Research Article

Statistics in Medicine

Received: 27 January 2014, Accepted: 31 January 2016 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6919

A multivariate method for meta-analysis and comparison of diagnostic tests

Niki L. Dimou, Maria Adam and Pantelis G. Bagos*†

We present here an extension of the classic bivariate random effects meta-analysis for the log-transformed sensitivity and specificity that can be applied for two or more diagnostic tests. The advantage of this method is that a closed-form expression is derived for the calculation of the within-studies covariances. The method allows the direct calculation of sensitivity and specificity, as well as, the diagnostic odds ratio, the area under curve and the parameters of the summary receiver operator's characteristic curve, along with the means for a formal comparison of these quantities for different tests. There is no need for individual patient data or the simultaneous evaluation of both diagnostic tests in all studies. The method is simple and fast; it can be extended for several diagnostic tests and can be fitted in nearly all statistical packages. The method was evaluated in simulations and applied in a meta-analysis for the comparison of anti-cyclic citrullinated peptide antibody and rheumatoid factor for discriminating patients with rheumatoid arthritis, with encouraging results. Simulations suggest that the method is robust and more powerful compared with the standard bivariate approach that ignores the correlation between tests. Copyright © 2016 John Wiley & Sons, Ltd.

Keywords: meta-analysis; diagnostic tests; SROC method

1. Introduction

Laboratory diagnostic tests are routinely used in medical research to screen for, diagnose, grade and monitor the progression of disease. The most common way to describe the performance of a diagnostic test is the 2×2 table, which gives a number of positive and negative test results among the subjects with and without the disease. Diagnostic accuracy is commonly measured using true positive rate (*TPR*) and false positive rate (*FPR*). An equivalent parameterization is in terms of Sensitivity (*Se*) and Specificity (*Sp*). Thus, *TPR* is equivalent to Sensitivity and *FPR* to 1-Specificity of a diagnostic test. A well-established method of summarizing the performance of a diagnostic test is the receiver operating characteristic (ROC) curve, which indicates the relationship between the *TPR* and *FPR* at different diagnostic thresholds [1].

Meta-analysis constitutes a particular type of research, in which a set of original studies is synthesized and the potential diversity across them is explored using specific statistical methods [2–5]. Meta-analytic techniques have been proposed for combining the results of different studies that evaluated the accuracy of a given diagnostic test [6–10]. The major difference of diagnostic studies is that a pair of estimates (*TPR* and *FPR*) is usually reported rather than a single statistic and thus specialized (i.e. bivariate) methodology needs to be utilized.

The simplest (and definitely not recommended) method for meta-analysis of diagnostic tests suggests that the numbers of true positives (*TP*), false positives (*FP*), true negatives (*TN*) and false negatives (*FN*) from each study are summed resulting in a non-stratified meta-analysis design. A separate meta-analysis of sensitivity and specificity has been proposed based on logit transformations. This approach allows for heterogeneity in sensitivity and specificity but ignores their between-studies correlation (i.e. the correlation of the random effects). Moreover, separate meta-analysis of positive and negative likelihood ratios has been proposed [11], which also ignores the correlation between these two parameters. Other authors have proposed that an estimate of the diagnostic odds ratio (*DOR*) could be derived from each study and

Department of Computer Science and Biomedical Informatics, University of Thessaly, Papasiopoulou 2-4, Lamia 35100, Greece

**Correspondence to: Dr Pantelis G. Bagos, Department of Computer Science and Biomedical Informatics, University of Thessaly, Papasiopoulou 2–4, Lamia 35100, Greece.*

E-mail: pbagos@compgen.org

combined in a standard meta-analysis [12], which can be performed using either fixed or random effects models. In the latter case, the between studies heterogeneity is taken into account but we ignore the fact that in each study a different threshold may have been used; additionally, calculation of a pooled estimate of sensitivity and specificity is not feasible in this case. The most commonly used approach is the summary ROC (SROC) method [7,13], which uses logit-trasforms of *TPR* and *FPR,* and it is based on simple linear regression of their difference (*D,* which is equal to log*DOR*) on their sum (*S*). A summary ROC curve can then be derived from the fitted regression line. This method incorporates the use of a different threshold in each study, but even if a weighted analysis for the sum of the parameters is performed, the measurement error in *S* is not taken into account.

The hierarchical SROC method [8] expresses the logit-transformed *TPR* and *FPR* in terms of two parameters (accuracy and threshold). This method allows for between studies heterogeneity by modelling the accuracy parameter as a random effects term. The model was simplified by using Empirical Bayes estimates, and the results were close to those obtained using the Bayesian analysis [9] and was extended to account for the situation where no gold standard test is available [14]. Recently, a standard bivariate random effects meta-analysis of logit-transformed *TPR* and *FPR* has been proposed. This method allows for between studies heterogeneity and for the correlation of *TPR* and *FPR* [10,15] involving a hierarchical structure where the within-study variation refers to the variation in the repeated sampling of the studies' results if they were replicated and the between-study variation refers to any variation in the studies' true underlying estimates of *TPR* and *FPR*. Methods that directly model the binomial structure data are, in general, recommended [10,16]. Harbord and co-workers, showed that the various previously mentioned multivariate methods for meta-analysis of diagnostic tests are essentially equivalent models that use a different parameterization [6]. Recently, a composite likelihood method for bivariate meta-analysis has been proposed, which overcomes the nonconvergence problem of the standard likelihood estimation methods when the number of studies is small and is more robust than the standard likelihood inference to misspecifications of the joint distributions assumptions [17]. When the disease prevalence is known, as in cohort studies, a trivariate modelling of the disease prevalence, sensitivity and specificity has been proposed [18]. This approach accounts for the dependence of sensitivity and specificity on disease prevalence, which is more obvious when a continuous trait is used as a classifier. This dependence was assessed using a Pearson-type correlation coefficient [19]. A shortcoming of this trivariate modelling approach is that it can only include cohort studies with information estimating study-specific disease prevalence. Thus, two main alternative models have been introduced, including a novel Bayesian hierarchical model for combing cohort and case-control studies and correcting partial verification bias [20] and a hybrid model, where an alterative inference procedure based on composite likelihood is used [21]. Ma and coworkers have presented a recent overview of the existing multivariate methods for meta-analysis of diagnostic studies [22].

When one wants to compare two or more diagnostic tests, the situation is complicated. Several methods, which make use of the *DOR*, have been proposed. The traditional method of meta-analysis for the comparison of two diagnostic tests extracts *D* values (i.e. the difference of logit-trasformed *TPR* and *FPR*) from each study (or the estimated values using the SROC model) and then summarizes them. The diagnostic tests can then be compared using the difference of summary *Ds*. Another approach is based on the relative OR as the relative accuracy of one test against the other and makes the assumption that the two tests were performed on 'paired' subjects within each study. If sufficient data are available (which however, rarely is the case), the calculation of the conditional relative *OR* (*CROR*) can be used based on the discordant results of diagnostic tests [23]. Furthermore, statistical methodology for repeated measurements has also been adopted in order to combine several studies of diagnostic tests, where each study reports on more than one test [24]. In situations when a gold standard is not available, a multivariate random effects model has been used to model simultaneously the sensitivity, the specificity and the prevalence of the disease [25]. Mixed effects models or Bayesian hierarchical models can be used for the estimation of the parameters. When the conditional independence assumption does not hold [26], a correlation among the diagnostic tests can be imposed. The aforementioned approach and the hierarchical SROC framework by Dendukuri and coworkers [14] are closely related and some of their submodels are equivalent [27]. Finally, Trikalinos and coworkers introduced a Bayesian model that incorporates the relations between *TPR* and *FPR* across two or more tests when those are applied to the same patients [28].

In this work, we present a simple yet powerful approach for performing multivariate meta-analysis of diagnostic studies. The model we propose is a direct extension of the multivariate meta-analysis method of diagnostic tests [10,15]. The key element of our approach is the calculation of the within studies covariance of the parameters necessary for the multivariate meta-analysis method [29]. Because the

model is based on the general model for multivariate meta-analysis, it has some very important features: it can easily incorporate studies reporting only one of the tests under a missing at random assumption and, thus, allows for borrowing strength from external studies [30]; it can be easily fitted using standard software used for multivariate meta-analysis [31,32]. It allows after appropriate transformations the recovery of other important metrics such as the DOR, the SROC curve and the area under the curve (AUC), and finally, it allows the direct comparison of theses parameters of the different tests using formal techniques. In Section 2, we introduce the multivariate meta-analysis model for two or more diagnostic tests; we derive formulae for within-study covariance and recover the SROC method and the AUC. In Section 3, we perform a simulation study to investigate the properties of our method. In Section 4, we apply our method to data from a meta-analysis of diagnostic tests for rheumatoid arthritis. We close with discussion in Section 5.

