Synthesizing Evidence

Paul-Christian Bürkner

Abstract Synthesizing evidence of diagnostic studies comes with more complications than synthesizing evidence of clinical trials. This is because the performance of diagnostic tests is evaluated for participants who have the target condition as well as for participants who do not have the target condition, usually in terms of sensitivity and specificity, respectively. There is a natural trade-off between sensitivity and specificity as lowering the threshold, at which participants will be diagnosed as positive, will increase sensitivity but at the same time reduce specificity. Thus, appropriate methods for diagnostic meta-analysis deal with *pairs* of sensitivity and specificity to preserve the bivariate nature of diagnostic accuracy. In the present chapter, we present a number of approaches to diagnostic meta-analysis and focus on the most commonly applied methods that are able to incorporate systematic variation between studies in addition do differences in the applied thresholds.

1 Introduction

Results of diagnostic studies are typically reported in terms of 2×2 tables that capture the relation of the true state of participants with the state that is diagnosed by the test under evaluation (see Table 1). The approaches presented in this chapter all assume that the true state is known and measured without error. In practice, the true state may not always be known exactly, but we can assume that a sufficiently accurate *gold standard* test exists that can serve as a benchmark for other tests.

Synthesizing evidence in diagnostic meta-analyses comes with more challenges than typical meta-analyses such as those evaluating clinical trials [12]. This is due to the fact that diagnostic studies always have two relevant outcomes (1) the accuracy of the diagnostic test for participants who have the target condition and (2) the accu-

Paul-Christian Bürkner
Institute of Psychology, University of Münster, Fliednerstr. 21, 48149, Münster, e-mail: paul.buerkner@gmail.com

Table 1 Data from a diagnostic study in a 2×2 table.

		True State		
		With target condition	Without target condition	Total
Diagnostic Test	Positive	y_{11}	y_{01}	m_1
	Negative	y_{10}	y_{00}	m_0
	Total	n_1	n_2	N

racy for participants who do not have the target condition. The former is measured by the *sensitivity* that is the proportion of participants with the target condition who are correctly identified as such:

$$\text{Sen} = \frac{y_{11}}{n_1} \tag{1}$$

The latter is measured by the *specificity* that is the proportion of participants without the target condition who are correctly identified as such:

$$\text{Spe} = \frac{y_{00}}{n_0} \tag{2}$$

It is critical to incorporate both outcomes in the evaluation of the test's performance. For instance, one could easily achieve 100% sensitivity through diagnosing everyone as positive, but this would not lead to a meaningful test since the specificity would be 0% in this case. Generally, there is a trade-off between sensitivity and specificity: Depending on where we set the threshold at which participants are diagnosed as positive, we will favor sensitivity over specificity or vise versa. The fact that studies often use different thresholds, further complicates meta-analyses of diagnostic studies. Even in the same study, multiple thresholds with corresponding sensitivities and specificities might be reported, but in the present chapter, we assume that only one pair of sensitivity and specificity is selected for each diagnostic study.

In addition to sensitivity and specificity, there are other bivariate statistics used in diagnostic studies. The *positive predictive value* (PPV) measures the probability that, given a positive test result, the diagnosed participant indeed has the target condition, while the *negative predictive value* (NPV) measures the probability that, given a negative test result, the diagnosed participant does not have the target condition. With the above introduced notation, PPV and NPV can be written as follows:

$$\text{PPV} = \frac{y_{11}}{m_1} \tag{3}$$

$$\text{NPV} = \frac{y_{00}}{m_0} \tag{4}$$

An important property of these two measures is that they depend on the prevalence of the target condition, that is the proportion of participants in the population hav-

ing the target condition at a certain point in time. Thus, we always have to keep the prevalence in mind when interpreting PPV and NPV. Also, since the prevalence might vary across studies, using these quantities for meta-analyses is somewhat more complicated.

