Research Article

Statistics in Medicine

Received 22 July 2013, Accepted 27 March 2014 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6177

Testing for publication bias in diagnostic meta-analysis: a simulation study

Paul-Christian Bürkner*† and Philipp Doebler

The present study investigates the performance of several statistical tests to detect publication bias in diagnostic meta-analysis by means of simulation. While bivariate models should be used to pool data from primary studies in diagnostic meta-analysis, univariate measures of diagnostic accuracy are preferable for the purpose of detecting publication bias. In contrast to earlier research, which focused solely on the diagnostic odds ratio or its logarithm ($\ln \omega$), the tests are combined with four different univariate measures of diagnostic accuracy. For each combination of test and univariate measure, both type I error rate and statistical power are examined under diverse conditions. The results indicate that tests based on linear regression or rank correlation cannot be recommended in diagnostic meta-analysis, because type I error rates are either inflated or power is too low, irrespective of the applied univariate measure. In contrast, the combination of trim and fill and $\ln \omega$ has non-inflated or only slightly inflated type I error rates and medium to high power, even under extreme circumstances (at least when the number of studies per meta-analysis is large enough). Therefore, we recommend the application of trim and fill combined with $\ln \omega$ to detect funnel plot asymmetry in diagnostic meta-analysis. Copyright © 2014 John Wiley & Sons, Ltd.

Keywords: diagnostic meta-analysis; publication bias; diagnostic odds ratio; trim and fill; simulation study

1. Introduction

The interest in research synthesis and meta-analysis has rapidly grown over the last few decades. From today's point of view, it is difficult to think of scientific research without the possibility to integrate different findings into one big picture. This is especially true for research on the accuracy of diagnostic tests, because clinical and health policy decisions as well as technology development and evaluation in diagnostic medicine have to rely on a good empirical basis [1–3]. In general, studies of diagnostic accuracy compare the performance of an imperfect diagnostic test with an optimal diagnostic instrument, which is also known as gold standard. Usually, using the gold standard is time consuming, expensive, and/or invasive, so that it can only be applied in certain circumstances. Therefore, alternative tests that are more economic but less accurate have to be evaluated in studies of diagnostic accuracy by comparing them with the gold standard. These studies most often report pairs of *sensitivity* (Sen) and *specificity* (Spe), the former being the percentage of correctly diagnosed diseased individuals and the latter being the percentage of correctly diagnosed healthy individuals.

As in any other area of research synthesis, results in diagnostic meta-analysis may be biased by various effects such as study quality, heterogeneity of the examined populations, or, as most commonly cited, *publication bias* (PB) [4]. PB arises when studies *systematically* remain unpublished so that they cannot contribute to meta-analysis leading to biased and possible misleading results. 'Systematically' comprises that studies are not just missing at random, but that some of the study's characteristics (most often its outcomes) influence the probability of this study being published.

There are several reasons for researchers to refrain from publishing a study. Some of these reasons occur at random and do not depend on the outcome of the study (e.g., the retirement of the responsible researcher), so that they do not contribute to the emergence of PB, while other reasons do depend on the outcome: For example, a study's findings may be suppressed by the funding source supporting the

Psychology and Sport Sciences, Fliednerstr. 21, D-48149, Münster, Germany
**Correspondence to: Paul-Christian Bürkner, Psychology and Sport Sciences, Fliednerstr. 21, D-48149, Münster, Germany.*
†E-mail: p_buer02@uni-muenster.de

research. Besides, the possibly most prominent reason is the non-significance of the results [5]. However, this non-significance rarely applies to diagnostic studies. Usually, findings on test accuracy contain Sen and Spe (or univariate measures (UMs) computed on their basis) together with 95% confidence intervals, but they do not state a null hypothesis [6, 7]. Hence, in most cases, there is no statistical test that can potentially fail to be significant. Therefore, one might argue that PB is not a problem in diagnostic meta-analysis. Nevertheless, some approaches such as non-inferiority designs (in which a new treatment or diagnostic test is compared with an already established one) are based on significance tests even in diagnostic contexts [8–10]. In addition, as studies of diagnostic accuracy are often conducted as part of routine clinical practice, they may be abandoned more easily, which should increase the probability of PB [3, 11]. Another reason accounting for PB in all types of meta-analysis is that dramatic findings are more likely to occur in primary studies and meta-analyses that are methodically weaker [12]. As scientific journals may prefer such findings, it is plausible that meta-analyses not only tend to overestimate effects because of unpublished studies, but that researchers also base their results on less elaborate literature search. This assumption has been supported by findings of Song *et al.* [11] who showed that publication bias was increased in diagnostic meta-analyses in which only few databases were used to find relevant studies. In addition, we share the view of one reviewer of this paper who suspects that potential mechanisms of PB in the diagnostic context will depend on the type of test (e.g., imaging vs. biomarker vs. questionnaire) and the application (e.g., mass screening vs. confirmation of a screen). In light of the preceding discussion, it becomes clear that the examination and evaluation of PB in diagnostic meta-analysis should not be neglected.

An issue that complicates the detection of PB in diagnostic meta-analysis emerges from different and sometimes implicit cutoff values used in different studies in order to decide which test scores indicate disease and which do not. The choice of the cutoff value depends on the (somewhat subjective) importance of Sen and Spe in the respective research context. In one study, an individual may be diagnosed as diseased, while in another study, with a different cutoff value, the same person may be diagnosed as healthy. Therefore, even very different pairs of observed Sen and Spe might only be caused by different cutoff values and may not indicate that the test accuracy itself varied between studies. Using univariate effect measures such as the diagnostic odds ratio or its logarithm instead of Sen and Spe can reduce the cutoff value problem [13] at the cost of a loss of information. However, the heterogeneity caused by different cutoff values may still be interpreted as PB, even though there is none. In contrast, different cutoff values can also mask an existing PB, so that it cannot be detected. Given all these complications, we evaluated the performance of several statistical tests to detect PB in diagnostic meta-analysis in a comprehensive simulation study.

In general, models of meta-analysis simultaneously have to cope with between and within study variance. Therefore, random effects models are of critical importance for many types of meta-analysis, because they are able to separate within study variance from between-study variance. Rutter and Gatsonis [14] were the first to develop a specific approach of random effects for diagnostic meta-analysis by using hierarchical regression. A few years later, Reitsma *et al.* [15] formulated a different, bivariate model to cope with random effects. It has later been shown by Harbord *et al.* [16] and likewise and independently by Arends *et al.* [17] that the models of Rutter and Gatsonis [14] and Reitsma *et al.* [15] are very closely related and even identical in the absence of covariates. In the present study, the model of Reitsma *et al.* [15] was used to sample pairs of Sen and Spe, as a bivariate model should be used for the analysis of data from primary diagnostic studies [18–20]. However, *pairs* of Sen and Spe are difficult to use directly to detect PB, as the common graphical and statistical methods require *UMs*, and to our knowledge, bivariate tests have not been developed so far [19]. In our study, four UMs that were computed on the basis of Sen and Spe were each combined with every applied statistical test. The performance of these combinations was evaluated by means of simulation.

Section 2 reviews the bivariate model, different UMs used in diagnostic studies, and existing statistical tests to detect PB. The simulation process and its results are described in detail in Sections 3 and 4. In Section 5, the findings, implications, and limitations of the present study are summarized and discussed.

2. Theory and methods

2.1. Bivariate sampling model

As mentioned earlier, the bivariate model of Reitsma *et al.* [15] was used as a sampling model for our simulations. Let k be the number of studies summarized in the meta-analysis and $i = 1, \ldots, k$. The basic

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

		Diagnostic test		
		Positive	Negative	Total
Gold standard	Positive	x	w	n_1
	Negative	y	z	n_2
	Total	m_1	m_2	N

assumption is that the true pair of $\text{logit}(\text{Sen}_i) = \theta_{A,i}$ and $\text{logit}(1 - \text{Spe}_i) = \theta_{B,i}$ is bivariate normally distributed with mean $\mu = (\theta_A,\ \theta_B)^{\text{T}}$ and between-study covariance matrix Σ:

$$\begin{pmatrix} \theta_{A,i} \\ \theta_{B,i} \end{pmatrix} \sim \text{N}(\mu,\ \Sigma) \quad \text{with} \quad \Sigma = \begin{pmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{pmatrix}. \tag{1}$$

Despite that this model explicitly considers the bivariate nature of diagnostic data, it has the advantage that it directly accounts for the covariance between Sen and Spe through σ_{AB}.