2. Methods

2.1. The multivariate model

Let Y_i, X_{1i} and X_{2i} denote three categorical random variables with two levels (i.e. 0, 1) that are used to classify n_i individuals for study i ($i=1,2,\ldots,k$). Usually, Y_i denotes the disease status and X_{1i}, X_{2i} the test result (positive or negative) for study i ($i=1,2,\ldots,k$). In the general case, the data would be presented in the form of a three-dimensional ($2\times2\times2$) contingency table (Table I) where we denote the counts as n_{clpi} with $c,l,p\in\{0,1\}$ and this table is referred to as partial contingency table. Usually, the results of two diagnostic tests are represented in the form of Table II, which, in the terminology of contingency tables, is a marginal table because for each test, the outcome of the other test is ignored. The logit transformations of $\widehat{TPR}_{ji}(\widehat{\text{Se}}_{\text{ji}})$ and $\widehat{FPR}_{ji}$ ($1\text{-}\widehat{Sp}_{ji}$) for test j ($j=1,2$) and for study i ($i=1,2,\ldots,k$) are given by the following:

$$\widehat{y}_{1i}=\text{logit}\left(\widehat{TPR}_{1i}\right)=\text{logit}\left(\widehat{Se}_{1i}\right)=\log\left(\frac{TP_{1i}}{FN_{1i}}\right)=\log\left(\frac{n_{11+i}}{n_{10+\text{i}}}\right) \tag{1}$$

$$\widehat{y}_{2i}=\text{logit}\left(\widehat{FPR}_{1i}\right)=\text{logit}\left(1-\widehat{Sp}_{1i}\right)=\log\left(\frac{FP_{1i}}{TN_{1i}}\right)=\log\left(\frac{n_{01+i}}{n_{00+i}}\right) \tag{2}$$

Table I. The table that defines the joint distribution for the association of the two diagnostic tests according to the disease status for study i ($i=1,2,\ldots,k$).

		Y_i			
		$Y_i=1$		$Y_i=0$	
		X_{2i}			
		$X_{2i}=1$	$X_{2i}=0$	$X_{2i}=1$	$X_{2i}=0$
X_{1i}	$X_{1i}=1$	n_{111i}	n_{110i}	n_{011i}	n_{010i}
	$X_{1i}=0$	n_{101i}	n_{100i}	n_{001i}	n_{000i}

We denote the counts as n_{clpi}, with c being the indicator for Y_i, l the indicator for X_{1i} and p the indicator for X_{2i}. This table is usually referred to as «partial contingency table». See also [29,39].

Table II. Classification of the findings of two diagnostic tests according to the test result (positive or negative) and the disease status for study i ($i=1,2,\ldots,k$).

		X_{1i}		X_{2i}	
		$X_{1i}=1$	$X_{1i}=0$	$X_{2i}=1$	$X_{2i}=0$
Y_i	$Y_i=1$	n_{11+i} (TP_{1i})	n_{10+i} (FN_{1i})	n_{1+1i} (TP_{2i})	n_{1+0i} (FN_{2i})
	$Y_i=0$	n_{01+i} (FP_{1i})	n_{00+i} (TN_{1i})	n_{0+1i} (FP_{2i})	n_{0+0i} (TN_{2i})

We denote the counts as n_{clpi}, with c being the indicator for Y_i, l the indicator for X_{1i} and p the indicator for X_{2i}. Thus, the interior cells of the table consist of the marginals of Table I (i.e. $n_{cl+i}=n_{cl0i}+n_{cl1i}$ and $n_{c+pi}=n_{c0pi}+n_{c1pi}$). In terms of contingency tables, these two tables are named «marginal tables». See also [29,39].
FN, false negative; FP, false positive; TN, true negative; TP, true positive.

$$\hat{y}_{3i} = \text{logit}\left(\widehat{TPR}_{2i}\right) = \text{logit}\left(\widehat{Se}_{2i}\right) = \log\left(\frac{TP_{2i}}{FN_{2i}}\right) = \log\left(\frac{n_{1+1i}}{n_{1+0i}}\right) \tag{3}$$

$$\hat{y}_{4i} = \text{logit}\left(\widehat{FPR}_{2i}\right) = \text{logit}\left(1 - \widehat{Sp}_{2i}\right) = \log\left(\frac{FP_{2i}}{TN_{2i}}\right) = \log\left(\frac{n_{0+1i}}{n_{0+0i}}\right) \tag{4}$$

with approximate variances estimated by the following:

$$s^2_{1i} = \frac{1}{TP_{1i}} + \frac{1}{FN_{1i}} = \frac{1}{n_{11+i}} + \frac{1}{n_{10+i}} \tag{5}$$

$$s^2_{2i} = \frac{1}{FP_{1i}} + \frac{1}{TN_{1i}} = \frac{1}{n_{01+i}} + \frac{1}{n_{00+i}} \tag{6}$$

$$s^2_{3i} = \frac{1}{TP_{2i}} + \frac{1}{FN_{2i}} = \frac{1}{n_{1+1i}} + \frac{1}{n_{1+0i}} \tag{7}$$

$$s^2_{4i} = \frac{1}{FP_{2i}} + \frac{1}{TN_{2i}} = \frac{1}{n_{0+1i}} + \frac{1}{n_{0+0i}} \tag{8}$$

It should be noted that in the traditional approach for bivariate meta-analysis for a single diagnostic test [6,10], a bivariate meta-analysis is performed for the pairs of outcomes (i.e. $\hat{y}_{1i}$, $\hat{y}_{2i}$ or $\hat{y}_{3i}$, $\hat{y}_{4i}$). When we want to perform a joint analysis, following the general framework for multivariate meta-analysis [33,34], we will denote by $\mathbf{y}_i$ the vector containing the four different estimates and by $\boldsymbol{\beta}$, the vector of the overall means given by the following:

$$\mathrm{y}_i = \begin{pmatrix} \hat{y}_{1i} \\ \hat{y}_{2i} \\ \hat{y}_{3i} \\ \hat{y}_{4i} \end{pmatrix}, \text{and } \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} \tag{9}$$

In a multivariate random-effects setting, we assume that the four $\hat{y}_i$' s are distributed following a multivariate normal distribution:

$$\begin{bmatrix} \hat{y}_{1i} \\ \hat{y}_{2i} \\ \hat{y}_{3i} \\ \hat{y}_{4i} \end{bmatrix} \sim MVN\left\{ \begin{bmatrix} \beta_{1i} \\ \beta_{2i} \\ \beta_{3i} \\ \beta_{4i} \end{bmatrix}, [\mathrm{C}_i] \right\} \tag{10}$$

where $\mathbf{C}_i$ is the within-studies covariance matrix:

$$\mathrm{C}_i = \begin{pmatrix} s^2_{1i} & \rho_{w12i}s_{1i}s_{2i} & \rho_{w13i}s_{1i}s_{3i} & \rho_{w14i}s_{1i}s_{4i} \\ \rho_{w12i}s_{1i}s_{2i} & s^2_{2i} & \rho_{w23i}s_{2i}s_{3i} & \rho_{w24i}s_{2i}s_{4i} \\ \rho_{w13i}s_{1i}s_{3i} & \rho_{w23i}s_{2i}s_{3i} & s^2_{3i} & \rho_{w34i}s_{3i}s_{4i} \\ \rho_{w14i}s_{1i}s_{4i} & \rho_{w24i}s_{2i}s_{4i} & \rho_{w34i}s_{3i}s_{4i} & s^2_{4i} \end{pmatrix} \tag{11}$$

The means ($\beta_{1i}\beta_{2i}\beta_{3i}\beta_{4i}$) are considered random terms, distributed similarly as follows:

$$\left[\begin{bmatrix} \beta_{1i} \\ \beta_{2i} \\ \beta_{3i} \\ \beta_{4i} \end{bmatrix} \right] \sim MVN\left\{ \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}, [\boldsymbol{\Sigma}] \right\} \tag{12}$$

where $\boldsymbol{\Sigma}$ is the between-studies covariance matrix, given by the following:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \tau_1^2 & \rho_{B12}\tau_1\tau_2 & \rho_{B13}\tau_1\tau_3 & \rho_{B14}\tau_1\tau_4 \\ \rho_{B12}\tau_1\tau_2 & \tau_2^2 & \rho_{B23}\tau_2\tau_3 & \rho_{B24}\tau_2\tau_4 \\ \rho_{B13}\tau_1\tau_3 & \rho_{B23}\tau_2\tau_3 & \tau_3^2 & \rho_{B34}\tau_3\tau_4 \\ \rho_{B14}\tau_1\tau_4 & \rho_{B24}\tau_2\tau_4 & \rho_{B34}\tau_3\tau_4 & \tau_4^2 \end{pmatrix} \tag{13}$$

