.

The Cochrane Collaboration (2014). Review Manager (RevMan). Copenhagen, The Nordic Cochrane Centre.

Van Houwelingen, H. C., K. H. Zwinderman and T. Stijnen 1993. "A bivariate approach to meta-analysis." *Stat Med* **12**(24): 2273-2284

White, I. R. 2009. "Multivariate random-effects meta-analysis." *The Stata Journal* **9**(1): 40-56

Whiting, P. F., A. W. Rutjes, M. E. Westwood, S. Mallett, J. J. Deeks, J. B. Reitsma, M. M. Leeflang, J. A. Sterne, P. M. Bossuyt and Q.-. Group 2011. "QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies." *Ann Intern Med* **155**(8): 529-536 DOI: 10.7326/0003-4819-155-8-201110180-00009.

Table 1: Summary of the 23 temperature studies in the fever dataset (Craig et al. 2002, Dodd et al. 2006).

First Author	Number of true positives	Number of diseased	Sensitivity	Logit(sens)	s.e.logit(sens)*	Number of true negatives	Number of non-diseased	Specificity	Logit(spec)	s.e.logit(spec)$
Akinyinka	77	105	0.73	1.01	0.22	259	273	0.95	2.92	0.27
Bernardo	0	3	0.13	-1.95	1.51	33	35	0.93	2.60	0.66
Brennan	150	203	0.74	1.04	0.16	155	167	0.93	2.56	0.30
Davis	9	18	0.50	0.00	0.47	46	48	0.96	3.14	0.72
Green	8	9	0.85	1.73	0.89	12	12	0.96	3.22	1.44
Greenes	53	109	0.49	-0.06	0.19	193	195	0.99	4.57	0.71
Hoffman 1999a	30	42	0.71	0.92	0.34	56	58	0.97	3.33	0.72
Hoffman 1999b	36	62	0.58	0.33	0.26	32	34	0.94	2.77	0.73
Hoffman 1999c	41	42	0.98	3.71	1.01	44	55	0.80	1.39	0.34
Hooker 1993	10	15	0.66	0.65	0.53	24	24	0.98	3.89	1.43
Hooker 1996	75	99	0.76	1.14	0.23	78	81	0.96	3.26	0.59
Lanham	53	103	0.51	0.06	0.20	74	75	0.99	4.30	1.01
Loveys 1999a	12	30	0.40	-0.41	0.37	44	46	0.96	3.09	0.72
Loveys 1999b	37	47	0.79	1.31	0.36	74	93	0.80	1.36	0.26
Muma	48	87	0.55	0.21	0.21	136	136	1.00	5.61	1.42
Nypaver	282	425	0.66	0.68	0.10	445	453	0.98	4.02	0.36
Petersen-Smith	9	10	0.90	2.20	1.05	214	222	0.96	3.29	0.36
Rhoads	7	27	0.27	-1.01	0.43	38	38	0.99	4.34	1.42
Robinson	1	2	0.50	0.00	1.15	13	13	0.96	3.30	1.44
Selfridge	16	18	0.89	2.08	0.75	75	84	0.89	2.12	0.35
Stewart	57	59	0.96	3.14	0.65	20	20	0.98	3.71	1.43
Terndrup	91	178	0.51	0.04	0.15	105	125	0.84	1.66	0.24
Wilshaw	16	16	0.97	3.50	1.44	60	104	0.58	0.31	0.20

Sens, sensitivity; spec, specificity; s.e., standard error; * s.e.logit(sensitivity) estimated based on $s^2_{A,i} = 1/N_{1i} \times \mu_{A,i} \times \left(1 - \mu_{A,i}\right)$ (equations (1) and (2)) (Reitsma et al. 2005); $ s.e.logit(specificity) estimated based on $s^2_{B,i} = 1/N_{0i} \times \mu_{B,i} \times \left(1 - \mu_{B,i}\right)$ (equations (1) and (2)) (Reitsma et al. 2005).

Table 2: Summary of the nine studies in the Alzheimer's dataset (Hamza et al. 2008).