2.2. *Univariate effect measures*

Comparing diagnostic tests only on the basis of Sen and Spe can be problematic, as one test may have a higher Sen, while the other test may have a higher Spe. If this is the case, it will be difficult to decide which test is to be preferred unless one indicator is obviously more important. The advantage of UMs combining the information on Sen and Spe is that different diagnostic tests are always comparable, at the expense of losing some information due to the reduction to one single measure. With respect to the present study, UMs are *only* of importance, as statistical tests to detect PB require UMs instead of pairs of Sen and Spe [19]. In the following, four different UMs are introduced. The notation is given in Table I.

The probably most prominent UM in the context of diagnostic studies is the diagnostic odds ratio (ω; or its natural logarithm $\ln \omega$). In general, one computes the *standard error* (SE) and the related confidence interval of $\ln \omega$ as it is approximately normally distributed unless the observed frequencies (i.e., w, x, y, and z) are too small. It holds that

$$\ln \omega := \ln \left(\frac{xz}{yw} \right) \tag{2}$$

with

$$\text{SE}(\ln \omega) = \sqrt{\frac{1}{x} + \frac{1}{y} + \frac{1}{w} + \frac{1}{z}}. \tag{3}$$

The range of $\ln \omega$ is $-\infty$ to $+\infty$, with higher values standing for higher test accuracy and values lower than zero representing tests that are worse than guessing. In case of zero cells in the underlying 2×2 table, it is generally recommended to add 0.5 to each observed frequency in order to calculate an approximation of $\ln \omega$ [13, 21], although there are other methods that may be less biased [22]. Also, $\ln \omega$ is not affected by unequal sample sizes [23]. Importantly, tests to detect PB in diagnostic meta-analysis are almost without exception based on ω or $\ln \omega$, while other UMs have not been discussed so far (see [24] for the only exception known to the authors, in which the risk ratio was examined and performed worse then $\ln \omega$. Therefore, the risk ratio was not included in our simulations). Among other aims, the present study investigated whether this focus on the diagnostic odds ratio can be regarded as justified.

Another UM proposed by Le [25] is derived from a model, which describes the relationship between Sen and Spe using the Lehmann family:

$$\text{Sen} = (1 - \text{Spe})^{\vartheta} \quad \text{with} \quad 0 < \vartheta \leqslant 1. \tag{4}$$

This model allows to approximate the receiver operating characteristic (ROC), which is often used to summarize diagnostic results that are continuous (i.e., not binary [25–27]). Solving for the natural logarithm of ϑ reveals

$$\ln \vartheta := \ln\left(\frac{\ln(x) - \ln(n_1)}{\ln(y) - \ln(n_2)}\right) \tag{5}$$

with the SE

$$\mathrm{SE}(\ln \vartheta) = \sqrt{\frac{\frac{1}{x} - \frac{1}{n_1}}{(\ln(x) - \ln(n_1))^2} + \frac{\frac{1}{y} - \frac{1}{n_2}}{(\ln(y) - \ln(n_2))^2}}. \tag{6}$$

Similar to $\ln \omega$, $\ln \vartheta$ is not affected by unequal sample sizes and ranges from $-\infty$ to $+\infty$ with values greater than zero representing tests that are worse than guessing. Thus, in contrast to $\ln \omega$, higher values represent lower test accuracy. However, some tests for PB require higher values to be associated with higher test accuracy. Therefore, $-\ln \vartheta$ was used in our simulations instead.

Youden's index [28], in our notation, is written as

$$Y := \frac{x}{n_1} + \frac{z}{n_2} - 1, \tag{7}$$

is yet another UM used in diagnostic studies. The value of Y is fairly constant, even if Sen and Spe are varying due to different cutoff scores [29]. Also, Y appears to be a better measure to choose an appropriate cutoff value than ω [30]. These are important properties when dealing with the cutoff value problem. Moreover, Y is very simple to calculate, it is unaffected by unequal sample sizes, and its values range from -1 to 1. Assuming that a diagnostic test is equal or better than a random decision, Y only ranges from 0 to 1 (the higher the Y, the better the test). It holds that

$$\mathrm{SE}(Y) = \sqrt{\frac{\frac{x}{n_1}\left(1 - \frac{x}{n_1}\right)}{n_1} + \frac{\frac{y}{n_2}\left(1 - \frac{y}{n_2}\right)}{n_2}}. \tag{8}$$

Due to the simplicity of Y and its usefulness when choosing an appropriate cutoff value, it has frequently been applied in diagnostic studies. With respect to meta-analysis, Böhning *et al.* [29] proposed an overall estimator for the performance of a diagnostic test based on Y, including the possibility to add mixtures to cope with unobserved heterogeneity.

A measure that behaves somewhat similar to Y is Cohen's kappa [31]:

$$K := \frac{2(xz - yw)}{n_1 m_2 + n_2 m_1}. \tag{9}$$

Although it can be applied in diagnostic contexts [32, 33], K is generally known as a measure of the agreement between two or (if generalized) more raters [34], but not as a measure of diagnostic accuracy. Accordingly, when compared with the other measures mentioned earlier, K is less often applied in diagnostic studies. Calculating $\mathrm{SE}(K)$ is quite complex. The SE used in the present study was proposed by Fleiss *et al.* [35] (under a multinomial assumption) and is the most common:

$$\mathrm{SE}(K) = \frac{\sqrt{A + B - C}}{(1 - (n_1 m_1 + n_2 m_2)/N^2)\sqrt{N}}, \tag{10}$$

where

$$\begin{aligned} A &:= \frac{1}{N^3}(x(N - (n_1 + m_1(1 - K))^2 + w(N - (n_2 + m_2(1 - K))^2), \\ B &:= \frac{1}{N^3}(1 - K)^2(w(n_2 + m_1)^2 + y(n_1 + m_2)^2), \\ C &:= (K - (n_1 m_1 + n_2 m_2)(1 - K)/N^2)^2. \end{aligned} \tag{11}$$

When sample sizes are equal in both groups, K is identical to Y with $\mathrm{SE}(K)$ and $\mathrm{SE}(Y)$ being very similar. However, when sample sizes are unequal, both measures differ, because in contrast to the other UMs, K depends on the sample sizes per group. For sample distributions of the UMs in the absence and presence of PB, see Figure 2.

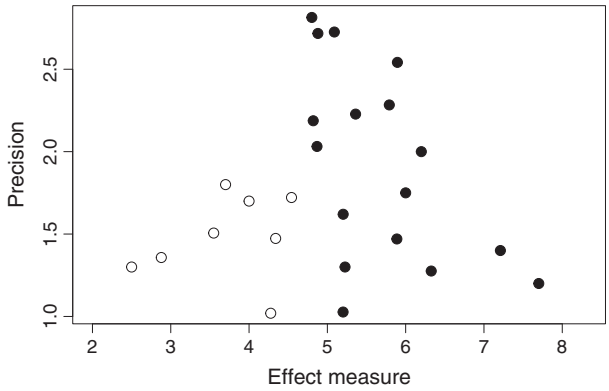

Figure 1. Hypothetical funnel plot illustrating published studies (solid points) and unpublished studies (white points). As the unpublished studies cannot be considered for meta-analysis, the funnel emerged from the solid points is asymmetric and hints at the presence of PB.

2.3. *Detecting publication bias using funnel plots*

A basic graphical method to detect PB are funnel plots, in which effect measures (such as those mentioned earlier) are plotted against the total sample size N or the precision SE^{-1} of the effect measure [36]. The shape of the funnel plot visualizes whether PB is existent or not. If not existent, the points will form a symmetrical funnel around an overall estimated effect. More precisely, studies with low N or high SE should spread more around the overall effect, while studies with high N or low SE should be closer to it [4]. However, if small studies with low or non-significant effects remain unpublished, the funnel plot should appear asymmetrical, which may lead to the conclusion that PB is existent. A sample funnel plot of a hypothetical meta-analysis in which PB is present is depicted in Figure 1.

Notably, funnel plots and statistical tests cannot discriminate between PB and other sources that cause funnel plot asymmetry [37]. The cutoff value problem in diagnostic meta-analysis further complicates the detection of PB. As funnel plots of Sen and Spe do particularly suffer from this cutoff value problem, UMs should be used instead. Glas *et al.* [13] showed that the diagnostic odds ratio might depend on the cutoff (as implied by the model in [14]) but is reasonably constant nevertheless. In this sense, UMs alleviate the cutoff value problem. Moreover, three-dimensional funnel plots displaying Sen and Spe at the same time are possible, but they may be too difficult for reasonable interpretation. Yet even without these complications, funnel plots cannot be seen as a reliable method to detect PB [38]. Therefore, statistical tests are needed.