A pair of diagnostic quantities derived directly from sensitivity and specificity are the *positive likelihood ratio* (PLR) and the *negative likelihood ratio* (NLR). They are defined as the odds that a positive and negative test result, respectively, is obtained for participants having the target condition versus those not having the target condition. More formally:

$$\text{PLR} = \frac{\text{Sen}}{1-\text{Spe}} \tag{5}$$

$$\text{NLR} = \frac{1-\text{Sen}}{\text{Spe}} \tag{6}$$

While intuitively appealing, [22] have argued against the use of likelihood ratios in meta-analyses, as summarizing them across studies may lead to impossible summary estimates for sensitivity and specificity.

Quite a few statistical methods have been proposed to tackle the problem of synthesizing evidence in diagnostic meta-analysis. In the present chapter, we will focus on the currently most common and important ones and briefly mention less common approaches at the end of the chapter.

2 The SROC curve

[NOTE TO THE EDITOR: If no chapter introduces ROC curves, this has to be done here or otherwise readers might not understand SROC curves]. One of the oldest methods developed to summarize diagnostic studies is the *summary receiver operating characteristic curve* (SROC curve; [14]). While the basic SROC approach is rarely applied in practice to date, because of the development of more advanced methods, understanding it is vital to the understanding of most other methods, and so we introduce the SROC approach first. The basic SROC curve can be obtained as follows. First, compute the quantities

$$D_i = \text{logit}(\text{Sen}_i) - \text{logit}(1-\text{Spe}_i) \tag{7}$$

$$S_i = \text{logit}(\text{Sen}_i) + \text{logit}(1-\text{Spe}_i) \tag{8}$$

for each study i. The logit transform $\text{logit}(p) = \log(p/(1-p))$ is used to transform probabilities or rates in the unit interval $[0,1]$ to values on the complete real line. The quantity D_i is the log of the diagnostic odds ratio that may also be used as a one dimensional measure of diagnostic accuracy (see Section 5). Second, fit a linear regression with D as response variable and S as predictor variable:

$$D_i = a + bS_i + e_i \tag{9}$$

The regression may also be weighted to account for differences in the measurement uncertainty between studies typically originating from varying sample sizes. Studies with less measurement uncertainty / higher sample sizes will receive higher weights. In Equation (9), a and b are the model intercept and slope respectively and e_i is the error term, which is assumed to be normally distributed. Third, having estimated a and b, we can back-transform values to the original scales to obtain the SROC curve capturing the relation between the sensitivity and the false positive rate (FPR), which is just one minus specificity.

$$\text{Sen}(\text{FPR}) = \left(1 + \exp(-\hat{a}/(1-\hat{b}))\left(\frac{1-\text{FPR}}{\text{FPR}}\right)^{(1+\hat{b})/(1-\hat{b})}\right)^{-1}. \tag{10}$$

In the above equation, $\hat{a}$ and $\hat{b}$ denote the estimates of a and b computed from the data. Consider the following example of four hypothetical diagnostic studies with sensitivities and specificities given in Table 2. Using linear regression, the estimates of a and b can be computed as $\hat{a} = 3.06$ and $\hat{b} = 0.05$. Applying formula (10) yields the SROC curve for the four diagnostic studies (see Figure 2). We see that studies differ to a non-negligible amount with respect to sensitivity and specificity, but apparently, the SROC curve provides a good fit to the data. Hence, it is plausible that differences between studies originate simply from different thresholds.

Table 2 Hypothetical data from four diagnostic studies.