Thus, the final marginal model on which we base the inference is as follows:

$$\mathrm{y}_i \sim MVN(\beta, \Sigma + \mathrm{C}_i) \tag{14}$$

or we could rewrite the marginal model described in Eqn 14 as follows:

$$\begin{bmatrix} \hat{y}_{1i} \\ \hat{y}_{2i} \\ \hat{y}_{3i} \\ \hat{y}_{4i} \end{bmatrix} \sim MVN \left\{ \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}, \begin{bmatrix} s_{1i}^2 + \tau_1^2 & \rho_{w12i}s_{1i}s_{2i} + \rho_{B12}\tau_1\tau_2 & \rho_{w13i}s_{1i}s_{3i} + \rho_{B13}\tau_1\tau_3 & \rho_{w14i}s_{1i}s_{4i} + \rho_{B14}\tau_1\tau_4 \\ \rho_{w12i}s_{1i}s_{2i} + \rho_{B12}\tau_1\tau_2 & s_{2i}^2 + \tau_2^2 & \rho_{w23i}s_{2i}s_{3i} + \rho_{B23}\tau_2\tau_3 & \rho_{w24i}s_{2i}s_{4i} + \rho_{B24}\tau_2\tau_4 \\ \rho_{w13i}s_{1i}s_{3i} + \rho_{B13}\tau_1\tau_3 & \rho_{w23i}s_{2i}s_{3i} + \rho_{B23}\tau_2\tau_3 & s_{3i}^2 + \tau_3^2 & \rho_{w34i}s_{3i}s_{4i} + \rho_{B34}\tau_3\tau_4 \\ \rho_{w14i}s_{1i}s_{4i} + \rho_{B14}\tau_1\tau_4 & \rho_{w24i}s_{2i}s_{4i} + \rho_{B24}\tau_2\tau_4 & \rho_{w34i}s_{3i}s_{4i} + \rho_{B34}\tau_3\tau_4 & s_{4i}^2 + \tau_4^2 \end{bmatrix} \right\} \tag{15}$$

The diagonal elements of $\mathbf{C}_\mathrm{i}$ are the study-specific estimates of the variance that are assumed known [33] and are given in Eqns 5–9. Assuming s_i^2 terms to be known instead of to be estimated has little impact on the results, which is the basis for the traditional meta-analysis [35]. The off-diagonal elements of $\mathbf{C}_\mathrm{i}$ correspond to the pairwise within-studies covariances, for instance $\rho_{w12i}s_{1i}s_{2i} = \mathrm{cov}(\hat{y}_{1i}, \hat{y}_{2i})$ (with ρ_{w12i} being the within-studies correlation of the *TPR* and *FPR* of the first diagnostic test for study i that has to be known beforehand). Because $\mathrm{cov}(\hat{y}_{1i}, \hat{y}_{2i}) = 0$ and $\mathrm{cov}(\hat{y}_{3i}, \hat{y}_{4i}) = 0$ the remaining covariances need to be calculated (see next section). The diagonal elements of $\boldsymbol{\Sigma}$ are the between studies variances, whereas the off-diagonal elements correspond to the between studies covariances that are to be estimated from the data during the fitting procedure (for instance ρ_{B12} is the between studies correlation of the *TPR* and *FPR* of the first diagnostic test). We propose an unstructured specification for $\boldsymbol{\Sigma}$ or we could eliminate the parameters that have to be estimated by imposing a structured variant $\boldsymbol{\Sigma}$ when the number of tests increases (for instance we could set between-studies variances and correlations among tests to be equal).

The parameters estimated from the multivariate model of Eqn 14 can be directly compared for instance, using a Wald test or a bivariate test based on a chi-square distribution. Other measures such as the DOR, the AUC and the parameters of the SROC curve can be easily constructed [10] and compared, as we shall show in the next sections. A direct extension of the method for more than two diagnostic tests is straightforward.

2.2. *Calculation of the within-studies covariances*

The method used here is a special case of a general methodology that has been recently proposed [29]. Thus, we may proceed with some probability calculations (Appendix A) using the properties of the covariance function in order to derive the following:

$$\begin{aligned} \mathrm{cov}(\hat{y}_{1i}, \hat{y}_{3i}) &= \frac{n_{111i}}{(n_{111i} + n_{110i})(n_{111i} + n_{101i})} - \frac{n_{110i}}{(n_{111i} + n_{110i})(n_{110i} + n_{100i})} \\ &- \frac{n_{101i}}{(n_{101i} + n_{100i})(n_{111i} + n_{101i})} + \frac{n_{100i}}{(n_{101i} + n_{100i})(n_{110i} + n_{100i})} \\ &= \sum_l \sum_p (-1)^{l-p} \left(\frac{n_{1lpi}}{n_{1l+i} n_{1+pi}} \right) \end{aligned} \tag{16}$$

$$\begin{aligned} \mathrm{cov}(\hat{y}_{2i}, \hat{y}_{4i}) &= \frac{n_{011i}}{(n_{011i} + n_{010i})(n_{011i} + n_{001i})} - \frac{n_{010i}}{(n_{011i} + n_{010i})(n_{010i} + n_{000i})} \\ &- \frac{n_{001i}}{(n_{001i} + n_{000i})(n_{011i} + n_{001i})} + \frac{n_{000i}}{(n_{001i} + n_{000i})(n_{010i} + n_{000i})} \\ &= \sum_l \sum_p (-1)^{l-p} \left(\frac{n_{0lpi}}{n_{0l+i} n_{0+pi}} \right) \end{aligned} \tag{17}$$

$$\mathrm{cov}(\hat{y}_{1i}, \hat{y}_{4i}) = \mathrm{cov}(\hat{y}_{2i}, \hat{y}_{3i}) = 0 \tag{18}$$

Attention should be paid when the number of the participants is not equal for the two diagnostic tests for a given published study. In such a case, we can extrapolate the counts of the diagnostic test with the smaller sum to match the larger one and proceed based on the extrapolated counts.

2.3. Handling missing information

As it is apparent from Eqns 16 and 17, the exact counts of Table I need to be known for the calculation of the within studies covariances. In the simplest case, where, for each study i, the counts of Table I are known, the calculation of the covariances is straightforward. In practical applications however, we do not expect all studies to report such information.

When this information is available only for a subset of the studies, we can proceed as follows. First, for these «complete» studies, we can calculate, the *conditional ORs*, which reflect the association of the two diagnostic tests for a given disease status:

$$\widehat{OR}_{Y_i=1} = \frac{n_{111i}n_{100i}}{n_{101i}n_{110i}} \tag{19}$$

$$\widehat{OR}_{Y_i=0} = \frac{n_{011i}n_{000i}}{n_{010i}n_{001i}} \tag{20}$$

Standard random effects meta-analysis can be used in order to derive a combined estimate of the *conditional ORs* for the diseased and the non-diseased individuals, across the «complete» subset of studies. After obtaining these combined estimates, we can use them on the «incomplete studies» (those for which the interior cells of Table I are not given) along with the knowledge of the marginal counts of Table I for each disease status and derive expressions of the counts for each cell solving a system of equations (Appendix B).

If the *conditional ORs* do not differ according to disease status or if we have some prior knowledge that this is the case, the iterative proportional fitting (IPF) algorithm [36] can be used in order to obtain the interior cells, even if no study reports the full partial table (Table I). It should be noted that the IPF algorithm assumes that there is no three-way interaction in the $2\times2\times2$ contingency table. In other words, it assumes that the two ORs of Eqns 19 and 20 are equal. Thus, this approach (although not perfect) is more general than the assumption of conditional independence that is usually employed in analysis [37,38] and meta-analysis of diagnostic tests with no gold standard [14,25]

2.4. Recovering the summary receiver operating characteristic method and the area under the curve

Under the SROC approach [7,13], the logit-trasformed TPR_{ji} and FPR_{ji} are used and a linear regression of their difference (D_{ji}) on their sum (S_{ji}) is performed. The parameters of the regression of D_{1i} on S_{1i} for the first test can be expressed using the parameters of the bivariate model with τ_1^2 and τ_2^2 denoting the between studies variances of y_{1i} and y_{2i} respectively, as well as $\rho_{B12}\tau_1\tau_2$ the between studies covariances of y_{1i} and y_{2i}. In particular, from Eqn 12, it turns out that the covariance of D_{1i} and S_{1i} is equal to $\tau_1^2-\tau_2^2$, and the variance of S_{1i} is equal to $\tau_1^2+\tau_2^2+2\rho_{B12}\tau_1\tau_2$. Thus, the slope is $b_1=\left(\tau_1^2-\tau_2^2\right)/\left(\tau_1^2+\tau_2^2+2\rho_{B12}\tau_1\tau_2\right)$; the intercept is $a_1=\beta_1-\beta_2-b(\beta_1+\beta_2)$, and the residual variance of the regression is given by $\sigma^2_{D_1|S_1}=\left(\tau_1^2+\tau_2^2-2\rho_{B12}\tau_1\tau_2\right)-\frac{\left(\tau_1^2-\tau_2^2\right)^2}{\tau_1^2+\tau_2^2+2\rho_{B12}\tau_1\tau_2}$ [10]. These can easily be estimated by plugging in the model parameter estimates; the standard errors are calculated using the Delta method [39]. Similar expressions can be derived for the second diagnostic test by using the appropriate estimates of the between studies variance–covariance matrix (i.e. τ_3^2, τ_4^2, $\rho_{B34}\tau_3\tau_4$). It is worth-mentioning that the coefficient b_j represents the dependence of the test accuracy on threshold. If $b_j\approx0$, then the studies are homogeneous and can be summarized by an overall DOR_j noting that $a_j=\ln(OR_j)$.