2.4. *Detecting publication bias using statistical tests*

The present study examined the performance of several statistical tests that were developed to detect PB by means of simulation. All of these tests were combined with each of the four UMs (i.e., $\ln \omega$, $\ln \vartheta$, Y, and K). The tests do all have in common that they are more or less based on funnel plots. Contrary to funnel plots, however, formal instead of visual criteria decide if PB is present. When PB is genuinely existent in diagnostic meta-analysis, studies that found low diagnostic accuracy remain unpublished and cannot be considered for meta-analysis. Therefore, by using UMs with higher values corresponding to higher diagnostic accuracy as in the present study, missing studies should appear on the left part of the funnel. Conclusively, it seems reasonable to apply tests for detecting PB with one-sided (instead of two-sided) hypotheses.

Concerning the performance of a statistical test, two types of errors are of importance. A type I error (e_1) occurs when the test is significant, even though there is no PB. In contrast, a type II error (e_2) occurs, when there is PB but the test fails to be significant. Let α be the nominal α level of a test and $P(x)$ the probability of an event x. A test is called conservative when $P(e_1) < \alpha$, whereas it is called liberal (or anti-conservative) when $P(e_1) > \alpha$. The power of tests is generally defined as $1 - P(e_2)$. Taken together, statistical tests are considered as good when they are not liberal and have a high power.

2.4.1. Egger regression. Egger *et al.* [39] used a linear regression approach to detect PB. They suggested to regress an UM divided by its SE as dependent variable on the precision of that UM as independent variable:

$$\mathrm{UM} \cdot \mathrm{SE}^{-1} = b_0 + b_1 \mathrm{SE}^{-1} + \varepsilon, \tag{12}$$

where ε represents the error arising from the prediction. In case of no PB, small studies are expected to be close to zero on both axes (due to their high SE), whereas larger studies have a high precision and are thus expected to deviate from zero. In this case, the regression line is assumed to go through the origin, so that b_0 does not differ significantly from zero. However, if PB is existent, most small studies will have relatively high UMs, and b_0 is thus assumed to be significantly greater than zero. The Egger regression can be altered by replacing SE^{-1} in (12) by the total sample size N. In addition, each study can be weighted by the inverse of the variance under a fixed or random effects assumption to control for possible heteroscedasticity (cf. [40]). These variations of the Egger regression have in common that they are mostly too liberal when combined with (diagnostic) odds ratio or its logarithm [4, 24, 41]. For this reason, Harbord *et al.* [42] proposed another variation of Egger's regression specifically designed for meta-analysis of binary outcomes, which uses the efficient score of an UM and the variance of the efficient score (for details, see [42]). Furthermore, there are alternative regression tests for meta-analysis of binary outcomes, which utilize the arcsine transformation in order to hold the nominal α level [43].

2.4.2. Macaskill regression. Macaskill *et al.* [41] introduced a different regression approach to detect PB. They suggested to regress an UM as dependent variable on the total sample size N as independent variable and weight the studies by the inverse of the variance:

$$\mathrm{UM} = b_0 + b_1 N + \varepsilon. \tag{13}$$

In case of no PB, the values of the UM should be constant across different sample sizes, and b_1 is thus assumed to be close to zero. However, when PB is present, small studies should have greater UMs on average, which causes b_1 to be below zero. Deeks *et al.* [4] suggested to replace N by $1/\sqrt{\mathrm{ESS}}$ (i.e., the *effective sample size* calculated as $\mathrm{ESS} = (4n_1n_2)/(n_1 + n_2)$) and weight the studies by ESS to achieve a better performing test, especially when there are only few diseased persons in the sample. A similar approach was suggested by Peters *et al.* [44], who replaced N by $1/N$ (for details on the weighting, see [44, 45]). Note that, when ESS or $1/N$ are used, $b_1 > 0$ has to be tested. Although it seems reasonable to further variate the Macaskill regression by replacing N by SE, Sterne *et al.* [46] demonstrated that the resulting test is equivalent to Egger's regression. Notably, in the approach of Egger *et al.* [39], b_0 is tested for significance in order to detect PB, whereas the method of Macaskill *et al.* [41] uses b_1.

2.4.3. Begg's rank correlation. Begg and Mazumdar [47] proposed a non-parametric rank correlation method to detect PB, which is based on Kendall's tau (τ [48]). Let t_i be the effect measure and $\mathrm{Var}_i := \mathrm{Var}(t_i) = \mathrm{SE}(t_i)^2$ the related variance of study i in the meta-analysis. Then one calculates

$$t_i^* := (t_i - \bar{t})\,/\mathrm{SE}_i \tag{14}$$

with

$$\bar{t} := \left(\sum \mathrm{Var}_i^{-1} t_i\right) / \sum \mathrm{Var}_i^{-1} \tag{15}$$

being the common fixed effects estimator of the overall effect. It is then tested whether t_i^* and Var_i are significantly associated (i.e., if τ differs significantly from zero). In the absence of PB, the variance should be independent of the effect measure, and thus, τ is assumed to be close to zero. In the presence of PB, some small studies will have inflated effect measures and hence large values t_i^*. For that reason, a τ that is significantly greater than zero indicates the presence of PB. It has been shown that Begg's rank correlation is a little conservative and has less power than Egger's regression when used with (diagnostic) odds ratio or its logarithm [4, 41, 46]. Possible variations of Begg's rank correlation can be obtained by replacing Var by N^{-1} or ESS^{-1}, respectively [4]. Further variation specifically designed for binary data were proposed by Schwarzer *et al.* [49] and Rücker et al. *et al.* [43].

2.4.4. Trim and fill. Trim and fill is another non-parametric method to detect PB and was developed by Duval and Tweedie [50]. It is based on the idea that there are k studies present in the meta-analysis and k_0 studies missing due to PB, which implies an asymmetrical funnel. With respect to trim and fill, funnel plots are applied with an UM on the x-axis and its precision (or alternatively N) on the y-axis. If we assume to know the true overall effect Θ, we will be able to estimate k_0. Let $t_i^+ := t_i - \Theta$, r_i^+ be the

Table II. Short forms of the tests to detect PB.

Test	Short form	Additional notations
Egger	$E(t, v, w)$	w: weights of each study
Macaskill	$M(t, v)$	
Begg	$B(t, v)$	
Trim and fill	$T(t, v, m)$	m: estimator of k_0

rank of $|t_i^+|$ $(1 \leqslant r_i^+ \leqslant k)$ and $\gamma^+ \geqslant 0$ be the rightmost run of ranks associated with positive values of t_i^+. Then k_0 can be estimated by

$$R := \gamma^+ - 1 \tag{16}$$

or

$$L := \frac{4\sum_{t_i^+>0} r_i^+ - k(k+1)}{2k-1}. \tag{17}$$

In addition, the authors of trim and fill proposed a third estimator of the number of missing studies but did not recommend its usage [50]. Therefore, this estimator was not considered in the present study.

As Θ is usually unknown, it has to be estimated as well. Using an easily computed iterative algorithm (discussed in detail in [50]), one arrives at a random effects estimator $\hat{\Theta}$ that is used in the preceding formula. In the absence of PB (i.e., $k_0 = 0$), the approximate distributions of R and L are known. Thus, R and L can be tested for significance to decide whether PB is present or not. As $\hat{\Theta}$ can be seen as corrected for PB [50], trim and fill is not only a method to test for PB, but it also offers to correct the overall estimator (for further discussion, see [50, 51]). Note that trim and fill can only be applied with one-sided hypotheses as apposed to all other tests mentioned earlier.

In literature, several alternative methods other than trim and fill were proposed to estimate the number of missing studies [52–54], but they are far more complex than any of the methods discussed in this section. As a method's success rests not only upon its performance but also upon its simplicity and applicability in practice, these complex alternative methods were not considered in the present study.

All other statistical tests discussed in this section were included in our simulation, and each of them was combined with every of the four UMs (and with every of the accuracy measures SE, N, and ESS) as far as possible and feasible. In view of the large number of tests for PB simulated in our study, it makes sense to introduce a short form at least for those combinations, which are frequently mentioned in our results and discussion (Table II; for notational convenience, UMs are denoted as t, whereas measures of its accuracy are denoted as v). Note that not all test variations have a short form in order to keep the number of notations at an acceptable level.

For example, if Y is used in the original version of Egger's regression with every study being equally weighted (so that the weight argument can be suppressed), the resulting test will be shortened to

$$E(Y, \mathrm{SE}). \tag{18}$$

If $\ln \vartheta$ is plotted against N and tested with trim and fill while k_0 is estimated by the statistic R stated in (16), we will write

$$T(\ln \vartheta, N, R). \tag{19}$$

3. Simulations

In order to evaluate the performance of the tests to detect PB discussed earlier, diagnostic data were simulated with and without PB. In addition, several other parameters such as the number of studies per meta-analysis k or the mean quality μ of the diagnostic test were systematically varied. In the present simulation, k took on values of $k = 30$ (many studies) or $k = 10$ (few studies). The total sample size N of each of the k studies was randomly sampled from a discrete uniform distribution and varied between $N = 50$ and $N = 1000$. The prevalence π (i.e., the rate of diseased individuals) took on values of

Table III. Summary of all parameters varied.