Study ID	Number of true positives	Number of diseased	Sensitivity	Logit(sens)	s.e.logit(sens)*	Number of true negatives	Number of non-diseased	Specificity	Logit(spec)	s.e.logit(spec)$
1	33	39	0.85	1.70	0.44	35	40	0.88	1.95	0.48
2	18	24	0.75	1.10	0.47	10	15	0.67	0.69	0.55
3	20	33	0.60	0.42	0.35	41	41	0.99	4.42	1.42
4	19	19	0.98	3.66	1.43	19	19	0.98	3.66	1.43
5	44	50	0.88	1.99	0.44	19	29	0.66	0.64	0.39
6	18	21	0.86	1.79	0.62	9	10	0.90	2.20	1.05
7	27	28	0.96	3.30	1.02	21	25	0.84	1.66	0.55
8	21	21	0.98	3.76	1.43	9	10	0.86	1.85	0.88
9	18	19	0.95	2.89	1.03	20	21	0.95	3.00	1.02

Sens, sensitivity; spec, specificity; s.e., standard error; * s.e.logit(sensitivity) estimated based on $s^2_{A,i} = 1/N_{1i} \times \mu_{A,i} \times \left(1 - \mu_{A,i}\right)$ (equations (1) and (2)) (Reitsma et al. 2005); $ s.e.logit(specificity) estimated based on $s^2_{B,i} = 1/N_{0i} \times \mu_{B,i} \times \left(1 - \mu_{B,i}\right)$ (equations (1) and (2)) (Reitsma et al. 2005).

Table 3: Percentage study weights for the pooled sensitivity and specificity estimates in the fever dataset.

First author	Percentage study weight based on equation (3)				Relative precision based on sample size, irrespective of model choice	
	Bivariate random effects normal model (1)		Hierarchical logistic regression model (2)			
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
Akinyinka	5.5	5.9	5.1	5.6	6.1	11.4
Bernardo	2.0	4.4	2.7	3.8	0.2	1.5
Brennan	5.6	5.8	5.2	5.5	11.9	7.0
Davis	4.5	4.2	4.4	4.1	1.1	2.0
Green	3.0	2.3	3.3	2.6	0.5	0.5
Greenes	5.5	4.2	5.1	4.6	6.4	8.2
Hoffman 1999a	5.0	4.2	4.8	4.3	2.5	2.4
Hoffman 1999b	5.3	4.2	5	4	3.6	1.4
Hoffman 1999c	2.8	5.6	3.9	5.4	2.5	2.3
Hooker 1993	4.2	2.5	4.2	2.8	0.9	1.0
Hooker 1996	5.4	4.7	5.1	4.7	5.8	3.4
Lanham	5.5	3.3	5.1	3.9	6.0	3.1
Loveys 1999a	4.9	4.2	4.7	3.9	1.8	1.9
Loveys 1999b	4.9	5.9	4.7	5.6	2.8	3.9
Muma	5.4	2.6	5.1	3.8	5.1	5.7
Nypaver	5.8	5.6	5.3	5.4	24.9	18.9
Petersen-Smith	2.7	5.5	3.5	5.4	0.6	9.3
Rhoads	4.6	2.5	4.6	2.5	1.6	1.6
Robinson	2.3	2.3	2.1	2	0.1	0.5
Selfridge	3.5	5.6	3.8	5.4	1.1	3.5
Stewart	3.8	2.4	4.3	3.5	3.5	0.8
Terndrup	5.7	6.0	5.2	5.7	10.4	5.2
Wilshaw	2.2	6.0	2.8	5.7	0.9	4.3
TOTAL	100.0	100.0	100	100	100.0	100.0
Summary estimate (95% CI)	0.79 (0.35 to 1.22)	2.83 (2.30 to 3.36)	0.88 (0.37 to 1.38)	3.14 (2.54 to 3.74)	-	-
Between-study standard deviation estimate (95% CI)	0.91 (0.48 to 1.33)	1.07 (0.67 to 1.47)	1.11 (0.73 to 1.69)	1.21 (0.83 to 1.79)		

* Summary estimates are logit(sensitivity) and logit(specificity); $ between-study standard deviation estimates for logit(sensitivity) and logit(specificity); CI, confidence interval.

Table 4: Percentage study weights for the pooled sensitivity and specificity estimates in the Alzheimer's dataset (Hamza et al. 2008).

