Check for updates

Tutorial

A Tutorial for Meta-Analysis of Diagnostic Tests for Low-Prevalence Diseases: Bayesian Models and Software

Johny J. Pambabay-Calero [a], Sergio A. Bauz-Olvera [a], Ana B. Nieto-Librero [b], Maria Purificación Galindo-Villardón [b], Ana B. Sánchez-García [c]

[a] *Faculty of Natural Sciences and Mathematics, ESPOL, Polytechnic University, Guayaquil, Ecuador.* [b] *Department of Statistics, University of Salamanca, Salamanca, Spain.* [c] *INICO, Faculty of Education, University of Salamanca, Salamanca, Spain.*

Corresponding Author: Johny J. Pambabay-Calero, Facultad de Ciencias Naturales y Matemáticas, Escuela Superior Politécnica del Litoral, Km. 30.5 Vía Perimetral, Guayaquil, Ecuador 09-01-5863. E-mail: jpambaba@espol.edu.ec

Abstract

Although measures such as sensitivity and specificity are used in the study of diagnostic test accuracy, these are not appropriate for integrating heterogeneous studies. Therefore, it is essential to assess in detail all related aspects prior to integrating a set of studies so that the correct model can then be selected. This work describes the scheme employed for making decisions regarding the use of the R, STATA and SAS statistical programs. We used the R Program Meta-Analysis of Diagnostic Accuracy package for determining the correlation between sensitivity and specificity. This package considers fixed, random and mixed effects models and provides excellent summaries and assesses heterogeneity. For selecting various cutoff points in the meta-analysis, we used the STATA module for meta-analytical integration of diagnostic test accuracy studies, which produces bivariate outputs for heterogeneity.

Keywords

bivariate models, heterogeneity, meta-analysis, statistical software

Diagnostic accuracy plays a central role in the evaluation of diagnostic tests, where accuracy can be expressed as sensitivity, specificity, positive predictive value, negative predictive value, and reasons of probability. However, predictive values depend directly on the prevalence of the disease in question and, therefore, cannot be directly compared in different situations. By contrast, it is believed that test sensitivity and specificity do not vary with the prevalence of disease.

This is also the case for reasons of probability. Since they depend on sensitivity and specificity they are believed to remain constant, although variability with regard to prevalence does exist. However, some studies (Brenner & Gefeller, 1997; Leeflang et al., 2013; Ransohoff & Feinstein, 1978) have shown contradictory findings.

Several studies have indicated that the variability of sensitivity and specificity may be related to differences in thresholds (Holling, Böhning, & Böhning, 2007; Jiang, 2018; Mulherin & Miller, 2002; Szklo & Nieto, 2014). One study has reported differences in cutoff points or in the definition of the disease (Brenner & Gefeller, 1997). Therefore, it is necessary to analyze to what extent, in what form and why the sensitivity and specificity of diagnostic tests vary with respect to prevalence before performing a meta-analysis (Holling et al., 2007). However, these factors are quite difficult to identify and often warrant the use of models that consider this fact when generating summary estimates of sensitivity and specificity. The bivariate model of random effects captures the correlation between sensitivity and specificity and models the logits of both factors (Reitsma et al., 2005).

In situations of low prevalence, where the test being employed provides a high number of true negatives and a small number of true positives, the percentage of cases correctly classified does allow different tests to be compared. This is because true positives will be very high, even when the number of false positives is equal to or greater than the number of true positives, which is a situation that can cause the test to be rejected and declared as being inefficient.

Meta-Analysis of Diagnostic Accuracy (MADA) libraries are among the statistical packages available that can be used with the most relevant models (Doebler, 2020). These include the Hierarchical Summary Receiver Operating Characteristic (HSROC; Schiller & Dendukuri, 2013) by R, Meta-analytical Integration of Diagnostic Test Accuracy Studies (MIDAS; Dwamena, 2007; Wang & Leeflang, 2019) by STATA, and the Macro for Meta-analysis of Diagnostic accuracy studies (MetaDas; Takwoingi & Deeks, 2010; Wang & Leeflang, 2019) by SAS.

This paper describes the main models used in this context, as well as the available software. Since it is not always easy for researchers to decide on the model most appropriate for their study or to choose the correct software for interpreting their results, we have created a guide for carrying out a meta-analysis on diagnostic tests. According to the assumptions that fulfill the analyzed data, we present findings regarding the most suitable model, the software that allows this model to be used and how the results obtained can be interpreted.

Evaluation of Heterogeneity

To investigate the effect of a cutoff point on sensitivity and specificity, the results have been presented in the form of a receiver operating characteristic (ROC) curve. In addition, one way to summarize the behavior of a diagnostic test from multiple studies

PsychOpen GOLD

is to calculate the mean sensitivity and specificity (Bauz-Olvera et al., 2018); however, these measures are invalid if heterogeneity exists (Midgette, Stukel, & Littenberg, 1993). The sensitivity and specificity within each study are inversely related and depend on the cutoff point, which implies that sensitivity and medium specificity are not acceptable (Irwig, Macaskill, Glasziou, & Fahey, 1995). On the other hand, the Biplot has proved to be an extremely useful multivariate tool in the analysis of data from the meta-analysis of diagnostic tests, both in the descriptive phase and in the search for the causes of variability (Pambabay-Calero et al., 2018).

In diagnostic tests, the assumption of methodological homogeneity in studies is not met and thus it becomes important to evaluate heterogeneity. Assessing the possible presence of statistical heterogeneity in the results can be done (in a classical way) by presenting the sensitivity and specificity of each study in a forest plot.