Study	Sensitivity	Specificity	logit(Sen)	logit(Spe)	D	S
1	0.74	0.88	1.05	1.99	3.04	−0.95
2	0.84	0.83	1.66	1.59	3.24	0.07
3	0.95	0.57	2.94	0.28	3.23	2.66
4	0.82	0.79	1.52	1.32	2.84	0.19

A common method to summarize (S)ROC curves is the *area under the curve* (AUC). It can be interpreted as the average sensitivity of the diagnostic test taken over all possible values of the specificity [5]. The higher the AUC, the higher the accuracy of the diagnostic test, with AUC $= 0.5$ describing a useless test, classifying participants at random, and AUC $= 1$ describing a perfect test, classifying all participants correctly. The AUC might also interpreted as follows: If pairs of participants – one with and one without the target condition – are randomly drawn and tested, the AUC is equal to the probability that the participant getting the higher test result is the one with the target condition. When only a certain range of specificity values is of interest, one may compute the partial AUC, which is simply the average sensitivity over the desired range of specificity values.

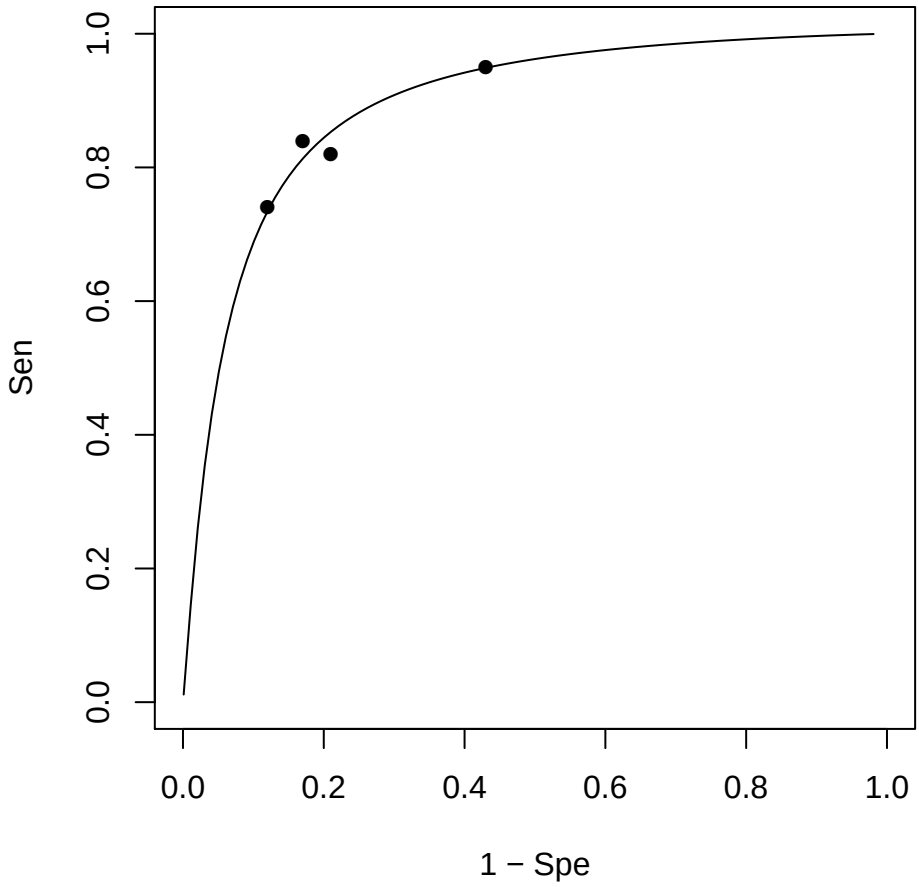

Fig. 1 SROC curve based on four hypothetical diagnostic studies. Dots indicate pairs of sensitivity and false positive rate obtained from the studies.

The basic SROC approach is easy to apply and provides a nice visualization of the relationship of sensitivity and specificity across studies. The main problem with this method is, however, that it assumes all variation between studies to originate from differences in the applied thresholds. Hence, we call the basic SROC approach a *fixed effects* model to indicate that we do not allow for systematic variation between studies apart from the applied threshold. This assumption is quite unrealistic since studies often differ considerably in other aspects such as the investigated population or the exact testing procedure. Thus, more advanced methods have been developed over the years and we will explain the two most important ones in the two upcoming sections.