When the parameters a_j and b_j are estimated, the relationship between TPR_{ji} and FPR_{ji} can be deduced:

$$TPR_{ji} = \frac{\exp\left(\frac{a_j}{1-b_j}\right)\left(\frac{FPR_{ji}}{1-FPR_{ji}}\right)^{\left(1+b_j\right)/\left(1-b_j\right)}}{1+\exp\left(\frac{a_j}{1-b_j}\right)\left(\frac{FPR_{ji}}{1-FPR_{ji}}\right)^{\left(1+b_j\right)/\left(1-b_j\right)}} \tag{21}$$

Equation 21 gives TPR_{ji} at any given value of FPR_{ji} and hence the entire SROC curve. While there may be some interest in identifying particular points on the curve, it is often useful to have an overall

summary measure of the behaviour of the curve. Perhaps, the most appropriate such measure is the AUC, which can be calculated as [40]:

$$AUC_j = \int_0^1 \frac{\exp\left(\frac{a_j}{1-b_j}\right)\left(\frac{x}{1-x}\right)^{(1+b_j)/(1-b_j)}}{1+\exp\left(\frac{a_j}{1-b_j}\right)\left(\frac{x}{1-x}\right)^{(1+b_j)/(1-b_j)}}dx \tag{22}$$

Using the delta method, an approximate variance for $\hat{AUC}_j$ is as follows:

$$\begin{aligned}\operatorname{var}\left(\hat{AUC}_j\right) = &\left(\frac{\partial AUC_j}{\partial a_j}\right)^2 \operatorname{var}\left(\hat{a}_j\right) + \left(\frac{\partial AUC_j}{\partial b_j}\right)^2 \operatorname{var}\left(\hat{b}_j\right) \\ &+ 2\left(\frac{\partial AUC_j}{\partial a_j}\right)\left(\frac{\partial AUC_j}{\partial b_j}\right)\operatorname{cov}\left(\hat{a}_j, \hat{b}_j\right)\end{aligned} \tag{23}$$

where:

$$\frac{\partial\left(AUC_j\right)}{\partial a_j} = \left(\frac{1}{1-b_j}\right)\exp\left(\frac{\alpha_j}{1-b_j}\right)\int_0^1 \frac{\left(\frac{x}{1-x}\right)^{(1+b_j)/(1-b_j)}}{\left(1+\left(\frac{x}{1-x}\right)^{(1+b_j)/(1-b_j)}\exp\left(\frac{a_j}{1-b_j}\right)\right)^2}dx \tag{24}$$

$$\frac{\partial\left(AUC_j\right)}{\partial b_j} = \left(\frac{1}{1-b_j}\right)^2\exp\left(\frac{a_j}{1-b_j}\right)\int_0^1 \frac{\left(\frac{x}{1-x}\right)^{(1+b_j)/(1-b_j)}\left(a_j+2\ln\left(\frac{x}{1-x}\right)\right)}{\left(1+\left(\frac{x}{1-x}\right)^{(1+b_j)/(1-b_j)}\exp\left(\frac{a}{1-b_j}\right)\right)^2}dx \tag{25}$$

In the homogeneous case, $b_j = 0$, and the general expression (Eqn 22) becomes [40]:

$$AUC_{\text{hom}_j} = \frac{DOR_j}{\left(DOR_j-1\right)^2}\left[\left(DOR_j-1\right) - \ln\left(DOR_j\right)\right] \tag{26}$$

where AUC_{hom_j} indicates the AUC_j for homogeneous studies and $DOR_j = \exp(a_j)$. If $a_j = 0$, then $AUC_{\text{hom}_j} = \frac{1}{2}$. Although only valid for homogeneous studies, Eqn 26 is a useful upper bound and provides a good approximation for AUC_j in heterogeneous studies. Using the delta method, we may also obtain:

$$SE\left(\hat{AUC}_{\text{hom}_j}\right) = \frac{1}{\left(DOR_j-1\right)^3}\left[\left(DOR_j+1\right)\ln DOR_j - 2\left(DOR_j-1\right)\right]SE\left(\hat{DOR}_j\right) \tag{27}$$

Finally, a formal test for the equality of the AUC for the two tests (i.e. $AUC_1 = AUC_2$) could be derived using the properties of the AUC [40] and the Delta method. This would lead to the formulation of the null hypothesis $H_0: d = AUC_1 - AUC_2 = 0$, $H_a: d \neq 0$. The variance of $\hat{d}$ will be equal to **GVG'** where **V** is the estimated variance-covariance matrix and **G** is the derivative matrix of d with respect to the vector of estimated coefficients a_j, b_j $\left[\frac{\partial(d)}{\partial a_1} \quad \frac{\partial(d)}{\partial b_1} \quad \frac{\partial(d)}{\partial a_2} \quad \frac{\partial(d)}{\partial b_2}\right]$. Thus, **GVG'** will be equal to

$$\begin{aligned}\mathbf{GVG'} = &\frac{\partial(d)}{\partial a_1}\left\{\frac{\partial(d)}{\partial a_1}\operatorname{var}(\hat{a}_1) + \frac{\partial(d)}{\partial b_1}\operatorname{cov}\left(\hat{a}_1, \hat{b}_1\right) + \frac{\partial(d)}{\partial a_2}\operatorname{cov}(\hat{a}_1, \hat{a}_2) + \frac{\partial(d)}{\partial b_2}\operatorname{cov}\left(\hat{a}_1, \hat{b}_2\right)\right\} \\ &+\frac{\partial(d)}{\partial b_1}\left\{\frac{\partial(d)}{\partial a_1}\operatorname{cov}\left(\hat{a}_1, \hat{b}_1\right) + \frac{\partial(d)}{\partial b_1}\operatorname{var}\left(\hat{b}_1\right) + \frac{\partial(d)}{\partial a_2}\operatorname{cov}\left(\hat{a}_2, \hat{b}_1\right) + \frac{\partial(d)}{\partial b_2}\operatorname{cov}\left(\hat{b}_1, \hat{b}_2\right)\right\} \\ &+\frac{\partial(d)}{\partial a_2}\left\{\frac{\partial(d)}{\partial a_1}\operatorname{cov}(\hat{a}_1, \hat{a}_2) + \frac{\partial(d)}{\partial b_1}\operatorname{cov}\left(\hat{a}_2, \hat{b}_1\right) + \frac{\partial(d)}{\partial a_2}\operatorname{var}(\hat{a}_2) + \frac{\partial(d)}{\partial b_2}\operatorname{cov}\left(\hat{a}_2, \hat{b}_2\right)\right\} \\ &+\frac{\partial(d)}{\partial b_2}\left\{\frac{\partial(d)}{\partial a_1}\operatorname{cov}\left(\hat{a}_1, \hat{b}_2\right) + \frac{\partial(d)}{\partial b_1}\operatorname{cov}\left(\hat{b}_1, \hat{b}_2\right) + \frac{\partial(d)}{\partial a_2}\operatorname{cov}\left(\hat{a}_2, \hat{b}_2\right) + \frac{\partial(d)}{\partial b_2}\operatorname{var}\left(\hat{b}_2\right)\right\}\end{aligned} \tag{28}$$

with the partial derivatives for diagnostic test j ($j = 1,2$) given by the following:

$$\frac{\partial(d)}{\partial a_j} = (-1)^{j+1}\left(\frac{1}{1-b_j}\right)\exp\left(\frac{\alpha_j}{1-b_j}\right)\int_0^1 \frac{\left(\frac{x}{1-x}\right)^{(1+b_j)/(1-b_j)}}{\left(1+\left(\frac{x}{1-x}\right)^{(1+b_j)/(1-b_j)}\exp\left(\frac{a_j}{1-b_j}\right)\right)^2}dx \tag{29}$$

$$\frac{\partial(d)}{\partial b_j} = (-1)^{j+1}\left(\frac{1}{1-b_j}\right)^2\exp\left(\frac{a_j}{1-b_j}\right)\int_0^1 \frac{\left(\frac{x}{1-x}\right)^{(1+b_j)/(1-b_j)}\left(a_j+2\ln\left(\frac{x}{1-x}\right)\right)}{\left(1+\left(\frac{x}{1-x}\right)^{(1+b_j)/(1-b_j)}\exp\left(\frac{a_j}{1-b_j}\right)\right)^2}dx \tag{30}$$