μ	Σ	k	π	Bias
Random decision: $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$	Fixed effects: $\Sigma = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$	Few studies: $k = 10$	Balanced: $\pi = 0.5$	None: No PB
Low accuracy: $\mu = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$	Small random effects: $\Sigma = \begin{pmatrix} 0.5 & 0.3 \\ 0.3 & 0.5 \end{pmatrix}$	Many studies: $k = 30$	Unbalanced: $\pi = 0.2$	Selection small: Removal of $0.2k$ studies with the lowest Youden index
High accuracy: $\mu = \begin{pmatrix} 2 \\ -2 \end{pmatrix}$	Large random effects: $\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$			Selection large: Removal of $0.4k$ studies with the lowest Youden index
High sensitivity: $\mu = \begin{pmatrix} 2 \\ -1 \end{pmatrix}$				Mixture small: Mixture of two distributions with the mean of the second shifted by $(0.75, -0.75)^{\mathrm{T}}$
				Mixture large: Mixture of two distributions with the mean of the second shifted by $(1.25, -1.25)^{\mathrm{T}}$

$\pi = 0.5$ ($n_1 = n_2 = 0.5N$; balanced) or $\pi = 0.2$ ($n_1 = 0.2N$, $n_2 = 0.8N$; unbalanced). The resulting n_1 and n_2 were always rounded to nearest integer. Simulations were implemented in R [55–57] with parts of the program coded in C++ in order to decrease the simulations' duration.

In contrast to earlier simulation models concerning diagnostic accuracy (cf. [4]), for each study, the true logits of Sen and $1 - \mathrm{Spe}$ were directly sampled using the model of Reitsma *et al.* [15] stated in (1), to consider the bivariate structure of diagnostic data. Both μ and Σ were systematically varied (see Table III for the exact values). With regard to the true mean μ of the logit diagnostic accuracies, four different values were selected representing a random decision, a diagnostic test with overall low accuracy, a test with overall high accuracy, or a test with only high Sen. The alternative of a test with only high Spe was not considered, because its results were assumed to be similar to the high Sen condition due to symmetry. Regarding the between-study variance Σ, which is, among others, assumed to be caused by different cutoff values, three different values were selected representing a fixed effects assumption (in which no between-study heterogeneity is present) or small/large random effects, respectively. The values of Σ in the random effects conditions were chosen to be similar to real data of diagnostic meta-analysis.

Each point sampled from (1) can unambiguously be transformed back to its related Sen and $1 - \mathrm{Spe}$ by taking the inverse of the logit. After the true pair of Sen_i and $1 - \mathrm{Spe}_i$ of study i was sampled, a binomial error was added in order to include the error of measurement:

$$x_i \sim \mathrm{Binom}(\mathrm{Sen}_i, n_{1,i}) \quad \text{and} \quad y_i \sim \mathrm{Binom}(1 - \mathrm{Spe}_i, n_{2,i}). \tag{20}$$

Taken together, these values allowed to fill in the diagnostic 2×2 table (Table I) for each study, and thus, every UM and its respective SE could be calculated on that basis.

3.1. *Introducing publication bias*

In general, PB is modeled in such a way that the probability of a study to be published depends on its effect measure or its p-value [4, 41, 47]. The lower the effect measure (or the higher the p-value), the lower the probability of the study to be published. For the present simulations, two different methods to introduce PB were considered.

The first method was similar to that proposed in literature. It was based on the idea to exclude l studies with the lowest Youden index from the meta-analysis. In order to finally arrive at k studies (more precisely at $k = 30$ or $k = 10$), $k + l$ studies were simulated in a first step. In a second step, those l studies with the lowest Y were excluded. In the following, this procedure is referred to as *selection*. As not only the existence of PB but also its strength was varied, l took on values rounded to the nearest integer of $l = 0.2k$ (small PB) or $l = 0.4k$ (large PB). Contrary to methods proposed in literature, the decision to exclude a study from the meta-analysis depended on the outcomes of all other studies.

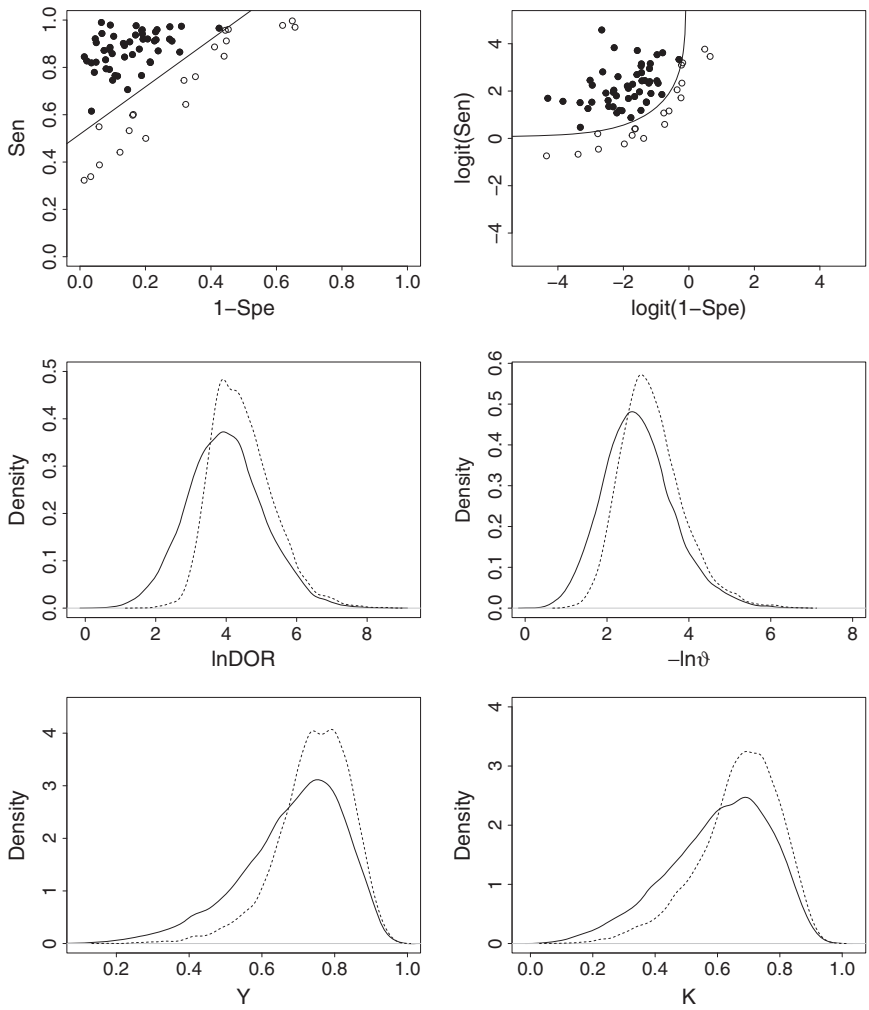

Figure 2. In the first row, 70 simulated studies (under the condition of high accuracy, large random effects, and unbalanced prevalence) are plotted at the ROC and logit ROC space. Solid points symbolize published studies, and white points symbolize unpublished studies. The lines depict the cutoff criterion that decides which studies are treated as unpublished. In the second and third rows, the related smoothed densities are illustrated in the absence of PB (solid line) and its presence (dotted line).

In the second method, no studies were excluded from meta-analysis to introduce PB. Instead, it was assumed that some studies do report systematically higher diagnostic accuracy than other studies. For instance, as developers of diagnostic tests are usually interested in presenting their test in a good light, they may choose certain experimental settings in order to obtain (possibly unrealistic) high diagnostic accuracies. To model this assumption, two-thirds of the k studies were sampled from $\mathrm{N}(\mu, \Sigma)$, whereas one-third was sampled from $\mathrm{N}(\mu + \eta, \Sigma)$. Here, η describes the strength of the PB. In the following, this method is referred to as *mixture*. In the present simulations, η took on values of $\eta = (0.75, -0.75)^{\mathrm{T}}$ (small PB) or $\eta = (1.25, -1.25)^{\mathrm{T}}$ (large PB). Although it may be argued that this kind of systematic heterogeneity can be addressed by adding moderators to the meta-analysis, well-performing tests to detect PB should be able to detect this type of bias. A summary of all varied parameters is provided in Table III.

To obtain an impression on how the UMs are distributed when PB is either existent or not, Figure 2 illustrates the selection method and the resulting smoothed densities. One realizes that the UMs may be asymmetrically distributed even in the absence of PB (the amount of asymmetry dependents on the parameter values in the underlying simulation).