3 The Hierarchical Model

To overcome the main problem of the basic SROC approach and incorporate systematic variation between studies, Rutter & Gatsonis [17] introduced their hierarchical model, which we will call the HSROC (hierarchical SROC) model in the following. It formally originates from a binomial regression model for the number of positively diagnosed participants in both study arms. We assume that y_{ij1} – where $j \in \{0,1\}$ indexes study arms (i.e. participants with and without the target condition) – are binomially distributed with probabilities π_{ij} and sample sizes n_{ij}:

$$y_{ij1} \sim \text{Binomial}(\pi_{ij}, n_{ij}) \tag{11}$$

The probabilities π_{i1} and π_{i0} are to be predicted in a regression model. The former probability refers the sensitivity of study i, whereas π_{i0} refers to the false positive rate of study i. The logit of π_{ij} should be regressed as

$$\text{logit}(\pi_{ij}) = (\theta_i + \alpha_i X_{ij}) \exp(-\beta X_{ij}). \tag{12}$$

The dummy variable X_{ij} indicates the true state of the participants being coded as $X_{i0} = -1/2$ and $X_{i1} = 1/2$. The parameters of the model, which are to be estimated, are θ_i, α_i as well as β. The model intercepts θ_i are often called cutpoint or threshold parameters since they increase with increasing thresholds and thus model the trade-off between sensitivity and specificity. This slopes α_i are called accuracy parameters since the model the difference between sensitivity and false positive rate. Both θ_i and α_i are allowed to vary between studies (as indicated by the index i) and are assumed to come from independent normal distributions:

$$\theta_i \sim N(\theta, \sigma_\theta) \tag{13}$$

$$\alpha_i \sim N(\alpha, \sigma_\alpha) \tag{14}$$

This is a so called *random effects* assumption contrasting with the fixed effects assumption of the basic SROC approach. The parameters θ, α, σ_θ, and σ_α are hyper-parameters, which are also estimated from the data. Finally, the parameter β in Equation (12) is a scale parameter, which allows for differences in the variance of diagnostic outcomes in the populations having / not having the target condition. It requires information from multiple studies and hence cannot be assumed to vary across studies.

The authors of the HSROC model proposed to fit it using fully Bayesian techniques, but it may also be fitted using classical statistical methods [11]. Having estimated the model parameters, the HSROC curve can be computed as

$$\text{Sen}(\text{FPR}) = \text{logit}^{-1}\left(\left(\text{logit}(\text{FPR}) \exp\left(\hat{\beta}/2\right) + \hat{\alpha}\right) \exp\left(\hat{\beta}/2\right)\right), \tag{15}$$

where $\text{logit}^{-1}(x) = \frac{\exp(x)}{1+\exp(x)}$ is the inverse of the logit-transform.

Adding covariates – sometimes also called (effect) moderators – to the HSROC model is straight forward: One may simply replace θ and / or α with linear predictor terms that is

$$\theta = \theta_0 + \sum_{k=1}^{K_\theta} \theta_k Z_k \tag{16}$$

$$\alpha = \alpha_0 + \sum_{k=1}^{K_\alpha} \alpha_k Z_k, \tag{17}$$

where Z_k are the covariates and θ_k as well as α_k are the corresponding regression parameters. For instance, if the aim of the test is to diagnose lung cancer, one may add the type of lung cancer investigated in each study as a dummy coded covariate. Of course, one may choose to use different covariates for θ and α or only predict one of them.

The HSROC model is a flexible and powerful approach for performing diagnostic meta-analysis. However, another (under certain conditions equivalent) random effects model is more frequently applied and is introduced in the upcoming section.