A characteristic source of heterogeneity is that which arises because the studies included in the analysis may have considered different thresholds for defining positive results; this effect is known as the threshold effect.

The most robust statistical methods proposed for meta-analysis take this threshold effect into account and do so by estimating a summary ROC curve (SROC) of the studies being analyzed. However, on some occasions the results of the primary studies are homogeneous and the presence of both threshold effect and other sources of heterogeneity can be ruled out. This statistical modelling can be done using either a fixed-effect model or a random-effects model, depending on the magnitude of heterogeneity. Several statistical methods for estimating the SROC curve have been proposed. The first, proposed by (Moses, Shapiro, & Littenberg, 1993), is based on estimating a linear regression between two variables created from the validity indices of each study.

Accuracy Analysis in Diagnostic Tests

The discriminatory capacity of a test is commonly expressed in terms of two measures (sensitivity and specificity) and there is usually an inverse relationship between the two due to the variability of the thresholds. Some of the recommended methods for meta-analysis of diagnostic tests, such as the bivariate model, focus on estimating a summary sensitivity and specificity at a common threshold. The HSROC model, on the other hand, focuses on estimating a summary curve from studies that have used different thresholds.

The test can be based on a biomarker or a more complex diagnostic procedure. However, the value of the index that the test provides may not be completely reliable. The starting information is a 2×2 table showing the concordance between the test results in binary form and information associated with disease (see Table 1; Deeks, Macaskill, & Irwig, 2005).

Table 1

Generic Representation of the Data of a Study of Accuracy in Diagnostic Test

| Test result | Disease state | | Total |
	D+	D−	
T+	TP	FP	TP + FP
T−	FN	TN	FN + TN
Total	n_1	n_2	n

Note. n = sample size; n_1 = patients who actually have the disease; n_2 = patients who are disease free. T+ = a positive result; T− = a negative result; TP = true positives; FP = false positives; TN = true negatives; FN = false negatives (Deeks et al., 2005).

The results of a meta-analysis of diagnostic tests are usually reported as a pair, representing both sensitivity and specificity. However, some attempts have been made to consolidate the result as a single number. The most common approach is the use of diagnostic odds ratio (DOR; Lee, Kim, Choi, Huh, & Park, 2015). But other important measures are positive LR+ or negative likelihood LR− ratios, which are estimated from sensitivity and specificity. Summary graphs can be generated that show the variability among the studies based on sensitivity and specificity. Thus, we have (see Figure 1 for a more detailed explanation):

1. **Forest plot** for sensitivity and specificity help in assessing the heterogeneity of individual aspects of test accuracy, but do not allow immediate assessment of whether the observed variation is that expected from the relationship of two variables, i.e. sensitivity decreases as specificity increases (Brenner & Gefeller, 1997).
2. **Crosshair**, shows the bivariate relationship and the degree of heterogeneity between sensitivity and the rate of false positives. These "cross-hair" graphs reflect the results of individual studies in the ROC space with confidence intervals denoting sensitivity and specificity. They also allow meta-analysis studies to be overlaid on the graph.
3. **RocEllipse**, shows a region of confidence that describes the uncertainty of the pair of sensitivity and specificity of each study. The clinical utility or relevance of the diagnostic test carried out on a patient is evaluated using maximum likelihood ratios to calculate post-test probability. On the basis of Bayes' theorem, this concept is shown in Fagan's nomogram (Fagan, 1975). The nomogram is a tool that allows estimating the post-test probability once the prevalence of the disease and the likelihood ratio are known. This graph has three columns with numbers: the first corresponds to the pre-test probability, the second to the likelihood ratio (LR) and the third to the post-test. This post-test allows likelihood to be quantified after testing if an individual will be affected by a specific condition. Also, the result of an

PsychOpen GOLD

observed test and the likelihood that the individual will have the condition before the test is performed is taken into account.

Figure 1

Summary Graphs That Show the Variability Among the Studies

(a) Sensitivity

(b) Specificity

(c) Crosshair

(d) RocEllipse
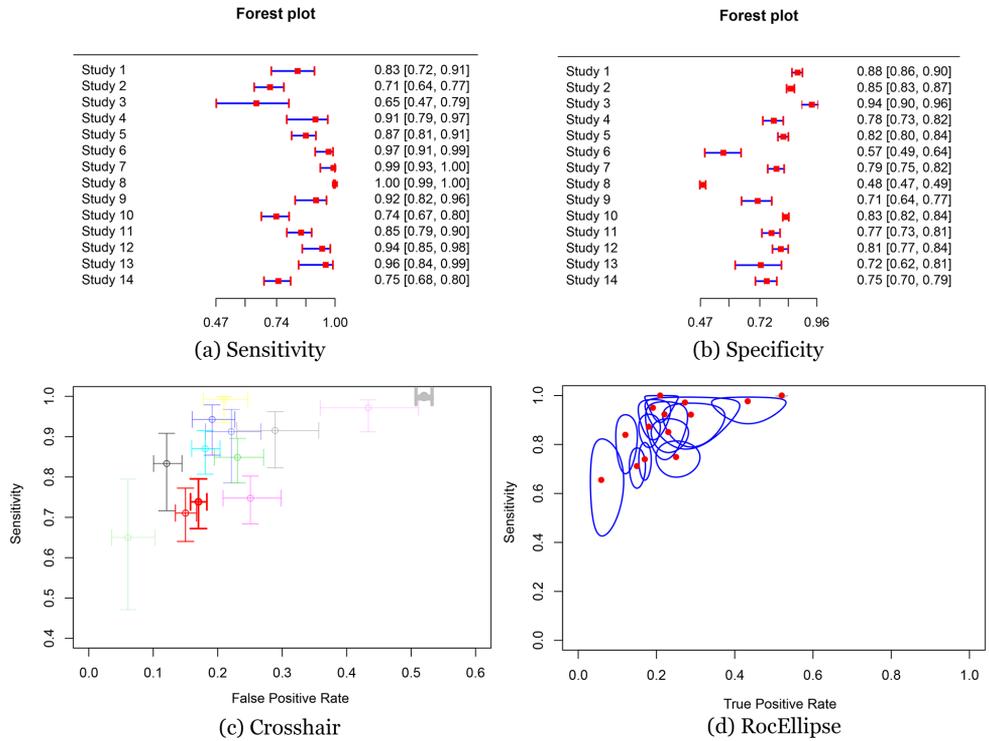
Tests Based on a Continuous Marker