In the homogeneous case (i.e. $b_j=0$), the general expression (Eqn 28) for estimating the variance of $\widehat{d}$ is simplified because the terms including b_j's are cancelled out. In such a case, the variance of $\widehat{d}_{\text{hom}}$is:

$$\text{var}\left(\widehat{d}_{\text{hom}}\right) = \left(\frac{\partial d_{\text{hom}}}{\partial a_1}\right)^2\text{var}(\widehat{a}_1) + \left(\frac{\partial d_{\text{hom}}}{\partial a_2}\right)^2\text{var}(\widehat{a}_2) + 2\left(\frac{\partial d_{\text{hom}}}{\partial a_1}\right)\left(\frac{\partial d_{\text{hom}}}{\partial a_2}\right)\text{cov}(\widehat{a}_1,\widehat{a}_2) \tag{31}$$

and the partial derivatives for diagnostic test j ($j=1,2$) can be calculated as follows:

$$\frac{\partial(d_{\text{hom}})}{\partial a_j} = (-1)^{j+1}\frac{\exp(\alpha_j)\left[\exp(\alpha_j)(\alpha_j-2)+\alpha_j+2\right]}{\left[\exp(\alpha_j)-1\right]^3} \tag{32}$$

3. Simulation study

To evaluate the proposed methodology, we conducted simulation studies. Section C of our supplementary material provides the simulation procedures and the results from this study. We compare the proposed multivariate method with the standard bivariate approach. Briefly, our simulations suggest that in the scenarios with 20 studies and 200 participants per study, the multivariate method produces

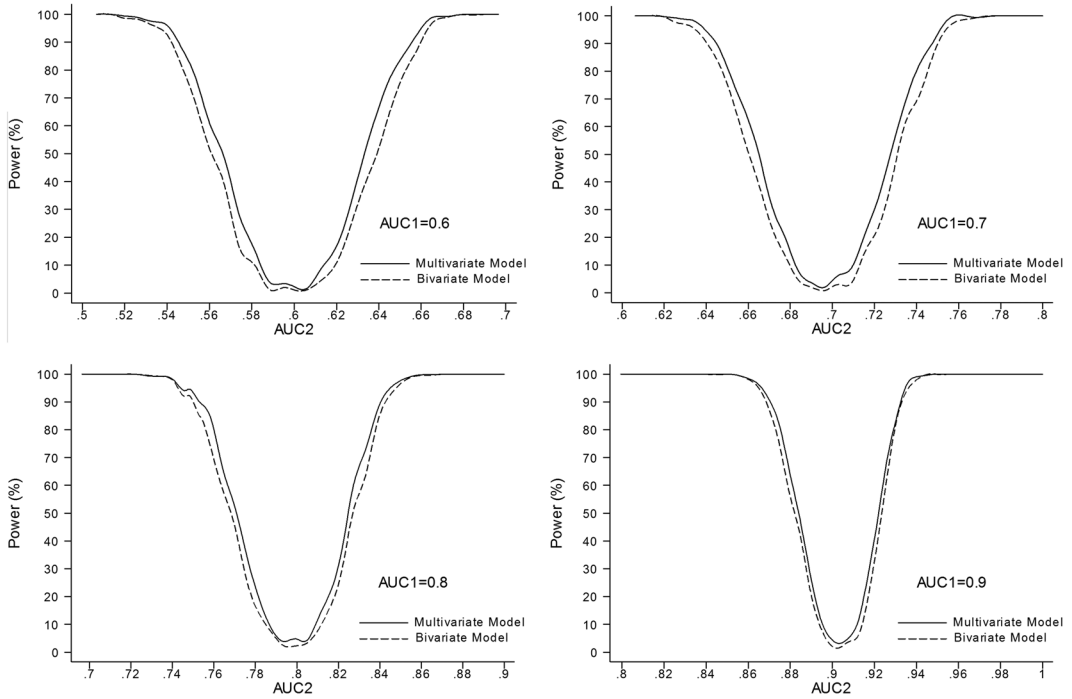

Figure 1. Power calculations to detect differences in $\widehat{AUC}_1$-$\widehat{AUC}_2$ under the multivariate and bivariate method for meta-analysis of diagnostic studies. We conducted 500 replications with 20 studies in each meta-analysis. The disease prevalence was set at 50% with 200 participants in each study. For simplicity, we used the exact analytical expression (Eqn 26) in order to calculate the area under the curve (*AUC*) assuming that our studies are homogeneous.

unbiased estimates and has greater power to detect small differences in sensitivity or specificity preserving the nominal type I error rate when compared with the bivariate approach. Moreover, the multivariate method has always greater power (up to 12%) to detect small differences in the $\widehat{AUC}$ (Figure 1). The gain of our method was less pronounced in the scenarios with large number of participants (>500) per study that report results of both tests (no missing data) and when there are large differences in either sensitivity or specificity.

4. Application of the method

The method was applied in the data obtained from a meta-analysis that aimed to determine whether anti-cyclic citrullinated peptide antibody (anti-CCP antibody) identifies more accurately patients with rheumatoid arthritis than does of rheumatoid factor (RF) [41] (Appendix D). As far as anti-CCP antibodies are concerned, we included studies using the CCP2 assay. In particular, a total of 50 and 29 studies provided information concerning only RF or anti-CCP2 antibody, respectively. Twenty two studies assessed both RF and anti-CCP2 antibody for diagnosing rheumatoid arthritis.

From the 22 studies, which assessed simultaneously both diagnostic tests, detailed data for the counts of Table I were available for two studies. A total of six studies reported estimates for sensitivity and specificity when both diagnostic tests where positive or when at least one of the tests is positive. In such a case, we show that the counts of Table I can be accurately reconstructed (Appendix E). One study provided information for sensitivity, specificity, positive and negative predictive value for those individuals who had a negative test result for one of the tests (the RF diagnostic test in particular). Similarly, a set of corresponding equations can be derived for this case, and the counts for Table I were also calculated (Appendix E). The total number of the participants was not equal for both diagnostic tests in one study, so we extrapolated the counts of the smaller sum to match the larger one.

Finally, detailed data for the counts of Table I were available for six studies for diseased participants only. There was lack of data for a total of seven studies that evaluated both tests, and these numbers needed to be imputed. A combined estimate of the *conditional OR* for the association of the two diagnostic tests was calculated for non-diseased and diseased individuals ($\widehat{OR}_{Y_i=0} = 6.537$with $I^2=52.7\%$ and $\tau^2=0.7817$, $\widehat{OR}_{Y_i=1} = 12.112$ with $I^2=52.8\%$ and $\tau^2=0.2298$, respectively) using a standard random effects meta-analysis [42]. These estimates differ significantly, indicating that there is a three-way interaction in the contingency table. Consequently, our best guess is to assume that these estimates are close to the true values for the respective *ORs* in the studies that did not provide sufficient information. Then, using the equations illustrated in Appendix B, we can derive the counts of Table I. We used mvmeta command [43] in Stata (Appendix F) under an unstructured specification for $\mathbf{\Sigma}$. We need to emphasize that the largest portion of the code refers to the estimation of the counts for the studies that did not provide sufficient information, the calculation of the parameters of the regression of D_{ji} on S_{ji} and the calculation of the AUC_j. For comparison, the parameters of the bivariate model were also estimated.

The $\widehat{TPR}$ of the two diagnostic tests does not differ significantly (Wald test: $z=-0.20$; p-value $=0.834$, Table III). Anti-CCP2 antibody, however, had lower $\widehat{FPR}$ and thus higher specificity than RF (Wald test: $z=5.57$; p-value <0.001) (Figure 2). From the 22 studies, which assessed simultaneously both tests, the within-studies correlation of the $\widehat{TPR}$ of the two tests (i.e. $\widehat{\rho}_{w13i}$) varied from 0.148–0.684 with a mean value equal to 0.463 while the correlation of the $\widehat{FPR}$ of the two tests (i.e. $\widehat{\rho}_{w24i}$) varied from 0–0.669 with

Table III. Estimates of the multivariate and bivariate model of $\widehat{TPR}_{ji}$ and $\widehat{FPR}_{ji}$ (*logits*) for the two diagnostic tests ($\widehat{\beta}_1$-$\widehat{\beta}_2$ for rheumatoid factor and $\widehat{\beta}_3$-$\widehat{\beta}_4$ for anti-cyclic citrullinated peptide 2 antibody).