4 The Bivariate Model

Similar to the hierarchical model, the bivariate model proposed by [20], extended in [19], brought to greater attention by [15], and refined by [3] preserves the bivariate nature of diagnostic outcomes and allows for systematic variation (i.e. random effects) between studies. As noted earlier, diagnosticians have to make a trade-off between sensitivity and specificity since lowering the threshold increases sensitivity but at the same time decreases specificity. Thus, there is an inherent negative correlation between sensitivity and specificity, which is explicitly considered in the bivariate model. Formally, the logit-transformed sensitivities of each study are assumed to come from a bivariate normal distribution with mean vector $(\theta_{\text{Sen}}, \theta_{\text{Spe}})$ and covariance matrix Σ estimated from the data. We write

$$\begin{pmatrix} \text{logit}(\text{Sen}_i) \\ \text{logit}(\text{Spe}_i) \end{pmatrix} \sim N\left(\begin{pmatrix} \theta_{\text{Sen}} \\ \theta_{\text{Spe}} \end{pmatrix}, \Sigma \right) \tag{18}$$

and

$$\Sigma = \begin{pmatrix} \sigma^2_{\text{Sen}} & \rho\,\sigma_{\text{Sen}}\sigma_{\text{Spe}} \\ \rho\,\sigma_{\text{Sen}}\sigma_{\text{Spe}} & \sigma^2_{\text{Spe}} \end{pmatrix}, \tag{19}$$

where σ_{Sen} and σ_{Spe} denote the standard deviation across studies of logit sensitivity and specificity respectively and ρ denotes their correlation. Sensitivities and specificities are measured with different precision due to varying sample sizes both within and between studies. Studies including more participants achieve higher precision / lower variance and should thus receive higher weights in the meta-analysis. The within study variances $s^2_{\text{Sen}_i}$ and $s^2_{\text{Spe}_i}$ of logit sensitivity and specificity of study i

can be approximated as follows:

$$s^2_{\text{Sen}_i} = \frac{1}{n_{i1}\ \text{Sen}_i(1-\text{Sen}_i)} \tag{20}$$

$$s^2_{\text{Spe}_i} = \frac{1}{n_{i0}\ \text{Spe}_i(1-\text{Spe}_i)} \tag{21}$$

For smaller sample sizes or in case of zero sensitivity or specificity, these approximations my not be accurate [18]. Therefore, one should rather use the binomial parameterization of the bivariate model as noted by [3]. Denoting with S_i the within study variance matrix that is

$$S_i = \begin{pmatrix} s^2_{\text{Sen}_i} & 0 \\ 0 & s^2_{\text{Spe}_i} \end{pmatrix}, \tag{22}$$

the complete bivariate model for diagnostic meta-analysis is given by

$$\begin{pmatrix} \text{logit}(\text{Sen}_i) \\ \text{logit}(\text{Spe}_i) \end{pmatrix} \sim N\left(\begin{pmatrix} \theta_{\text{Sen}} \\ \theta_{\text{Spe}} \end{pmatrix}, \Sigma + S_i \right). \tag{23}$$

The parameters θ_{Sen} and θ_{Spe} denote the meta-analytic logit sensitivity and specificity respectively. Together with their estimated confidence intervals, they may be transformed back to the original metric by applying the inverse of the logit-transform. The standard deviations σ_{Sen} and σ_{Spe} provide information on the variation of sensitivity and specificity across studies, for instance due to different thresholds or other systematic differences between studies. Finally, ρ captures the (usually negative) correlation between sensitivity and specificity. Similar to the HSROC model, moderators may easily be introduced by replacing θ_{Sen} and θ_{Spe} with linear predictors, that is

$$\theta_{\text{Sen}} = \theta_{\text{Sen}0} + \sum_{k=1}^{K_{\text{Sen}}} \theta_{\text{Sen}k} Z_k \tag{24}$$

$$\theta_{\text{Spe}} = \theta_{\text{Sen}0} + \sum_{k=1}^{K_{\text{Spe}}} \theta_{\text{Spe}k} Z_k. \tag{25}$$