Study ID	Percentage study weight based on equation (3)				Relative precision based on sample size, irrespective of model choice	
	Bivariate random effects normal model (1)		Hierarchical logistic regression model (2)			
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
1	15.6	16.5	13.4	13.9	15.4	19.0
2	15.1	15.1	13	13	9.4	7.1
3	17.3	4.9	14.4	9.5	13.0	19.5
4	4.5	4.8	7.5	8.1	7.5	9.0
5	15.8	18.4	13.6	14.7	19.7	13.8
6	12.4	7.6	11.4	8.7	8.3	4.8
7	7.4	15.1	9.8	13.1	11.0	11.9
8	4.5	9.6	7.7	8.8	8.3	4.8
9	7.3	7.9	9.2	10.2	7.5	10.0
TOTAL	100.0	100.0	100	100	100.0	100.0
Summary estimate (95% CI)*	1.82 (1.03 to 2.62)	1.77 (0.99 to 2.56)	2.20 (1.33 to 3.07)	2.27 (1.31 to 3.23)	-	-
Between-study standard deviation estimate (95% CI)$	0.75 (0.01 to 1.49)	0.73 (0 to 1.53)	0.99 (0.43 to 2.28)	1.11 (0.49 to 2.48)	-	-

* Summary estimates are logit(sensitivity) and logit(specificity); $ between-study standard deviation estimates for logit(sensitivity) and logit(specificity); CI, confidence interval.

Table 5: Percentage study weights for the meta-regression with binary variable (FirstTemp) in the fever dataset (Craig et al. 2002).

First author	Percentage study weight based on equation (3) for hierarchical logistic regression model (3)			
	Sensitivity for FirstTemp=0	Change in Sensitivity when FirstTemp=1	Specificity for FirstTemp=0	Change in Specificity when FirstTemp=1
Akinyinka	14.2	9.1	13.5	7.7
Bernardo	7.2	4.6	9.4	5.4
Brennan	0.0	2.9	0.0	4.3
Davis	0.0	2.5	0.0	2.8
Green	0.0	1.9	0.0	1.5
Greenes	0.0	2.9	0.0	3.4
Hoffman 1999a	0.0	2.7	0.0	3.1
Hoffman 1999b	13.9	8.9	9.8	5.6
Hoffman 1999c	10.8	6.9	12.9	7.4
Hooker 1993	0.0	2.4	0.0	1.7
Hooker 1996	14.1	9.0	11.5	6.6
Lanham	0.0	2.9	0.0	2.7
Loveys 1999a	13.2	8.4	9.9	5.6
Loveys 1999b	13.2	8.5	13.5	7.7
Muma	0.0	2.9	0.0	2.6
Nypaver	0.0	2.9	0.0	4.2
Petersen-Smith	0.0	2.0	0.0	4.1
Rhoads	0.0	2.6	0.0	1.5
Robinson	5.8	3.7	5.7	3.3
Selfridge	0.0	2.2	0.0	4.1
Stewart	0.0	2.4	0.0	2.2
Terndrup	0.0	2.9	0.0	4.4
Wilshaw	7.7	4.9	13.8	7.9
TOTAL	100.0	100.0	100.0	100.0
Summary estimate* (95% CI)	Sens (FirstTemp=0):0.74 (0.55 to 0.87)	OR for FirstTemp=1 vs FirstTemp=0: 0.74 (0.26 to 2.10)	Spec (FirstTemp=0):0.91 (0.82 to 0.96)	OR for FirstTemp=1 vs FirstTemp=0: 3.34 (1.17 to 9.53)

* When FirstTemp=0, summary estimate represents sensitivity and specificity. When FirstTemp=1, summary estimate represents an odds ratio comparing the accuracy of the FirstTemp device to other devices; CI, confidence interval.