	Multivariate model			Bivariate model		
	Mean (95% CI)	z	$\widehat{\tau}^2$	Mean (95% CI)	z	$\widehat{\tau}^2$
$\widehat{\beta}_1$	0.750 (0.538, 0.962)	6.94	0.537	0.708 (0.491, 0.925)	6.39	0.534
$\widehat{\beta}_2$	−1.845 (−2.141, -1.550)	-12.23	0.977	−1.876 (−2.173, -1.579)	-12.37	0.990
$\widehat{\beta}_3$	0.774 (0.534, 1.015)	6.31	0.449	0.823 (0.571, 1.075)	6.41	0.406
$\widehat{\beta}_4$	−2.986 (−3.310, -2.663)	-18.08	0.519	−2.909 (−3.227, -2.590)	-17.90	0.527
$\widehat{\rho}_{B12}=0.212$, $\widehat{\rho}_{B13}=0.779$, $\widehat{\rho}_{B14}=0.386$, $\widehat{\rho}_{B23}=0.194$, $\widehat{\rho}_{B24}=0.410$, $\widehat{\rho}_{B34}=0.487$				$\widehat{\rho}_{B12}=0.220$, $\widehat{\rho}_{B34}=0.477$		

TPR, true positive rate; FPR, false positive rate.

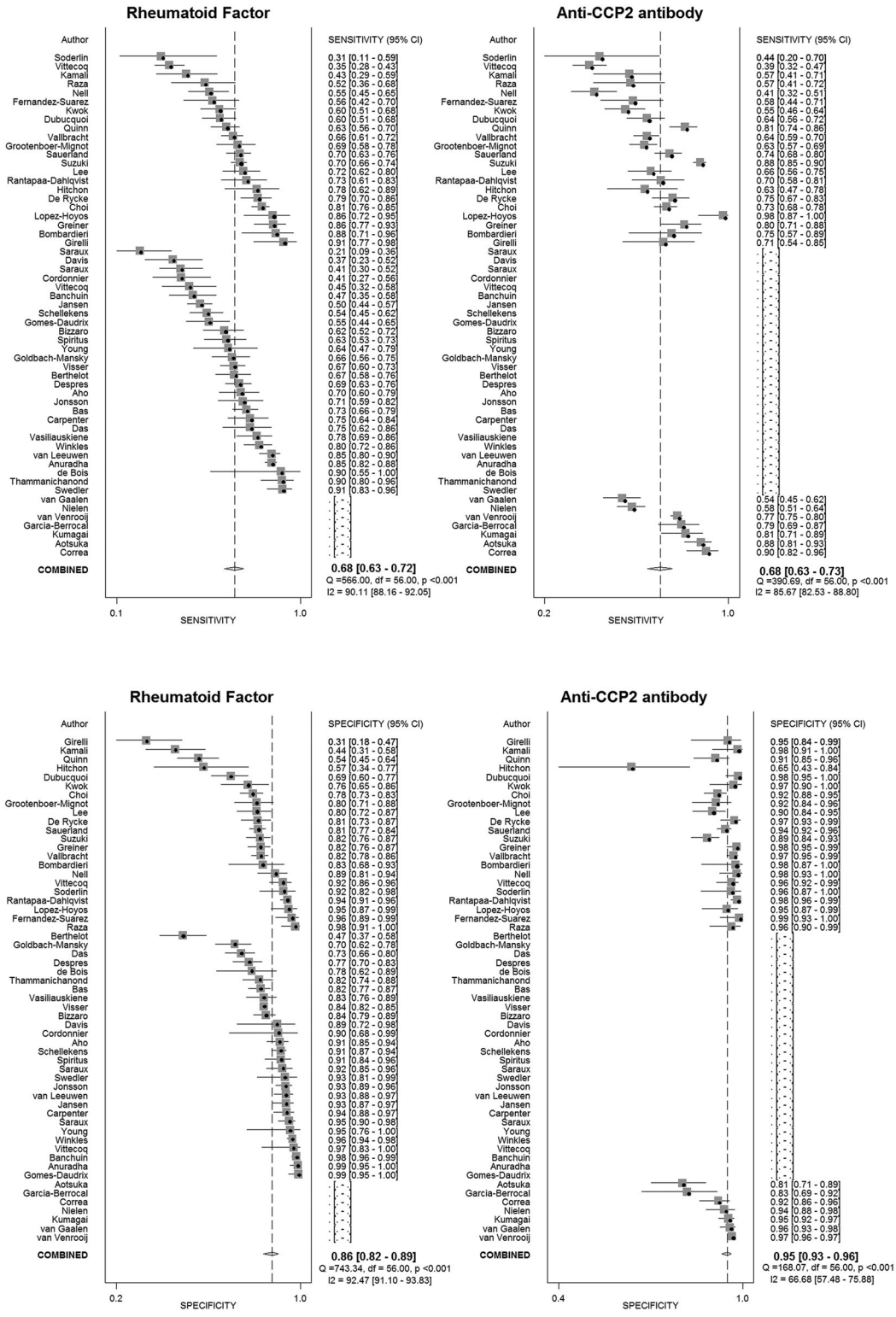

Figure 2. Forest plots of the sensitivity and specificity of the two diagnostic tests, obtained by the multivariate meta-analysis. CCP, cyclic citrullinated peptide.

a mean value equal to 0.197. The between-studies correlation of the $\widehat{TPR}$ and $\widehat{FPR}$ of the two tests were estimated as $\widehat{\rho}_{B13} = 0.779$ and $\widehat{\rho}_{B24} = 0.410$. Furthermore, using the estimates of the between studies covariances, the parameters of the regression of D_{ji} on S_{ji} were calculated as well as the SROC curve [10] with $\widehat{b}_1 = -0.241$ [95% confidence interval (CI): -0.505, 0.022], $\widehat{b}_2 = -0.049$ (95% CI: -0.344,

0.246), $\hat{a}_1 = 2.331$ (95% CI: 1.901, 2.761) and $\hat{a}_2 = 3.653$ (95% CI: 2.984, 4.323) (Figure 3). $\widehat{AUC}$ for RF and Anti-CCP2 antibody was calculated as 0.825 and 0.927 respectively with these two values showing a statistically significant difference, suggesting that in overall, the anti-CCP2 antibody is better compared with RF ($z = -3.54$; p-value < 0.001) (Figure 4). We should also mention that model misspecification (i.e. ignoring the differences in the conditional ORs in patients and controls) results in no significant changes in the pooled estimates or in the conclusions regarding the superiority of anti-CCP2 antibody.

5. Discussion

We described here a simple yet powerful method for meta-analysis and comparison of diagnostic tests by extending the bivariate approach [10,15]. The key point of this method is that the within-studies covariances can be calculated via a closed form expression using a recently published method [29]. The method inherits all the advantages derived from the bivariate model; that is, the between-studies correlation among sensitivity and specificity is accounted for. Point estimates and confidence intervals for these two parameters can be calculated, and a construction of SROC curve is also feasible. Parameters such as the *DOR*, the *AUC* and the likelihood ratios can also be calculated and compared for the two tests. Most importantly, the method can easily incorporate (under a missing at random assumption) studies

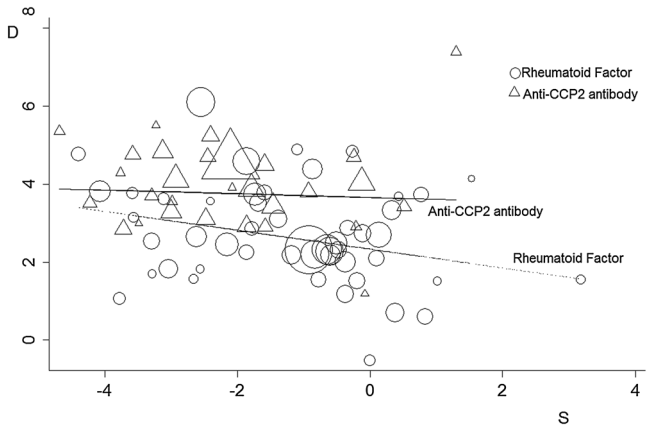

Figure 3. Plot of the sum of logit-trasforms of $\widehat{TPR}_{ji}$ and $\widehat{FPR}_{ji}(\hat{S}_{ji})$ on their difference ($\widehat{D}_{ji}$) of the two diagnostic tests under the summary receiver operating characteristic method. The regression coefficients are obtained by the multivariate meta-analysis model as described in the text. CCP, cyclic citrullinated peptide; TPR, true positive rate; FPR, false positive rate.