Let X be the continuous diagnostic marker that underlies a test, which must take into account two different probability distributions for X between diseased and healthy individuals respectively. It is assumed that the diagnostic marker tends to be higher in diseased individuals than in healthy individuals. A graphical representation is shown in Figure 2.

Figure 2

The Distributions of a Continuous Biomarker of Diseased and Healthy Individuals With a Specific Cut-Off Point
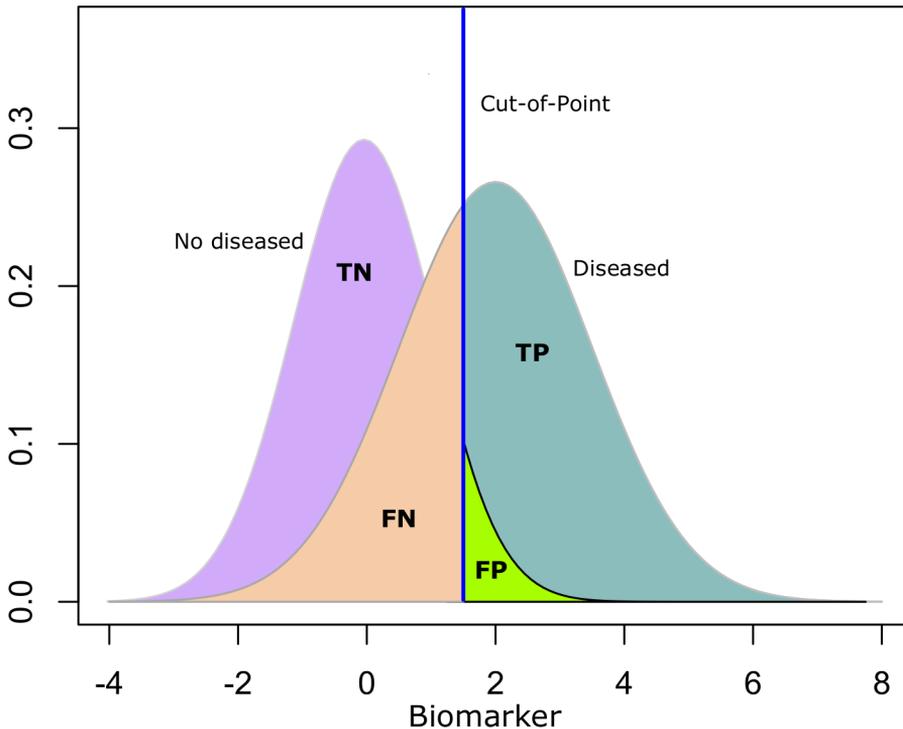

Note. TP = true positives; FP = false positives; TN = true negatives; FN = false negatives.

A consequence of the overlapping of the distributions of Figure 2 is that the cutoff point may not have been defined correctly. If, for example, the cutoff point moves to the left, TP and FP will increase, whereas both TN and FN decrease. The variation in the sensitivity–specificity pair and the cutoff point is shown in Figure 3.

Figure 3

SROC Curve Showing the Impact of the Heterogeneity on the Performance of the Test
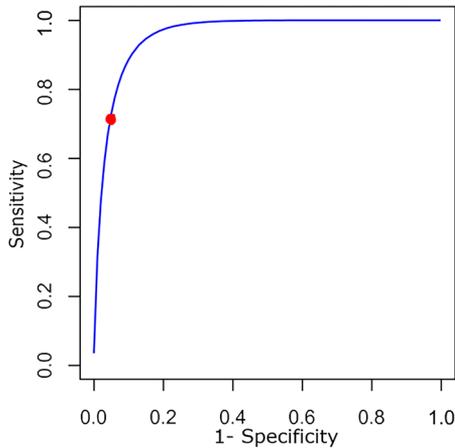
Model of Moses (SROC Model)

The objective of this model is to transform true positive rate (TPR) and false positive rate (FPR) so that the relationship becomes linear; thus making an adjustment for the points given (Moses et al., 1993). It is based on estimating a linear regression between two variables created from the validity indices of each study. These variables are *D* and *S*, respectively, and represent the DOR. The model is adjusted using either weighted or unweighted least squares.

Various useful statistical methods have been proposed to summarize a SROC curve. The most common is the area under the curve (AUC), which summarizes the diagnostic performance of the test in a single number (Walter, 2002): perfect tests have an AUC close to 1 whereas unusable tests have an AUC close to 0 (de Llano et al., 2007).