In the present simulations, all combinations of the values for the parameters were taken into account, resulting in 240 unique combinations. For each of these combinations, 10,000 meta-analyses were simulated, so that the accuracy of the results was ensured up to the third decimal place. Each of these meta-analyses was tested for PB with $\alpha = 10\%$ (which was recommended for tests to detect PB; [39, 41, 47]). It was recorded which tests did indicate PB and which tests did not.

4. Results

Although all combinations of UMs and tests to detect PB were simulated, only those findings of central importance are discussed in this section. In a first step it is discussed which UM performs best when combined with the common versions of Egger, Begg, and trim and fill. In a second step, the performance of a broader variety of tests was each combined with the best UM resulting from the first step.

4.1. Comparison of the UMs

In case of random decisions and fixed effects, all common tests combined with each UM nearly held the α level of 10%, irrespective of the other parameters (Figure 3). However, the more μ and Σ differed from zero, the more liberal most tests combined with $\ln \omega$ or $\ln \vartheta$ were (for instance see plots j, k, l, and p in Figure 3). In contrast, one-sided tests combined with Y and K mostly had zero type I error rates and zero power when the diagnostic test was better than a random decision and when random effects were

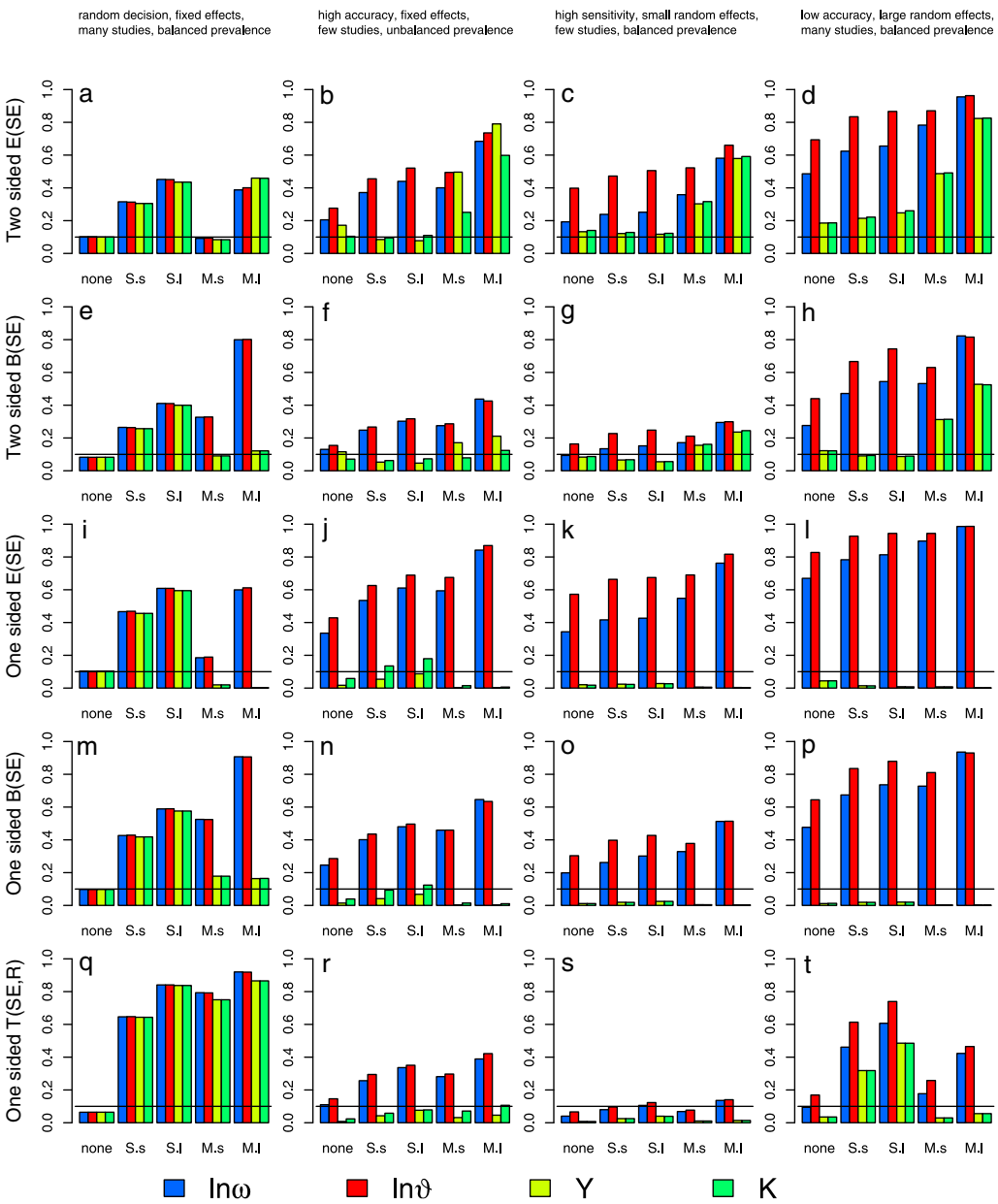

Figure 3. Type I error rates and statistical power for the comparison of the UMs. The conditions pictured in this figure were selected for being most representative. S.s = selection with small PB, S.l = selection with large PB, M.s = mixture with small PB, M.l = mixture with large PB.

present (see third to fifth rows of Figure 3). As this was not found for two-sided tests, further explanations follow in Section 5. For better comparison of the UMs, Figure 3 does also illustrate the results of two-sided tests.

The UMs Y and K were not considered to be adequate for detecting PB for two reasons. First, they cannot reasonably be used with one-sided tests, and second, they rarely performed better than $\ln \omega$ or $\ln \vartheta$ when two-sided tests were applied. Although $\ln \vartheta$ had higher power than $\ln \omega$ when used with E(SE), B(SE), or T(SE,R), it was also more liberal. More precisely, the greater the power difference between $\ln \vartheta$ and $\ln \omega$, the greater the type I error rate difference between $\ln \vartheta$ and $\ln \omega$. Importantly, T($\ln \omega$, SE, R) had non-inflated or only slightly inflated α levels (see fifth row of Figure 3). For these