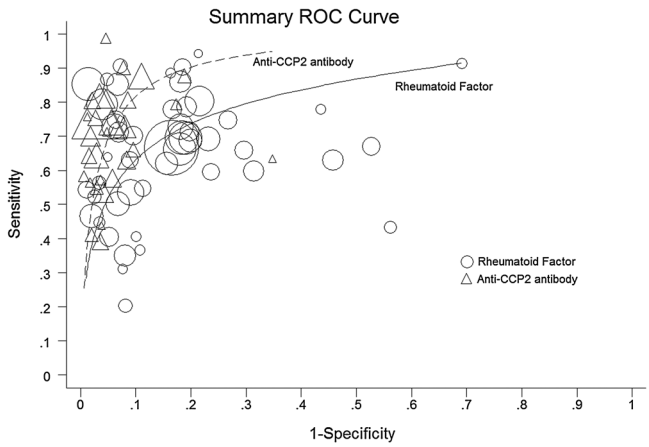

Figure 4. Summary receiver operating characteristic (SROC) curve obtained from the multivariate model of the two diagnostic tests ($\widehat{AUC} = 0.825$ and 0.927 for rheumatoid factor (RF) and anti-cyclic citrullinated peptide (CCP) 2, respectively). AUC, area under the curve.

reporting only one of the tests and thus allows borrowing strength from external studies [30]; it can be easily fitted using standard software used for multivariate meta-analysis [31,32], and it allows for the direct formal comparison of the different tests using formal techniques (Wald test). As a standard multivariate model, it can also incorporate covariates, which may be correlated with the disease process, the diagnostic testing procedure or account for strata with different disease prevalence. This approach thus yields a random-effects meta-regression [44], where various study level characteristics can be used as linear predictors, in order to estimate the extent to which these covariates explain the observed heterogeneity.

Our simulations reveal that in the realistic scenario that we evaluated, our method is more likely to produce unbiased estimates and to preserve the nominal type-I error rate, whereas it has greater power to detect small differences in sensitivity or specificity. When the meta-analysis includes a large number of studies, each having a large number of participants, that report the results of both tests and when these tests have large differences in either sensitivity or specificity, we expect that the gain of using the method will be negligible. However, we do not expect in any case the method to be worse compared with standard bivariate meta-analysis. Moreover, simulations under a misspecified model (i.e. assuming that there is no three-way interaction and using the IPF algorithm) showed that the method is still better compared with the standard bivariate approach. These conclusions are in agreement with previous simulation results concerning the superiority of bivariate meta-analysis over univariate one but also have theoretical justification [45–47]. Algebraic calculations reveal that for $\widehat{TPR}$ and $\widehat{FPR}>0.5$, the covariance is always positive, and thus, we expect the variance of the differences $(\widehat{\beta}_1\text{-}\widehat{\beta}_3,\ \widehat{\beta}_2\text{-}\widehat{\beta}_4)$ to be always smaller when we take this covariance into account. However, the covariance is rather small compared with the variance, and its contribution decreases with increased sample size or increased between studies heterogeneity. All these taken together, explain the fact that we do not see remarkably better results (at least for the range of values that we are interested in) but also provide the assurance that the multivariate method will be at least as accurate as the bivariate one, under any circumstances. The same rationale holds also for the power to detect differences in the $\widehat{AUC}$. When the real difference between the tests is large, both methods have 100% power to detect it, but the multivariate method has always greater power (up to 12%) to detect small differences that may be important in clinical practice.

Some other methods for comparing in a meta-analysis of two diagnostic tests have been proposed in the literature. As far as diagnostic *ORs* are concerned [12], the simple comparison of their log-transformations or those derived using SROC method ignores the correlation of the two diagnostic tests, in cases where the same patients are measured. When the tests are applied to different populations (which does not usually happen), the test is valid, but still, it involves only the *DOR* and not sensitivity or specificity. The CROR method [23] is simple but requires individual patient data for both tests which are not always available. A very important side-effect of our work that needs to be emphasized is that the methods for reconstructing the tables that we presented in the Appendix E can be directly used with the CROR method. Thus, a method originally developed for use mainly with individual data can be used using summary data collected from the literature. The particular approach, after the imputation procedure, has the advantage of being easy to use and requires no more than a standard software for univariate meta-analysis. As a proof of principle, from the original data, nine studies reported data sufficient to calculate CROR. Combining these studies, yield an estimate of CROR of 0.144 (95% C.I.: 0.066, 0.317) for the RF over the anti-CCP2 antibody. However, using the equations of the Appendix E, we were able to reconstruct the data for an additional 13 studies, and this enables us to calculate a more precise estimate of CROR of 0.152 (95% C.I.: 0.090, 0.256), suggesting a superiority of the anti-CCP2 antibody over RF. Nevertheless, this method has the additional disadvantage that it cannot incorporate studies that report only one test, whereas a direct comparison of sensitivity, specificity or AUC is not feasible and neither is it the construction of a summary SROC curve. We also need to comment on the Bayesian approach proposed by Trikalinos and coworkers [28], an approach that shares many common features with this work. The main difference besides the Bayesian formalism is the fact that Trikalinos and coworkers model the correlation of the two tests as a random parameter, whereas in this work, we use directly the actual study-specific correlations. When all studies report the cross-tabulation table, we expect that the two approaches will yield nearly identical results. Moreover, the simulations we conducted clearly showed that the model is robust even under severe misspecification. Thus, the main advantage of our approach is the simplicity and the fact that it can be fitted with standard software, without however losing in accuracy. One additional advantage as we already noted is that it inherits all the advantages derived from the standard bivariate model and additionally we made a lot of effort to derive tests for the parameters of the SROC curve and AUC.

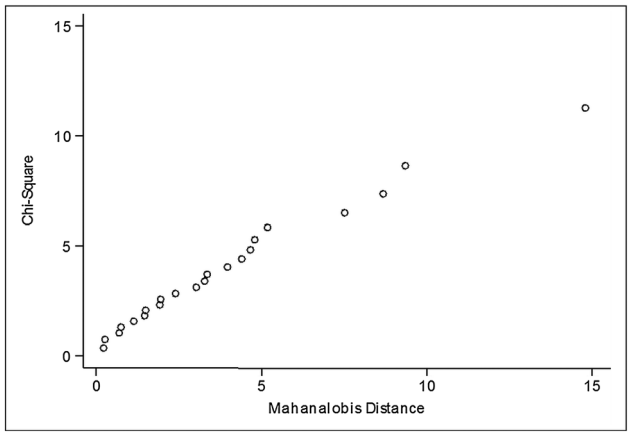

Figure 5. Plot check for multivariate normality of logit$\left(\widehat{TPR}_{ji}\right)$ and logit$\left(\widehat{FPR}_{ji}\right)$ for the 22 studies, which assessed simultaneously both diagnostic tests for discriminating patients with rheumatoid arthritis. The estimates are approximately normal as the graph approximates a 45° line.

We have to emphasize that our method is designed to be used for comparison of diagnostic tests in situations when a gold standard exists. On the contrary, some other multivariate random effects methods have been proposed to model simultaneously the sensitivity and the specificity of the two tests along with the prevalence of the disease, in situations when a gold standard is not available [14,25]. These methods are based on the conditional independence assumption [26,37,38], which is rather strong, but it is not needed in our approach. When this assumption is not met, a correlation among the diagnostic tests can be imposed, but nevertheless, our approach is more direct because it calculates the correlation directly from the joint distribution of the two tests and the disease status. Assumptions about the missing data mechanism become crucial as the amount of missing data increases when multiple diagnostic tests are used and the missing at random assumption may hold only in special situations, excluding of course cases where test ordering depends on health status. Because we used the classical model for multivariate meta-analysis, the method can handle data missing completely at random or missing at random [32,47]. In general, simple diagnostics for the case of informative missing mechanisms can be constructed, for instance by performing subgroup analyses for the studies that report both tests versus studies that report only one of the tests.

Our method uses summary data available on the published reports. The statistical theory behind the proposed methodology is very simple and is based on standard large sample approximations and normality assumptions [39,48] that are in everyday use by researchers performing meta-analyses of published data. In our illustrative example with the two diagnostic tests used for discriminating patients with Rheumatoid Arthritis, Henze–Zirkler's test [49] provided by the multnorm command in Stata [50,51], provided no indication that the multivariate normality assumption is violated (Figure 5). In some extreme cases, some practical issues may arise using the approximate likelihood inference [52,53]. Theoretically, the method is expected to fail when the sample size is very small or when there are several studies with small, or even zero, cell counts. In such situations, relying on the usual continuity correction that consists of adding ½ to the cell counts is the only option that additionally seems to perform quite well [54]. Another option could be to use multinomial distributions but to the authors' knowledge of multinomial likelihood for the particular problem cannot be fit using xtmelogit routine in Stata or lmer in R. Optimizing the likelihood for performing multivariate meta-analysis using the multinomial likelihood involves calculating complicated integrals numerically, which is outside the scope of the present work.

Overall, we presented a simple and powerful method for performing meta-analysis and comparison of diagnostic tests. The method can be fitted in nearly all statistical packages. In Appendix F, we give illustrative code in Stata, and we hope that this method will be widely used in future studies.

Acknowledgements

The authors would like to thank the two anonymous reviewers and the associate editor whose constructive criticism helped in improving the quality of the manuscript. The authors are financially supported by the project 'Integration of Data from Multiple Sources' (IntDaMus), which is implemented under the 'ARISTEIA II' Action

of the 'OPERATIONAL PROGRAMME EDUCATION AND LIFELONG LEARNING' and is co-funded by the European Social Fund (ESF) and National Resources.