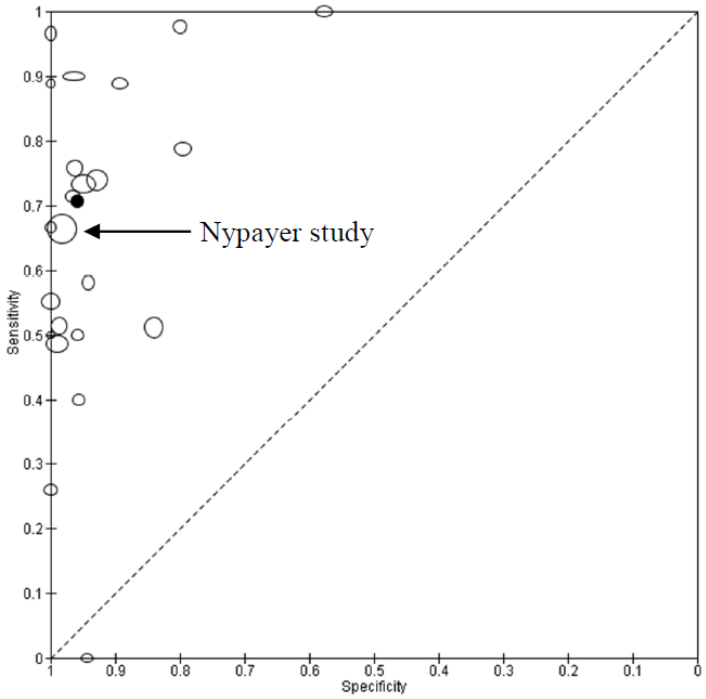

Figure 1: ROC plot for the fever data including study specific estimates and summary point with size of circles relative to sample size.