The Moses model does present some limitations. On one hand, it does not take into account the different levels of precision with which sensitivity and specificity are estimated in each study, nor does it incorporate heterogeneity between studies. To overcome these limitations, more complex regression models have been proposed. The first of these is a bivariate random effects model (Reitsma et al., 2005) that assumes that the logit of sensitivity and specificity follow a bivariate normal distribution. The model contemplates the possible correlation between both indices, models the different precision with which sensitivity and specificity have been estimated and incorporates a source of additional heterogeneity due to variance between studies. The second model refers to the HSROC approach or hierarchical model (Rutter & Gatsonis, 2001). It is similar to the previous model, except that it clearly defines the relationship between sensitivity and specificity across the threshold (Doi & Williams, 2013).

Bivariate Model

It should be noted that the SROC model does not quantify the error in *S* (Baker, Kim, & Kim, 2004). An alternative approach for the construction of a SROC curve has been described by Reitsma et al. (2005). This author proposes the use of a bivariate model through a joint distribution of sensitivity and specificity, which allows the linear correlation throughout the studies to be modelled. This model follows an approach developed for meta-analysis of binary results (Van Houwelingen, Zwinderman, & Stijnen, 1993), the same that has been improved by other authors (Arends et al., 2008). At the study level, this model assumes that the TP and FP within the study k, k = 1, 2, ..., K follow binomial distributions. For the levels among the studies, a bivariate random effect model is assumed, $logit(Se_k)$ and $logit(1-Sp_k)$, in which normal distributions of the specific parameters of the study are assumed a priori, Equation 1:

$$\begin{pmatrix} logit(Se_k) \\ logit(1 - Sp_k) \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} ; \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \right) \tag{1}$$

Alternatively, the covariance can be parameterized by the correlation coefficient ρ and standard errors in such a way that $\sigma_{12} = \rho\sigma_1\sigma_2$. This model has five parameters:

- Means, μ_1, μ_2
- Variances σ_1^2, σ_2^2
- Covariance σ_{12}

The inclusion of covariates in the sensitivity or specificity, or both, is done by replacing one or both means μ_1 and μ_2 by linear variables in the covariates. For example, if there is only one covariate that affects sensitivity and specificity, we could substitute μ_1 by $\mu_1 + \nu_1 Z_i$ and μ_2 by $\mu_2 + \nu_2 Z_i$ (Harbord, Deeks, Egger, Whiting, & Sterne, 2007; Takwoingi, Guo, Riley, & Deeks, 2017).

Hierarchical Model HSROC

Rutter and Gatsonis (2001) were the first to develop a model that quantifies the size of heterogeneity. These authors proposed a hierarchical model with a Bayesian empirical version added by Macaskill (2004). The model includes random effects for cutoff points and the accuracy of the test and focuses on estimating the ROC curve (Schwarzer, Carpenter, & Rücker, 2015). The objective is to obtain significant estimates of sensitivity and specificity and better manages variability, applying a Bayesian approach for the estimation of the parameters. Rutter and Gatsonis (2001) divided the model into three levels. At the study level, it is assumed that within each study k, k = 1, 2, ..., K; the TP and the FP follow binomial distributions (Schwarzer et al., 2015), where Index 1 indicates the sick people and Index **2** denotes those without the disease, Equation 2.

PsychOpen^GOLD

$$TP \sim Binomial(n_{k1}, Se_k)$$
$$FP \sim Binomial(n_{k2}, (1 - Sp_k))$$

(2)

The authors parameterized the sensitivities and specificities as follows (Schwarzer et al., 2015), Equation 3,

$$logit(Se_k) = \left(\theta_k + \frac{\alpha_k}{2}\right)e^{-\beta/2} logit(1 - Sp_k) = \left(\theta_k - \frac{\alpha_k}{2}\right)e^{\beta/2}$$

(3)

where θ_k is the random threshold in the study k, α_k is the random accuracy in the study k, and β is a parameter of the shape (asymmetry) of the ROC curve (Schwarzer et al., 2015). Normal distributions are used to model variation in the specific parameters of the study among the studies, Equation 4, which corresponds to the second level of modelling, i.e. variation between studies.

$$\theta_k \sim N(\theta, \tau_\theta^2)$$
$$\alpha_k \sim N(\lambda, \tau_\lambda^2)$$

(4)

Finally, the specification of the hierarchical model is completed by choosing a priori the distributions of the parameters. In short, the model has five parameters (Schwarzer et al., 2015):

- the mean and variance of the cutoff points (θ, τ_θ^2);
- the mean and variance of the accuracy $(\lambda, \tau_\lambda^2)$; and
- the shape parameter β.

A value of $\beta = 0$ would represent a symmetric curve in the ROC space (Schwarzer et al., 2015). The ROC curve is calculated by applying the inverse function *logit* to a function that is linear in $logit(1 - Sp_k)$, Equation 5 (Schwarzer et al., 2015).

$$Se = logit^{-1}\{e^{-\beta}logit(1 - Sp) + \lambda e^{-\beta/2}\}$$

(5)

The above expression is equivalent to Equation 6.

$$Se = \left\{1 + exp\left(-e^{-\beta}logit\left(\frac{1 - Sp}{Sp}\right) - \lambda e^{-\beta/2}\right)\right\}^{-1}$$

(6)

Further details can be found in some related papers (Macaskill, 2004; Rutter & Gatsonis, 2001). To understand the operation of the analyzed models, it is necessary to use a statistical program. In our case, we will analyze R, STATA, and SAS to facilitate the necessary analysis.

In more generally, the mean sensitivity and specificity can be modeled through linear regressions of study-level covariates (Harbord et al., 2007). This could be achieved, for

example, by using a single covariate Z that affects both the cutoff points and accuracy parameters such as Equation 7,

$$\theta_k \sim N\left(\theta + \gamma Z, \tau_\theta^2\right)$$
$$\alpha_k \sim N\left(\lambda + \nu Z, \tau_\lambda^2\right)$$

(7)

where the coefficients γ and ν quantify the weight of the covariate Z on the cutpoint and precision respectively. This model allows to include more than one covariate, also, allows to model the covariates independently in the parameters of accuracy and cutoff points (Harbord et al., 2007).