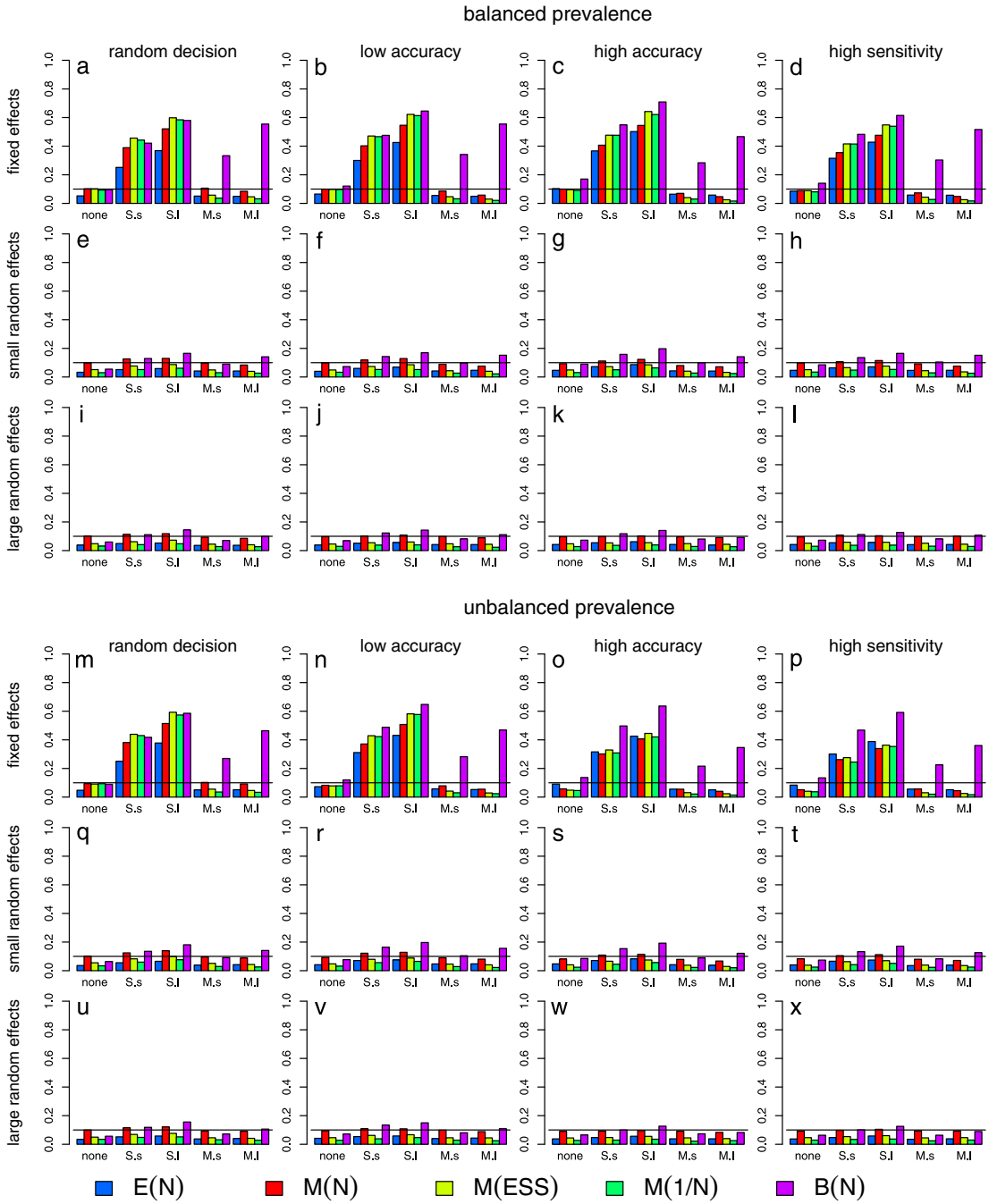

Figure 4. Type I error rates and statistical power in case of $k = 30$ for linear regression and rank correlation tests using N or ESS combined with $\ln \omega$. S.s = selection with small PB, S.l = selection with large PB, M.s = mixture with small PB, M.l = mixture with large PB.

reasons, $\ln \omega$ indeed seemed to the best UM for detecting PB in diagnostic meta-analysis. Therefore, its frequent application for research purposes can be considered as justified.

4.2. *Comparison of the tests*

The findings suggest to concentrate on $\ln \omega$, so that this section only focuses on tests combined with $\ln \omega$. In the following, the UM argument in the short forms is suppressed for notational convenience. As can be seen in Figure 3, E(SE) as well as B(SE) had highly inflated α levels when the diagnostic test was better than a random decision and when random effects were present, especially in case of many studies

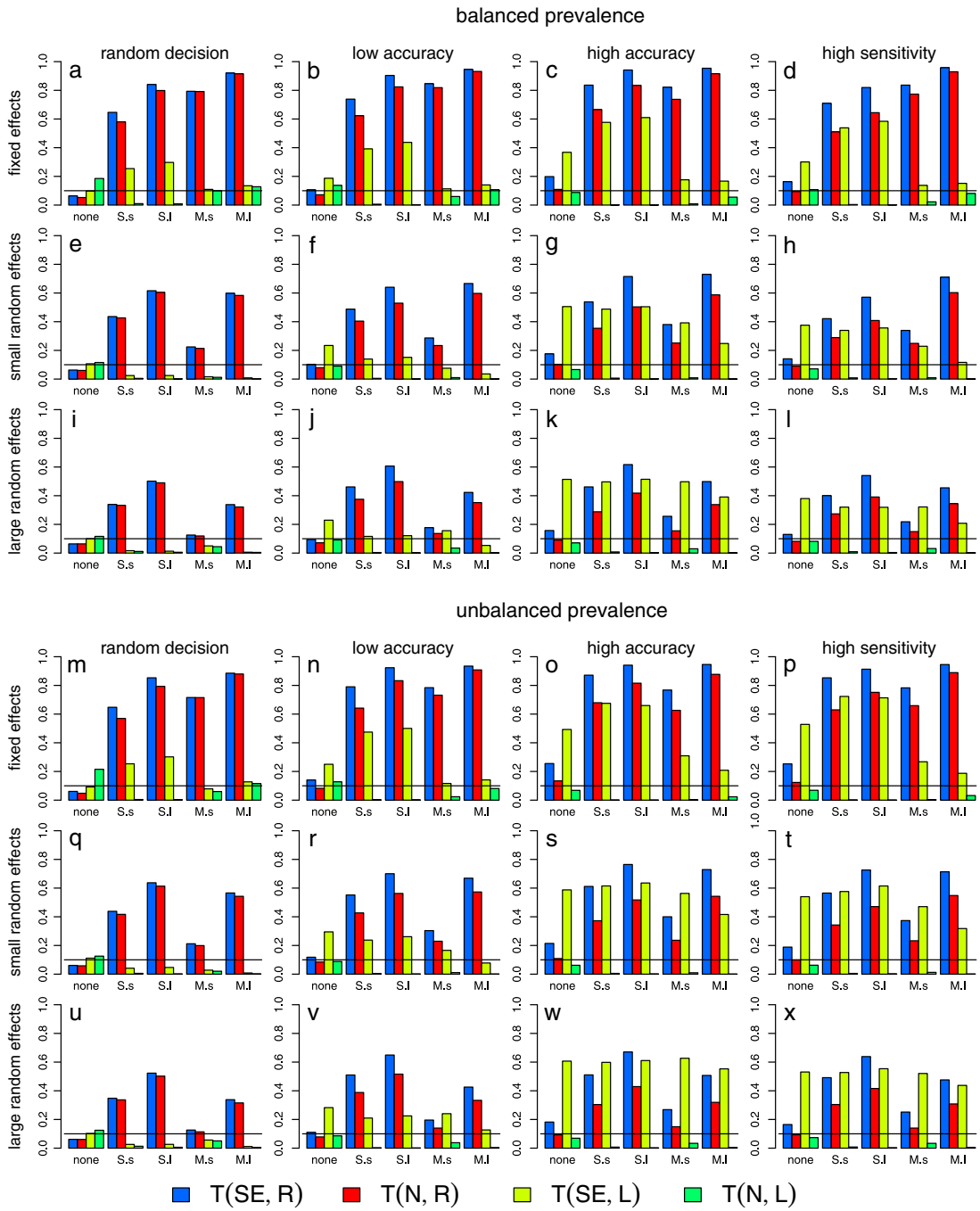

Figure 5. Type I error rates and statistical power in case of $k = 30$ for trim and fill combined with $\ln \omega$. S.s = selection with small PB, S.l = selection with large PB, M.s = mixture with small PB, M.l = mixture with large PB.

included in the meta-analysis. The highly inflated α levels were also found when Egger's regression was weighted by the inverse of the variance. Furthermore, the variation of Egger's regression proposed by Harbord *et al.* [42] and the alternative rank correlation test proposed by Schwarzer *et al.* [49] (both not shown in the figures) were less liberal and much less powerful than E(SE) or B(SE) but still had inflated α levels in case of random effects.

However, tests based on Egger, Macaskill, or Begg that used the total sample size N or similarly used ESS or $1/N$ instead of SE did not suffer from high α inflation. Instead, they appeared more or less