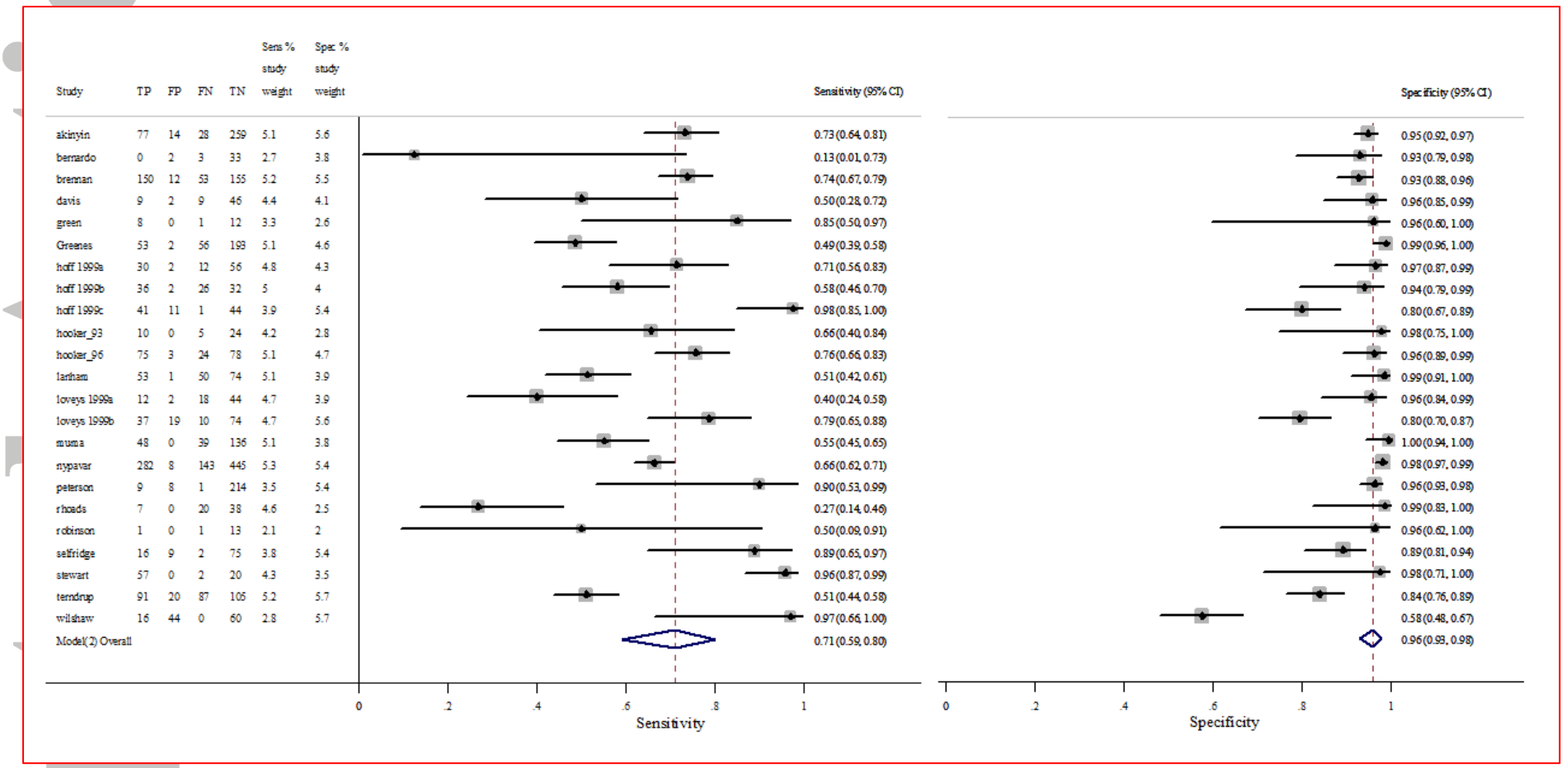

Figure 2: Forest plot for study-specific sensitivity and specificity estimates with percentage study weights from the logistic regression model (2) in the fever dataset.

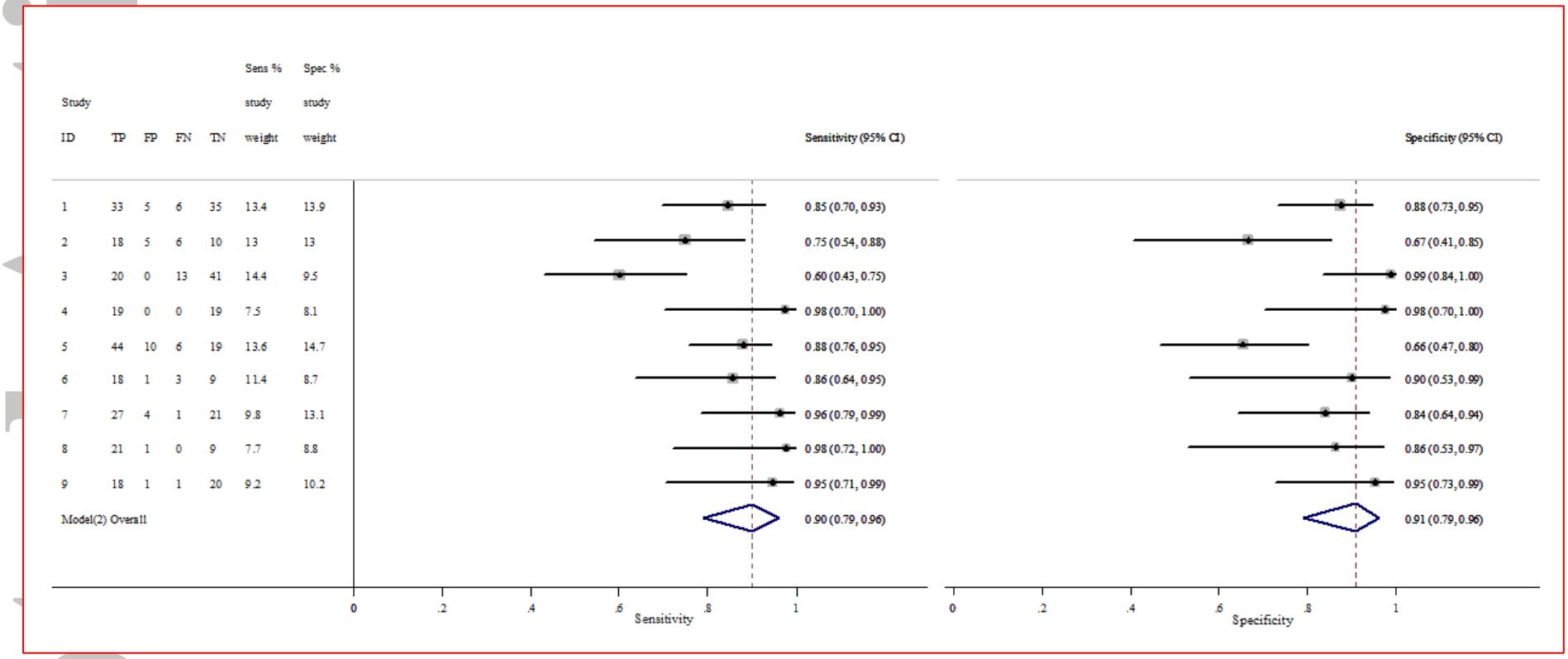

Figure 3: Forest plot for study-specific sensitivity and specificity estimates with percentage study weights from the hierarchical logistic regression model (2) in the Alzheimer's dataset.