Software to Integrate the Meta-Analysis of Diagnostic Tests

R Language

The MADA package of the statistical program R is a tool that allows the meta-analysis of diagnostic tests to be accurately carried out. Although there are many methods for diagnostic meta-analysis, it is still not a routine procedure. One of the reasons may be due to the complexity of the bivariate approach. The MADA statistical package offers some current approaches to diagnostic meta-analysis, as well as functions that allow for statistical methods for a data set include sensitivity, specificity, true/false positives, true/false negatives, and their DOR (Glas et al., 2003). These statistical methods can be employed using the madad function of the MADA library and the mslSROC function of the META library. Prior to the advent of the bivariate approach, some univariate approaches were very popular. This approach is characterized by the separate estimation of sensitivity and specificity. There are three methods in R for this approach:

- the Mantel–Haenszel (MH) method, for a fixed effect model (Deeks, 2001);
- the model is formulated in terms of DOR logarithms and is a weighted estimator;
- the proportional model of Hazards (Holling, Böhning, & Böhning, 2012), which is constructed on the assumption of a simple ROC curve and assumes the conditions of the Lehmann model.

In meta analysis of diagnostic tests, the relationship between sensitivity and specificity is negative. Since these quantities are related to each other, the bivariate approach for meta-analysis in the accuracy of the diagnosis has been welcomed. Using the Reitsma function in the MADA library, it is possible to use the aforementioned model. Finally, the HSROC library that contains the HSROC function is used to estimate the HSROC hierarchical model, which makes the necessary adjustments in the model.

STATA

The MIDAS package is a comprehensive program of statistical and graphical routines used to understand the meta-analysis of diagnostic tests in STATA, which is a statistical software package that was created by StataCorp in 1985. It provides statistical and graphical functions that allow us to study the accuracy of diagnostic tests. The modeling of primary data is done through a binary regression of bivariate mixed effects. Model fitting, estimation, and prediction are performed by adaptive quadrature. Using the values of the coefficients and the variance-covariance matrices, the sensitivity and specificity are estimated with their respective zones of confidence and prediction in the ROC space (Dwamena, 2007).

SAS

MetaDas is a high-performance SAS program, which adjusts the parameters of bivariate and HSROC models to analyze the accuracy of diagnostic tests using Proc nonlinear mixed models (NLMIXED; Takwoingi & Deeks, 2010). NLMIXED adjusts the parameters of the models using likelihood functions through optimization algorithms, the main ones being adaptive Gaussian quadrature and a first-order Taylor series approach (Takwoingi & Deeks, 2010).

Steps for Performing a Meta-Analysis of Diagnostic Tests in Low Prevalence

Once the systematic review of the diagnostic tests has been performed, it is necessary to integrate the results using the approaches described above. For this reason, we propose the following four steps:

1. Perform a descriptive statistical analysis of the studies using the R language and the MADA and META libraries together with the madad and mslSORC functions, respectively, which provide the following results and graphs.
   - Sensitivity per study with their respective confidence intervals (IC)
   - Specificity per study, IC
   - DOR per study, IC
   - Chi-square test that allows comparing the sensitivity and specificity of the studies
   - LR+ and LR− *per* study with their respective confidence intervals
   - Correlation between sensitivity (Se) and specificity (Sp)
   - Rate of false-positive (RFP) per study, IC
   - Forest plot for sensitivity and specificity
   - Crosshair and RocEllipse chart
   - SROC Curve of Moses model

2. If there is independence between sensitivity and specificity, a univariate analysis is then performed using the madauni and phm functions of the MADA library of the R language. This analysis uses the Mantel–Haenszel (fixed effects), DerSimonian-Laird (random effects) models and the Hazards proportional approach (fixed and random effects), which generate the following results.
   - DOR and DOR logarithm with their respective confidence interval
   - Forest plot for sensitivity and specificity with their respective confidence intervals
   - τ^2 with confidence interval
   - Q test, I^2
   - AUC
   - Forest plot with summary measures for DOR, LR+ and LR– log
   - Chi-square test of homogeneity between studies
   - Chi-square test of heterogeneity between studies
   - Curve SROC with RocEllipse
3. If the sensitivity and specificity are related, i.e., there are different cutoff points in the meta-analysis and the data is adjusted to a normal bivariate distribution, a bivariate analysis is performed using the R and STATA languages using the MADA and MIDAS libraries. Note that for using the bivariate approach in R, the reitsma function is used. This bivariate analysis generates the following results, see Table 2.

Table 2

Main Outputs of the Bivariate Model

R Language	Stata Language
Logit of consensus sensitivity with confidence interval	Forest plot for sensitivity with and with- out measure summary and their confi- dence intervals
Logit of false-positive rate with confidence interval	Forest plot for specificity with and with- out measure summary and their confi- dence intervals
Sensitivity consensus with confidence interval	DOR, LR+, and LR– consensus with their respective confidence intervals
False-positive consensus rate with confidence intervals	Q test, I^2
SROC curve with sensitivity and false- positive consensus rate	AUC
Matrix of variances between studies	Sensitivity and specificity, consensus with their respective
Correlation matrix	SROC curve with sensitivity, specificity consensus and confidence intervals
HSROC model parameters	Fagan plot

4. If the effect of the characteristics or the study on the threshold, accuracy, and shape of the SROC curve must be determined, a hierarchical approach HSROC should be used. The data must conform to this hierarchical approach using the HSROC and MetaDas packages of the R and SAS languages, respectively, which generate the following main outputs, see Table 3.