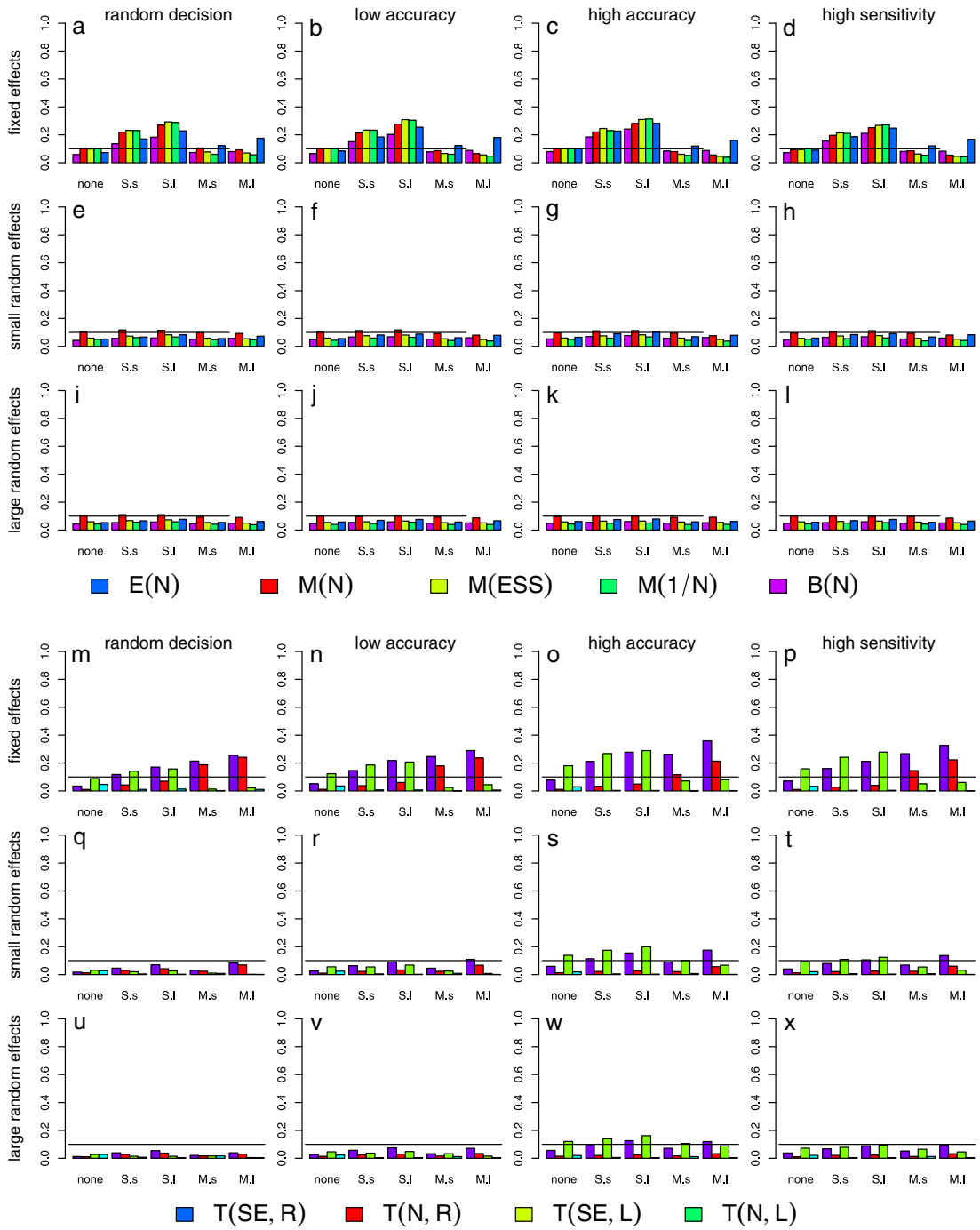

Figure 6. Type I error rates and statistical power in case of $k = 10$ and balanced prevalence for linear regression and rank correlation tests using N or ESS as well as trim and fill each combined with $\ln \omega$. Results for $k = 10$ and unbalanced prevalence are similar. S.s = selection with small PB, S.l = selection with large PB, M.s = mixture with small PB, M.l = mixture with large PB.

conservative even when many studies were included in the meta-analysis (Figure 4). In case of fixed effects (see first and fourth rows of Figure 4), all tests had an acceptable amount of power, when PB was simulated by the selection method. Notably, only rank correlation tests were able to identify PB, which was simulated by the mixture approach. Unfortunately, all of these tests generally had low power in case of random effects (rarely above the nominal α level; see second, third, fifth, and sixth rows of Figure 4). The same was true for the arcsine tests of Rücker *et al.* [43]. The prevalence π did only have small effects (compare first to third with fourth to sixth rows of Figure 4), at least for the non-extreme values of π that were chosen in this study. Taken together, when the investigated diagnostic test is better than a random decision or when random effects such as different cutoff values are present, the methods based on linear regression or rank correlation cannot be recommended for diagnostic meta-analysis, because of inflated α levels or very low power.

In contrast, trim and fill had non-inflated or only slightly inflated α levels and medium to high power (when $k = 30$; Figure 5), at least when the number of missing studies was estimated by (16). Generally, the results of T(SE, R) and T(N, R) were independent of π, and they were able to detect both types of PB (i.e., mixture and selection). T(SE, R) had more power than T(N, R) but was slightly liberal when the diagnostic test had overall high accuracy or at least high Sen (see third and fourth columns of Figure 5), whereas T(N, R) was non-liberal in all cases. Both tests lost power in the presence of random effects. Importantly, however, these losses were much smaller compared with all other non-liberal or slightly liberal tests. When only a few studies were included in the meta-analysis ($k = 10$), trim and fill was rather conservative and had quite low power, which was similar to other tests with non-inflated or only slightly inflated α levels (Figure 6). Interestingly, the estimator of the number of missing studies stated in (17) did not work properly for almost all parameter combinations. Among others, this might be because (17) tends to underestimate the number of missing studies [50].

5. Discussion

The present simulations confirmed that statistical tests based on funnel plots are able to detect PB in diagnostic meta-analysis under diverse conditions. Summarizing the results, trim and fill combined with the log diagnostic odds ratio, more precisely T($\ln \omega$, SE, R) or T($\ln \omega$, N, R), were best to detect PB in diagnostic meta-analysis, although both lacked power when the number of studies per meta-analysis was small. Furthermore, T($\ln \omega$, N, R) was able to hold the nominal α level in all cases and did not have much less power than T($\ln \omega$, SE, R), even in situations with between-study heterogeneity. This finding demonstrates that funnel plots based on the total sample size N (but not on the SE) also provide enough information, so that PB can be detected.

In contrast to trim and fill, the common tests of Egger, Macaskill, and Begg were too liberal in the presence of random effects or they had very low power. Thus, earlier findings were replicated (for instance, see [4, 42, 44]), and it can be concluded that those common tests cannot be recommended for diagnostic meta-analysis. The advantage of the tests proposed by Deeks *et al.* [4], which use ESS, was that those tests were mostly able to hold the nominal α level, but at the expense of low power when between-study heterogeneity was large. As most diagnostic tests are way better than random decisions and as cutoff values often vary between studies, tests of Egger, Macaskill, and Begg will often be misleading. Accordingly, these tests and their numerous existing variations cannot be recommended in diagnostic meta-analysis.

Both the α inflation of certain tests that were combined with $\ln \omega$ and $\ln \vartheta$ and the odd performance of Y and K when combined with one-sided tests need some explanation. First, the distributions of the UMs, especially the ones of Y and K are more symmetric in the presence of PB than in its absence (Figure 2). Second, the SEs of all the four applied UMs are dependent on the underlying effect. With N held constant, SE($\ln \omega$) and SE($\ln \vartheta$) reach higher values, when the underlying effect differs more from zero. For SE(Y) and SE(K), it is contrariwise. For example, consider Egger's regression applied to a meta-analysis of a well-performing diagnostic test in which PB does not exist. If $\ln \omega$ or $\ln \vartheta$ is used, UM $\cdot$ SE^{-1} will be overly constant and greater than zero across different values of SE^{-1}, because the precision becomes higher when the UM is close to zero. Conclusively, the probability of $b_0 > 0$ is increased, which results in a liberal test. In contrast, if Y or K is applied, small values of the UMs will be associated with lower precision, so that the probability of $b_0 < 0$ is increased, which results in a conservative test with very low power (when the hypothesis is that $b_0 > 0$). Trim and fill combined with $\ln \omega$ seems to compensate for these problems. Therefore, we recommend its application in diagnostic meta-analysis.

Besides the fact that a simulation by design can never display reality exactly (and is therefore always slightly wrong), there are some limitations of our modeling assumptions that have to be discussed in the following. First, we decided to directly sample the true logits of Sen and 1−Spe from the random effects model of Reitsma *et al.* [15]. This approach was different from and probably less intuitive than earlier simulation models [4]. However, the bivariate model is a common and valid approach for performing diagnostic meta-analysis [18–20], and therefore, its usage as a sampling model appears reasonable. Second, it might be more realistic in some diagnostic settings to assume a long-tailed distribution for the sample size N of each study instead of a uniform distribution, although this should only have a minor impact on the results. Third, two distinct methods to introduce PB were applied in the present study with one of them being very different from the methods proposed in literature. Interestingly, the mixture approach to introduce PB yielded similar results to the selection approach, although both arose from different theoretical assumptions. Fourth, the PB mechanisms we studied are not exhaustive: A reviewer of the paper pointed out that a selection mechanism based solely on sensitivity would have been a possibility. Fifth, we assumed a perfect gold standard, which seems common in simulations of diagnostic accuracy, but might not be adequate for every diagnostical setting. As the modeling of an imperfect gold standard would have further complicated our simulations and might not have led to very different results, this limitation was considered as acceptable.

Sixth and last, as described in Section 1, there are several reasons other than PB (such as different cutoff values, study quality, or heterogeneity of the examined populations) that may cause funnel plot asymmetry. In the present study, these biases were modeled by the covariance matrix of the true logits of Sen and 1−Spe. As indicated by our results, only trim and fill could discriminate adequately between this heterogeneity and PB. In contrast to PB, no other bias have an explicit *directional* effect on the funnel plot asymmetry in our simulation model. Instead, the effect of the between-study heterogeneity largely depended on the respective applied UM and its SE. However, in real diagnostic meta-analysis, other biases may have directional effects on funnel plot asymmetry as well. Unfortunately, the direction is not always known. In meta-analysis of treatment effects, smaller studies often have a poorer methodological quality, which may lead to an overestimation of the effects in these studies. However, this may not be the case in diagnostic meta-analysis: As larger retrospective studies may obtain their test results from large clinical databases and do thus have more heterogeneous and possibly not fully appropriate samples for the respective research question, they may be of poorer quality in contrast to studies with small but more carefully chosen samples [4]. Depending on possible directions of other biases, the performance of tests to detect PB may vary heavily, which further complicates the detection of PB.