References

1. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology* 2010; **5**(9):1315–1316.
2. Glass G. Primary, secondary and meta-analysis of research. *Educational Research* 1976; **5**:3–8.
3. Greenland S. In *Meta-Analysis, in Modern Epidemiology*, Rothman KJ, Greenland S (eds). Lippincott Williams & Wilkins, 1998; 643–673.
4. Normand SL. Meta-analysis: formulating, evaluating, combining, and reporting. *Statistics in Medicine* 1999; **18**(3):321–359.
5. Petiti DB. Monographs in epidemiology and biostatistics. In *Meta-Analysis Decision Analysis and Cost-Effectiveness Analysis*, Vol. **24**. Oxford University Press, 1994.
6. Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA. A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics* 2007; **8**(2):239–251.
7. 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. *Statistics in Medicine* 1993; **12**(14):1293–1316.
8. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in Medicine* 2001; **20**(19):2865–2884.
9. Macaskill P. Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. *Journal of Clinical Epidemiology* 2004; **57**(9):925–932.
10. Arends LR, Hamza TH, van Houwelingen JC, Heijenbrok-Kal MH, Hunink MG, Stijnen T. Bivariate random effects meta-analysis of ROC curves. *Medical Decision Making* 2008; **28**(5):621–638.
11. Deeks JJ. Systematic reviews in health care: systematic reviews of evaluations of diagnostic and screening tests. *BMJ* 2001; **323**(7305):157–162.
12. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM. The diagnostic odds ratio: a single indicator of test performance. *Journal of Clinical Epidemiology* 2003; **56**(11):1129–1135.
13. Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. *Medical Decision Making* 1993; **13**(4):313–321.
14. Dendukuri N, Schiller I, Joseph L, Pai M. Bayesian meta-analysis of the accuracy of a test for tuberculous pleuritis in the absence of a gold standard reference. *Biometrics* 2012; **68**(4):1285–1293.
15. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology* 2005; **58**(10):982–990.
16. Chu H, Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *Journal of Clinical Epidemiology* 2006; **59**(12):1331–1332.
17. Chen Y, Liu Y, Ning J, Nie L, Zhu H, Chu H. A composite likelihood method for bivariate meta-analysis in diagnostic systematic reviews. *Statistical Methods in Medical Research* 2014. [Epub ahead of print].
18. Chu H, Nie L, Cole SR, Poole C. Meta-analysis of diagnostic accuracy studies accounting for disease prevalence: alternative parameterizations and model selection. *Statistics in Medicine* 2009; **28**(18):2384–2399.
19. Li J, Fine JP. Assessing the dependence of sensitivity and specificity on prevalence in meta-analysis. *Biostatistics* 2011; **12**(4):710–722.
20. Ma X, Chen Y, Cole SR, Chu H. A hybrid Bayesian hierarchical model combining cohort and case-control studies for meta-analysis of diagnostic tests: accounting for partial verification bias. *Statistical Methods in Medical Research* 2014. [Epub ahead of print].
21. Chen Y, Liu Y, Ning J, Cormier J, Chu H. A hybrid model for combining case-control and cohort studies in systematic reviews of diagnostic tests. *Journal of the Royal Statistical Society: Series c: Applied Statistics* 2015; **64**(3):469–489.
22. Ma X, Nie L, Cole SR, Chu H. Statistical methods for multivariate meta-analysis of diagnostic tests: an overview and tutorial. *Statistical Methods in Medical Research* 2013. [Epub ahead of print].
23. Suzuki S, Moro-oka T, Choudhry NK. The conditional relative odds ratio provided less biased results for comparing diagnostic test accuracy in meta-analyses. *Journal of Clinical Epidemiology* 2004; **57**(5):461–469.
24. Siadaty MS, Philbrick JT, Heim SW, Schectman JM. Repeated-measures modeling improved comparison of diagnostic tests in meta-analysis of dependent studies. *Journal of Clinical Epidemiology* 2004; **57**(7):698–711.
25. Chu H, Chen S, Louis TA. Random effects models in a meta-analysis of the accuracy of two diagnostic tests without a gold standard. *Journal of the American Statistical Association* 2009; **104**(486):512–523.
26. Gardner IA, Stryhn H, Lind P, Collins MT. Conditional dependence between tests affects the diagnosis and surveillance of animal diseases. *Preventive Veterinary Medicine* 2000; **45**(1-2):107–122.
27. Liu Y, Chen Y, Chu H. A unification of models for meta-analysis of diagnostic accuracy studies without a gold standard. *Biometrics* 2015; **71**(2):538–547.
28. Trikalinos TA, Hoaglin DC, Small KM, Terrin N, Schmid CH. Methods for the joint meta-analysis of multiple tests. *Research Synthesis Methods* 2014; **5**(4):294–312.
29. Bagos PG. On the covariance of two correlated log-odds ratios. *Statistics in Medicine* 2012; **31**(14):1418–1431.
30. Higgins JP, Whitehead A. Borrowing strength from external trials in a meta-analysis. *Statistics in Medicine* 1996; **15**(24):2733–2749.
31. Mavridis D, Salanti G. A practical introduction to multivariate meta-analysis. *Statistical Methods in Medical Research* 2012. [Epub ahead of print].
32. Jackson D, Riley R, White IR. Multivariate meta-analysis: Potential and promise. *Statistics in Medicine* 2011; **30**(20):2481–2498.

33. van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine* 2002; **21**(4):589–624.
34. Berkey CS, Hoaglin DC, Antczak-Bouckoms A, Mosteller F, Colditz GA. Meta-analysis of multiple outcomes by regression with random effects. *Statistics in Medicine* 1998; **17**(22):2537–2550.
35. Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects. *Statistics in Medicine* 1996; **15**(6):619–629.
36. Deming WaSF. On least square adjustment of sampled frequency tables when the expected marginal totals are known. *The Annals of Mathematical Statistics* 1940; **6**:427–444.
37. Enoe C, Georgiadis MP, Johnson WO. Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. *Preventive Veterinary Medicine* 2000; **45**(1-2):61–81.
38. Hui SL, Zhou XH. Evaluation of diagnostic tests without gold standards. *Statistical Methods in Medical Research* 1998; **7**(4):354–370.
39. Agresti A. Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics. In *Categorical Data Analysis* (2nd edn). John Wiley & Sons: New York, 2002.
40. Walter SD. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. *Statistics in Medicine* 2002; **21**(9):1237–1256.
41. Nishimura K, Sugiyama D, Kogata Y, Tsuji G, Nakazawa T, Kawano S, Saigo K, Morinobu A, Koshiba M, Kuntz KM, Kamae I, Kumagai S. Meta-analysis: diagnostic accuracy of anti-cyclic citrullinated peptide antibody and rheumatoid factor for rheumatoid arthritis. *Annals of Internal Medicine* 2007; **146**(11):797–808.
42. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**:177–188.
43. White IR. Multivariate random-effects meta-analysis. *Stata Journal* 2009; **9**:40–56.
44. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in Medicine* 1999; **18**(20):2693–2708.
45. Riley RD. Multivariate meta-analysis: the effect of ignoring within-study correlation. *Journal of the Royal Statistical Society: Series a (Statistics in Society)* 2009; **172**(4):789–811.
46. Riley RD, Abrams KR, Lambert PC, Sutton AJ, Thompson JR. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. *Statistics in Medicine* 2007; **26**(1):78–97.
47. Riley RD, Abrams KR, Sutton AJ, Lambert PC, Thompson JR. Bivariate random-effects meta-analysis and the estimation of between-study correlation. *BMC Medical Research Methodology* 2007; **7**:3.
48. Clayton D, Hills M. *Statistical Models in Epidemiology*. Oxford University Press, 1993.
49. Henze N, Zirkler B. A class of invariant consistent tests for multivariate normality. *Communications in Statistics - Theory and Methods* 1990; **19**(10):3595–3617.
50. Stevens J. In *Applied Multivariate Statistics for the Social Sciences*, Hillsdale P (ed). : L Erlbaum Assoc., 1986.
51. Thompson B. Multinor: a fortran program that assists in evaluating multivariate normality. *Educ Psychol Measurement* 1990; **50**:845–848.
52. Hamza TH, Reitsma JB, Stijnen T. Meta-analysis of diagnostic studies: a comparison of random intercept, normal-normal, and binomial-normal bivariate summary ROC approaches. *Medical Decision Making* 2008; **28**(5):639–649.
53. Hamza TH, van Houwelingen HC, Stijnen T. The binomial distribution of meta-analysis was preferred to model within-study variability. *Journal of Clinical Epidemiology* 2008; **61**(1):41–51.
54. Agresti A. On logit confidence intervals for the odds ratio with small samples. *Biometrics* 1999; **55**(2):597–602.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.