Table 3

Main Outputs of the HSROC Model

R Language	SAS Language
A priori values of the model parameters	Information on covariates
A posteriori values of the model parameters	Initial values of the model and state of convergence and adjustment of the model
Sensitivity and specificity by studies with their respective confidence intervals	Sensitivity, specificity, DOR, LR+, LR− consensus
Sensitivity and specificity, consensus with their respective confidence intervals	Confidence intervals and prediction of model parameters
SROC curve with sensitivity and specificity consensus and its confidence intervals	Predictive values of sensitivity and specificity for studies, histogram and normal probability graphs of Bayesian empirical estimates of random effects

A graphic representation of the above is detailed in Figure 4. The above mentioned statistical and graphical measurements can be obtained using the algorithms available in Supplementary Materials.

Figure 4

Steps to be Followed to Carry out a Meta-Analysis in Diagnostic Tests for Low-Prevalence Diseases

Discussion

The Moses model uses true and false positive rate logit functions to build a linear regression model where the response variable (test accuracy) is explained by the proportion of positive test results (relative to the threshold). The SROC curve is symmetrical if the statistical relationship between precision and threshold is zero, i.e. constant DOR. This modeling is characterized by a fixed effect since the variation is attributed to the threshold and the sampling error. This model generates errors, which makes the statistical inference invalid (Arends et al., 2008; Chu, Guo, & Zhou, 2010; Ma et al., 2016; Macaskill, 2004; Verde, 2010).

Hierarchical models capture the stochastic relationship between sensitivity, specificity, and variability of test accuracy in all studies by incorporating random effects into the modeling. Bivariate and HSROC models differ in their parameterization but are mathematically equivalent when covariates are not included (Harbord et al., 2007). The choice of model depends on the variation in reported thresholds in the studies, and the inference is given by a summary point or an SROC curve (Takwoingi et al., 2017).

The bivariate model models random effects to estimate sensitivity and specificity, as well as to construct 95% credibility intervals. The model is based on logit transformations of sensitivity and specificity as bivariate normal distributions. The estimation of the correlation parameter is achieved from the subsequent means of sensitivity and specificity (Launois, Le Moine, Uzzan, Navarrete, & Benamouzig, 2014). Random effects also follow a bivariate normal distribution. If the model is simplified by assuming that the covariance or correlation is zero, the model is reduced to two univariate random effects regression models for sensitivity and specificity (Bauz-Olvera et al., 2018).

The HSROC model is a reference in the study of diagnostic test accuracy and can be seen as a generalization of the Moses SROC approach, in which TPR and FPR are modeled directly. (Macaskill, 2004; Takwoingi et al., 2017).

The HSROC model and the bivariate model are different settings of the same underlying model, and both approaches can be used to calculate estimates of the SROC curve and random effects. Moreover, there is a difference in the software packages that can fit them. While the HSROC model requires a non-linear mixed model program like NLMIXED in SAS, the bivariant only requires a linear mixed model program and can be installed in R and Stata.

Since the bivariate model is parameterized in terms of sensitivity and mean specificity (logit), it is often claimed that this is the preferred model for estimating the mean operating points. However, in practice, it is possible to obtain estimates of both the average operating point and the summary ROC curve from both HSROC modes. Therefore, the estimation of average operating points depends on the homogeneity of the thresholds included in the analysis, not on the choice of the statistical model. The bivariate model allows covariates to be included in sensitivity and/or specificity, while the HSROC model

facilitates the inclusion of covariates that affect threshold and/or accuracy (Takwoingi & Deeks, 2010).

We suggest that meta-analysts carefully explore and inspect their data using a forest plot and an SROC curve before performing meta-analyses. These first analyses will quantify stochastic heterogeneity and the dispersion of study points in the ROC space (Lee, Kim, Choi, Huh, & Park, 2015). This visualization should provide information on the approach to be taken at the time of model selection. Although the Bayesian approach is complex in its parameterization it is not commonly used, but, it represents an alternative to the maximum likelihood approach. In an empirical evaluation, both approaches were found to be similar, although Bayesian methods suggest greater uncertainty around point estimates (Dahabreh, Trikalinos, Lau, & Schmid, 2012; Harbord et al., 2008).

The hierarchical approach can be used in different situations such as (1) the presence or absence of heterogeneity and (2) cutoff points being homogeneous among studies. This is the reason we recommend using this model in situations of low prevalence, because it better handles the variability between and within studies. Thus, this model is an approach suitable for fixed and random effects depending on the nature of the data.

The bivariate model allows covariates to be included in sensitivity and/or specificity, while the HSROC model facilitates the inclusion of covariates that affect threshold and/or accuracy (Takwoingi & Deeks, 2010).

The selection of the statistical model in the meta-analysis of diagnostic tests of low-prevalence diseases is essential for the integration of the study results. Regardless of the software used, the rigorous application of the decision-making scheme will help to guarantee high quality results and facilitate the analysis and interpretation of the results.