The bivariate model can easily be fit using standard statistical software (cf. chapter *Statistical Packages*). Although the HSROC and the bivariate model may look rather different at first glance, it has been shown by [8] and likewise and independently by [1] that they are very closely related and even equivalent in the absence of covariates. Since the bivariate model is easier to fit and perhaps also easier to understand, it has become the standard approach for meta-analysis of diagnostic studies and we highly recommend its application when only one diagnostic test is being evaluated [12]. The bivariate model can also be applied if quantities other than sensitivity and specificity – such as the positive and negative likelihood ratio – are of primary interest, as one can generate samples for observed sensitivities and specifities based

on the fitted model parameters and then use these samples to obtain estimates for other quantities [22]. In fact, models that directly target alternative quantities may be so complicated that using the bivariate model, instead, is advised even in these cases [22].

4.1 Other Bivariate Models

Several other methods have been proposed that provide either extentions or alternatives to the standard bivariate model. More advanced methods of fitting the bivariate model include composite likelihood [2] or simulation based methods [7]. Alternative models include the method of [9] using the Lehmann family, the method of [16] based on the Youden index, non-parametric approaches [13, 21], semi-parametric mixtures [4] or copulas [10]. Due to the large number of alternative models, we will not discuss them in more detail in the present chapter. [NOTE TO THE EDITOR: Antonia Zapf might discuss some of these methods in her chapter, and when she does, we should reference it here]. Generalizations of the bivariate model for the comparison of multiple diagnostic tests are introduced in Chapter *Network Meta-Analysis of Diagnostic Test Accuracy Studies*.

5 Univariate Approaches

Although it is highly recommended to keep the bivariate nature of diagnostic data in order not to lose information, we want to briefly note the possibility of univariate diagnostic meta-analysis. In order to apply standard meta-analytic techniques such as those used for clinical trials, one has to find a univariate measure of diagnostic accuracy that is approximately normally distributed. Glas et al. [6] proposed to use the log diagnostic odds ratio D for this purpose, which we have already introduced in Equation (7). The log diagnostic odds ratio has some favorable properties as compared to other unvariate measures such that it is relatively robust to varying threshold across studies [6]. The variance of its estimator can be computed as

$$\mathrm{Var}(\hat{D}_i) = \frac{1}{y_{i11}} + \frac{1}{y_{i10}} + \frac{1}{y_{i01}} + \frac{1}{y_{i00}}. \tag{26}$$

Meta-Analysis may be performed using the standard univariate model for normally distributed outcomes:

$$\hat{D}_i \sim N(\theta, \sigma_\theta^2 + \mathrm{Var}(\hat{D}_i)), \tag{27}$$

where θ is the meta-analytic estimate across studies and σ_θ^2 is the between study variance. We do not recommend using a univariate approach to diagnostic meta-analysis, but if one still wants to apply it for some reason, the log diagnostic odds ratio should be the measure of choice.

6 Conclusion

In the present chapter, we introduced several methods to synthesize evidence of diagnostic studies. Emphasize was put on the bivariate nature of diagnostic outcomes, as performance of diagnostic tests is evaluated for participants who have the target condition and participants who do not have the target condition. Thus, appropriate methods analyze pairs of sensitivity and specificity. Despite other reasonable methods, we recommend applying the bivariate model – or one of its extensions – as it allows for systematic variation between studies in addition to differences in the applied thresholds, while still being relatively easy to fit using standard statistical software.

7 Acknowledgments

I want to thank Prof. Philipp Doebler and Prof. Gerta Rücker for their very helpful comments on this chapter.