Importantly, however, our results for those tests already investigated in literature turned out to be quite similar to earlier findings [4, 42, 44], which supports our assumptions and also validates our findings.

In summary, the present study provides evidence that trim and fill combined with log diagnostic odds ratios is superior to other combinations of tests and UMs when testing for PB in diagnostic meta-analysis. Moreover, the tests based on linear regression or rank correlation cannot be recommended for diagnostic meta-analysis, because of either highly inflated α levels or very low power.

Acknowledgement

The work of the second author was funded by DFG grant HO 1286/7-2.

References

1. Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, Chalmers TC, Mosteller F. Guidelines for meta-analyses evaluating diagnostic tests. *Annals of Internal Medicine* 1994; **120**(8):667–676.
2. Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. *Journal of Clinical Epidemiology* 1995; **48**(1):119–130.
3. Tacconelli E. Systematic reviews: CRD's guidance for undertaking reviews in health care. *The Lancet Infectious Diseases* 2010; **10**(4):1–281.
4. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *Journal of Clinical Epidemiology* 2005; **58**(9):882–893.
5. Easterbrook PJ, Gopalan R, Berlin J, Matthews DR. Publication bias in clinical research. *The Lancet* 1991; **337**(8746):867–872.
6. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, De Vet H. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Standards for reporting of diagnostic accuracy. *Clinical Chemistry* 2003; **49**(1):68–73.

7. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press: Oxford, 2003.
8. Cheng D, Branscum AJ, Stamey JD. A Bayesian approach to sample size determination for studies designed to evaluate continuous medical tests. *Computational Statistics & Data Analysis* 2010; **54**(2):298–307.
9. Li CR, Liao CT, Liu JP. A non-inferiority test for diagnostic accuracy based on the paired partial areas under ROC curves. *Statistics in Medicine* 2008; **27**(10):1762–1776.
10. Liu JP, Ma MC, Wu Cy, Tai JY. Tests of equivalence and non-inferiority for diagnostic accuracy based on the paired areas under ROC curves. *Statistics in Medicine* 2006; **25**(7):1219–1238.
11. Song F, Khan KS, Dinnes J, Sutton AJ. Asymmetric funnel plots and publication bias in meta-analyses of diagnostic accuracy. *International Journal of Epidemiology* 2002; **31**(1):88–95.
12. Begg CB, Berlin JA. Publication bias: a problem in interpreting medical data. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 1988; **20**(3):419–463.
13. 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.
14. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in Medicine* 2001; **20**(19):2865–2884.
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. 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.
17. Arends L, Hamza T, Van Houwelingen J, Heijenbrok-Kal M, Hunink M, Stijnen T. Bivariate random effects meta-analysis of ROC curves. *Medical Decision Making* 2008; **28**(5):621–638.
18. Macaskill P, Gatsonis C, Deeks JJ, Harbord RM, Takwoingi Y. Chapter 10: analysing and presenting results. In *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy*, Deeks JJ, Bossuyt PM, Gatsonis C (eds), Version 1.0. The Cochrane Collaboration, 2010. Available from: http://srdta.cochrane.org/.
19. 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; **0**(0):1–24.
20. Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM. Systematic reviews of diagnostic test accuracy. *Annals of Internal Medicine* 2008; **149**(12):889–897.
21. Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports a new meta-analytic method. *Medical Decision Making* 1993; **13**(4):313–321.
22. Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in Medicine* 2004; **23**(9):1351–1375.
23. Haddock CK, Rindskopf D, Shadish WR. Using odds ratios as effect sizes for meta-analysis of dichotomous data: a primer on methods and issues. *Psychological Methods* 1998; **3**(3):339–353.
24. Schwarzer G, Antes G, Schumacher M. Inflation of type I error rate in two statistical tests for the detection of publication bias in meta-analyses with binary outcomes. *Statistics in Medicine* 2002; **21**(17):2465–2477.
25. Le CT. A solution for the most basic optimization problem associated with an ROC curve. *Statistical Methods in Medical Research* 2006; **15**(6):571–584.
26. Holling H, Böhning W, Böhning D. Likelihood-based clustering of meta-analytic SROC curves. *Psychometrika* 2012; **77**(1):106–126.
27. Holling H, Böhning W, Böhning D. Meta-analysis of diagnostic studies based upon SROC-curves: a mixed model approach using the Lehmann family. *Statistical Modelling* 2012; **12**(4):347–375.
28. Youden WJ. Index for rating diagnostic tests. *Cancer* 1950; **3**(1):32–35.
29. Böhning D, Böhning W, Holling H. Revisiting Youden's index as a useful measure of the misclassification error in meta-analysis of diagnostic studies. *Statistical Methods in Medical Research* 2008; **17**(6):543–554.
30. Böhning D, Holling H, Patilea V. A limitation of the diagnostic-odds ratio in determining an optimal cut-off value for a continuous diagnostic test. *Statistical Methods in Medical Research* 2011; **20**(5):541–550.
31. Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 1960; **20**(1):37–46.
32. Kraemer HC, Periyakoil V, Noda A. Kappa coefficients in medical research. *Statistics in Medicine* 2002; **21**(14):2109–2129.
33. Bloch DA. Comparing two diagnostic tests against the same "gold standard" in the same sample. *Biometrics* 1997; **53**(1):73–85.
34. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 1971; **76**(5):378–382.
35. Fleiss JL, Cohen J, Everitt BS. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin* 1969; **72**(5):323–327.
36. Richard J, Pillemer DB. *Summing Up: The Science of Reviewing Research*. Harvard University Press: Cambridge, Massachusetts, 1984.
37. Rothstein HR, Sutton AJ, Borenstein M. *Publication Bias in Meta-analysis: Prevention, Assessment and Adjustments*. Wiley: Chichester, UK, 2006.
38. Terrin N, Schmid CH, Lau J. In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. *Journal of Clinical Epidemiology* 2005; **58**(9):894–901.
39. Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997; **315**(7109):629–634.
40. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in Medicine* 1999; **18**(20):2693–2708.

41. Macaskill P, Walter SD, Irwig L. A comparison of methods to detect publication bias in meta-analysis. *Statistics in Medicine* 2001; **20**(4):641–654.
42. Harbord RM, Egger M, Sterne JA. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Statistics in Medicine* 2006; **25**(20):3443–3457.
43. Rücker G, Schwarzer G, Carpenter J. Arcsine test for publication bias in meta-analyses with binary outcomes. *Statistics in Medicine* 2008; **27**(5):746–763.
44. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Comparison of two methods to detect publication bias in meta-analysis. *JAMA: the journal of the American Medical Association* 2006; **295**(6):676–680.
45. Peters J, Sutton A, Jones D, Abrams K, Rushton L. Performance of tests and adjustments for publication bias in the presence of heterogeneity. *Technical Report 05-01*, Department of Health Sciences, University of Leicester, Leicester, England, 2005.
46. Sterne JA, Gavaghan D, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology* 2000; **53**(11):1119–1129.
47. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. *Biometrics* 1994; **50**(4):1088–1101.
48. Kendall MG. A new measure of rank correlation. *Biometrika* 1938; **30**(1/2):81–93.
49. Schwarzer G, Antes G, Schumacher M. A test for publication bias in meta-analysis with sparse binary data. *Statistics in Medicine* 2007; **26**(4):721–733.
50. Duval S, Tweedie R. Trim and fill: a simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics* 2000; **56**(2):455–463.
51. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. *Statistics in Medicine* 2007; **26**(25):4544–4562.
52. Dear KB, Begg CB. An approach for assessing publication bias prior to performing a meta-analysis. *Statistical Science* 1992; **7**(2):237–245.
53. Givens GH, Smith D, Tweedie R. Publication bias in meta-analysis: a Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate. *Statistical Science* 1997; **12**(4):221–240.
54. Hedges LV. Modeling publication selection effects in meta-analysis. *Statistical Science* 1992; **7**(2):246–255.
55. R Core Team. R: a language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 2013. http://www.R-project.org/ [Accessed on 6 December 2013], ISBN 3-900051-07-0.
56. Eddelbuettel D, François R. Rcpp: seamless R and C++ integration. *Journal of Statistical Software* 2011; **40**(8):1–18.
57. Schwarzer G. meta: meta-analysis with R, 2013. http://CRAN.R-project.org/package=meta [Accessed on 6 December 2013], r package version 3.1-2.