Supplementary Materials

For this article the following supplementary materials are available (Pambabay-Calero, Bauz-Olvera, Nieto-Librero, Galindo-Villardón, & Sánchez-García, 2020):

- **Scripts in R**: script_1.R (foresplot, crosshair and RocEllipse); script_2.R (models DSL, MH and PHM); script_3.R (model Reitsma); script_4.R (model HSROC)
- **Script in Stata**: script_stata.do
- **Scripts in SAS**: ejecHSROC.sas; metadas.sas; data_set.csv

PsychOpen GOLD

Index of Supplementary Materials

Pambabay-Calero, J. J., Bauz-Olvera, S. A., Nieto-Librero, A. B., Galindo-Villardón, P., & Sánchez-García, A. B. (2020). *Supplementary materials to: A tutorial for meta-analysis of diagnostic tests for low-prevalence diseases: Bayesian models and software* [Scripts]. PsychOpen. https://doi.org/10.23668/psycharchives.3485

References

Arends, L. R., Hamza, T. H., Van Houwelingen, J. C., Heijenbrok-Kal, M. H., Hunink, M. G., & Stijnen, T. (2008). Bivariate random effects meta-analysis of ROC curves. *Medical Decision Making, 28*(5), 621-638. https://doi.org/10.1177/0272989X08319957

Baker, F. B., Kim, S.-H., & Kim, S.-H. (2004). *Item response theory* (2nd ed.). Boca Raton; FL, USA: CRC Press. https://doi.org/https://doi.org/10.1201/9781482276725

Bauz-Olvera, S. A., Pambabay-Calero, J. J., Nieto-Librero, A. B., & Galindo-Villardón, M. P. (2018). Meta-analysis in DTA with hierarchical models bivariate and HSROC: Simulation study. In I. Antoniano-Villalobos, R. Mena, M. Mendoza, L. Naranjo, & L. Nieto-Barajas (Eds.), *Selected contributions on statistics and data science in Latin America* (pp. 33-42). https://doi.org/ https://doi.org/10.1007/978-3-030-31551-1_3

Brenner, H., & Gefeller, O. (1997). Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. *Statistics in Medicine, 16*(9), 981-991. https://doi.org/10.1002/(SICI)1097-0258(19970515)16:9<981::AID-SIM510>3.0.CO;2-N

Chu, H., Guo, H., & Zhou, Y. (2010). Bivariate random effects meta-analysis of diagnostic studies using generalized linear mixed models. *Medical Decision Making, 30*(4), 499-508. https://doi.org/10.1177/0272989X09353452

Dahabreh, I. J., Trikalinos, T. A., Lau, J., & Schmid, C. (2012). *An empirical assessment of bivariate methods for meta-analysis of test accuracy* (Report No. 12[13]-EHC136-EF). Agency for Healthcare Research and Quality (US). Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23326899

Deeks, J. J. (2001). Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests. *BMJ, 323*(7305), 157-162. https://doi.org/10.1136/bmj.323.7305.157

Deeks, J. J., Macaskill, P., & Irwig, L. (2005). The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *Journal of Clinical Epidemiology, 58*(9), 882-893. https://doi.org/10.1016/j.jclinepi.2005.01.016

de Llano, S. R., Delgado-Bolton, R., Jiménez-Vicioso, A., Pérez-Castejón, M., Delgado, J. C., Ramos, E., . . . Pérez-Vázquez, J. M. (2007). Meta-analysis of the diagnostic performance of 18F-FDG PET in renal cell carcinoma. *Revista Espanola de Medicina Nuclear, 26*(1), 19-29. https://doi.org/10.1016/S1578-200X(07)70049-9

Doebler, P. (2020). *Meta-analysis of diagnostic accuracy* (Version 0.5.10). Retrieved from https://repo.bppt.go.id/cran/web/packages/mada/mada.pdf

Doi, S. A. R., & Williams, G. M. (Eds.). (2013). *Methods of clinical epidemiology*. Heidelberg, Germany: Springer. https://doi.org/10.1007/978-3-642-37131-8

Dwamena, B. (2007). *MIDAS: Stata module for meta-analytical integration of diagnostic test accuracy studies*. Statistical Software Components, Boston College Department of Economics. Retrieved from https://ideas.repec.org/c/boc/bocode/s456880.html

Fagan, T. J. (1975). Nomogram for Bayes's theorem. *The New England Journal of Medicine, 293*(5), 257. https://doi.org/10.1056/NEJM197507312930513

Glas, A. S., Lijmer, J. G., Prins, M. H., Bonsel, G. J., & Bossuyt, P. M. (2003). The diagnostic odds ratio: A single indicator of test performance. *Journal of Clinical Epidemiology, 56*(11), 1129-1135. https://doi.org/10.1016/S0895-4356(03)00177-X

Harbord, R. M., Deeks, J. J., Egger, M., Whiting, P., & Sterne, J. A. C. (2007). A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics, 8*(2), 239-251. https://doi.org/10.1093/biostatistics/kxl004

Harbord, R. M., Whiting, P., Sterne, J. A., Egger, M., Deeks, J. J., Shang, A., & Bachmann, L. M. (2008). An empirical comparison of methods for meta-analysis of diagnostic accuracy showed hierarchical models are necessary. *Journal of Clinical Epidemiology, 61*(11), 1095-1103. https://doi.org/10.1016/j.jclinepi.2007.09.013

Holling, H., Böhning, D., & Böhning, W. (2007). Meta-analysis of binary data based upon dichotomized criteria. *Zeitschrift für Psychologie. The Journal of Psychology, 215*(2), 122-131.