References

1. Lidia R Arends, Hussien Hamza, Hans C Van Houwelingen, Majanka H Heijenbrok-Kal, Maria G M Hunink, and Theo Stijnen. Bivariate random effects meta-analysis of roc curves. *Medical Decision Making*, 28(5):621–638, 2008.
2. Yong Chen, Yulun Liu, Jing Ning, Lei Nie, Hongjian Zhu, and Haitao Chu. A composite likelihood method for bivariate meta-analysis in diagnostic systematic reviews. *Statistical Methods in Medical Research*, page 0962280214562146, 2014.
3. Haitao Chu and Stephen R Cole. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *Journal of Clinical Epidemiology*, 59(12):1331–1332, 2006.
4. Philipp Doebler and Heinz Holling. Meta-analysis of diagnostic accuracy and roc curves with covariate adjusted semiparametric mixtures. *Psychometrika*, 80(4):1084–1104, 2015.
5. Constantine Gatsonis and Prashni Paliwal. Meta-analysis of diagnostic and screening test accuracy evaluations: methodologic primer. *American Journal of Roentgenology*, 187(2):271–281, 2006.
6. Afina S Glas, Jeroen G Lijmer, Martin H Prins, Gouke J Bonsel, and Patrick MM Bossuyt. The diagnostic odds ratio: a single indicator of test performance. *Journal of Clinical Epidemiology*, 56(11):1129–1135, 2003.
7. Annamaria Guolo. A double simex approach for bivariate random-effects meta-analysis of diagnostic accuracy studies. *BMC Medical Research Methodology*, 17(1):6, 2017.
8. Roger M Harbord, Jonathan J Deeks, Matthias Egger, Penny Whiting, and Jonathan AC Sterne. A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics*, 8(2):239–251, 2007.
9. Heinz Holling, Walailuck Böhning, and Dankmar Böhning. Meta-analysis of diagnostic studies based upon sroc-curves: a mixed model approach using the lehmann family. *Statistical Modelling*, 12(4):347–375, 2012.

10. Oliver Kuss, Annika Hoyer, and Alexander Solms. Meta-analysis for diagnostic accuracy studies: a new statistical model using beta-binomial distributions and bivariate copulas. *Statistics in Medicine*, 33(1):17–30, 2014.
11. Petra Macaskill. 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, 2004.
12. Petra Macaskill, Constantine Gatsonis, Jonathan Deeks, Roger Harbord, and Yemisi Takwoingi. *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy*. London: The Cochrane Collaboration, 2010.
13. Pablo Martínez-Camblor. Fully non-parametric receiver operating characteristic curve estimation for random-effects meta-analysis. *Statistical Methods in Medical Research*, 26(1):5–20, 2017.
14. Lincoln E Moses, David Shapiro, and Benjamin Littenberg. 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, 1993.
15. Johannes B Reitsma, Afina S Glas, Anne WS Rutjes, Rob JPM Scholten, Patrick M Bossuyt, and Aeilko H Zwinderman. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology*, 58(10):982–990, 2005.
16. Gerta Rücker and Martin Schumacher. Summary roc curve based on a weighted youden index for selecting an optimal cutpoint in meta-analysis of diagnostic accuracy. *Statistics in Medicine*, 29(30):3069–3078, 2010.
17. Carolyn M Rutter and Constantine A Gatsonis. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in Medicine*, 20(19):2865–2884, 2001.
18. Michael J Sweeting, Alexander J Sutton, and Paul C Lambert. What to add to nothing? use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in Medicine*, 23(9):1351–1375, 2004.
19. Hans C Van Houwelingen, Lidia R Arends, and Theo Stijnen. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine*, 21(4):589–624, 2002.
20. Hans C Van Houwelingen, Koos H Zwinderman, and Theo Stijnen. A bivariate approach to meta-analysis. *Statistics in Medicine*, 12(24):2273–2284, 1993.
21. Antonia Zapf, Annika Hoyer, Katharina Kramer, and Oliver Kuss. Nonparametric meta-analysis for diagnostic accuracy studies. *Statistics in Medicine*, 34(29):3831–3841, 2015.
22. Aeilko H Zwinderman and Patrick M Bossuyt. We should not pool diagnostic likelihood ratios in systematic reviews. *Statistics in Medicine*, 27(5):687–697, 2008.