Holling, H., Böhning, W., & Böhning, D. (2012). Meta-analysis of diagnostic studies based upon SROC-curves: A mixed model approach using the Lehmann family. *Statistical Modelling, 12*(4), 347-375. https://doi.org/10.1177/1471082X1201200403

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

Jiang, Z. (2018). Using the linear mixed-effect model framework to estimate generalizability variance components in R. *Methodology, 14*(3), 133-142. https://doi.org/10.1027/1614-2241/a000149

Launois, R., Le Moine, J.-G., Uzzan, B., Navarrete, L. I. F., & Benamouzig, R. (2014). Systematic review and bivariate/HSROC random-effect meta-analysis of immunochemical and guaiac-based fecal occult blood tests for colorectal cancer screening. *European Journal of Gastroenterology & Hepatology, 26*(9), 978-989. https://doi.org/10.1097/MEG.0000000000000160

Lee, J., Kim, K. W., Choi, S. H., Huh, J., & Park, S. H. (2015). Systematic review and meta-analysis of studies evaluating diagnostic test accuracy: A practical review for clinical researchers-part II. Statistical methods of meta-analysis. *Korean Journal of Radiology, 16*(6), 1188-1196. https://doi.org/10.3348/kjr.2015.16.6.1188

Leeflang, M. M., Rutjes, A. W., Reitsma, J. B., Hooft, L., & Bossuyt, P. M. (2013). Variation of a test's sensitivity and specificity with disease prevalence. *Canadian Medical Association Journal, 185*(11), E537-E544. https://doi.org/10.1503/cmaj.121286

PsychOpen GOLD

Ma, X., Nie, L., Cole, S. R., Chu, H., Lawson, A. B., Lee, D., & MacNab, Y. (2016). Statistical methods for multivariate meta-analysis of diagnostic tests: An overview and tutorial. *Statistical Methods in Medical Research, 25*(4), 1596-1619. https://doi.org/10.1177/0962280213492588

Macaskill, P. (2004). Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. *Journal of Clinical Epidemiology, 57*(9), 925-932. https://doi.org/10.1016/j.jclinepi.2003.12.019

Midgette, A. S., Stukel, T. A., & Littenberg, B. (1993). A meta-analytic method for summarizing diagnostic test performances: Receiver-operating-characteristic - summary point estimates. *Medical Decision Making, 13*(3), 253-257. https://doi.org/10.1177/0272989X9301300313

Moses, L. E., Shapiro, D., & Littenberg, B. (1993). Combining independent studies of a diagnostic test into a summary roc curve: Data-analytic approaches and some additional considerations. *Statistics in Medicine, 12*(14), 1293-1316. https://doi.org/10.1002/sim.4780121403

Mulherin, S. A., & Miller, W. C. (2002). Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. *Annals of Internal Medicine, 137*(7), 598-602. https://doi.org/10.7326/0003-4819-137-7-200210010-00011

Pambabay-Calero, J. J., Bauz-Olvera, S. A., Nieto-Librero, A. B., Galindo-Villardon, M. P., & Hernandez-Gonzalez, S. (2018). An alternative to the cochran-(Q) statistic for analysis of heterogeneity in meta-analysis of diagnostic tests based on HJ biplot. *Investigación Operacional, 39*(4), 536-545.

Ransohoff, D. F., & Feinstein, A. R. (1978). Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *The New England Journal of Medicine, 299*(17), 926-930. https://doi.org/10.1056/NEJM197810262991705

Reitsma, J. B., Glas, A. S., Rutjes, A. W., Scholten, R. J., Bossuyt, P. M., & Zwinderman, A. H. (2005). Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology, 58*(10), 982-990. https://doi.org/10.1016/j.jclinepi.2005.02.022

Rutter, C. M., & Gatsonis, C. A. (2001). A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in Medicine, 20*(19), 2865-2884. https://doi.org/10.1002/sim.942

Schiller, I., & Dendukuri, N. (2013). *HSROC: An R package for Bayesian meta-analysis of diagnostic test accuracy*. Retrieved from https://core.ac.uk/download/pdf/23797204.pdf

Schwarzer, G., Carpenter, J. R., & Rücker, G. (2015). *Meta-analysis with R* (Vol. 4784). Cham, Switzerland: Springer.

Szklo, M. M., & Nieto, F. J. (2014). *Epidemiology: Beyond the basics*. Burlington, MA, USA: Jones & Bartlett Learning.

Takwoingi, Y., & Deeks, J. (2010). *METADAS: A SAS macro for meta-analysis of diagnostic accuracy studies* (User guide version 1.3). Retrieved from https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.470.1564&rep=rep1&type=pdf

PsychOpen GOLD

Takwoingi, Y., Guo, B., Riley, R. D., & Deeks, J. J. (2017). Performance of methods for meta-analysis of diagnostic test accuracy with few studies or sparse data. *Statistical Methods in Medical Research, 26*(4), 1896-1911. https://doi.org/10.1177/0962280215592269

Van Houwelingen, H. C., Zwinderman, K. H., & Stijnen, T. (1993). A bivariate approach to meta-analysis. *Statistics in Medicine, 12*(24), 2273-2284. https://doi.org/10.1002/sim.4780122405

Verde, P. E. (2010). Meta-analysis of diagnostic test data: A bivariate Bayesian modeling approach. *Statistics in Medicine, 29*(30), 3088-3102. https://doi.org/10.1002/sim.4055

Walter, S. D. (2002). Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. *Statistics in Medicine, 21*(9), 1237-1256. https://doi.org/10.1002/sim.1099

Wang, J., & Leeflang, M. (2019). Recommended software/packages for meta-analysis of diagnostic accuracy. *Journal of Laboratory and Precision Medicine, 4*, Article 22. https://doi.org/10.21037/jlpm.2019.06.01

Methodology is the official journal of the European Association of Methodology (EAM).

leibniz-psychology.org

PsychOpen GOLD is a publishing service by Leibniz Institute for Psychology Information (ZPID), Germany.

PsychOpen GOLD