Journal of Clinical Epidemiology 58 (2005) 882–893

Journal of Clinical Epidemiology

ORIGINAL ARTICLES

The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed

Jonathan J. Deeks[1],*, Petra Macaskill, Les Irwig

Screening and Test Evaluation Program, School of Public Health, University of Sydney, Sydney, New South Wales 2006, Australia

Accepted 17 January 2005

Abstract

Background and Objective: Publication bias and other sample size effects are issues for meta-analyses of test accuracy, as for randomized trials. We investigate limitations of standard funnel plots and tests when applied to meta-analyses of test accuracy and look for improved methods.

Methods: Type I and type II error rates for existing and alternative tests of sample size effects were estimated and compared in simulated meta-analyses of test accuracy.

Results: Type I error rates for the Begg, Egger, and Macaskill tests are inflated for typical diagnostic odds ratios (DOR), when disease prevalence differs from 50% and when thresholds favor sensitivity over specificity or vice versa. Regression and correlation tests based on functions of effective sample size are valid, if occasionally conservative, tests for sample size effects. Empirical evidence suggests that they have adequate power to be useful tests. When DORs are heterogeneous, however, all tests of funnel plot asymmetry have low power.

Conclusion: Existing tests that use standard errors of odds ratios are likely to be seriously misleading if applied to meta-analyses of test accuracy. The effective sample size funnel plot and associated regression test of asymmetry should be used to detect publication bias and other sample size related effects. © 2005 Elsevier Inc. All rights reserved.

Keywords: Publication bias; Diagnostic test accuracy; Funnel plots; Systematic reviews; Meta-analyses; Sensitivity; Specificity

1. Introduction

The validity of a systematic review depends on minimizing bias in the identification of studies. If the studies that are included in a review have results that systematically differ from relevant studies that are missed, then the findings will be compromised by publication bias [1,2]. Systematic reviewers are therefore advised to use comprehensive searches to attempt to locate all relevant studies [3–5].

In stark contrast to the substantial literature and empirical evidence available for randomized controlled trials [1,6–11], there has been little research into the determinants, magnitude, and impact of publication bias for studies of diagnostic test accuracy. Recently, funnel plot analyses developed for investigating publication bias in randomized trials have been recommended [12] and used for reviews of test accuracy [13,14]. Evidence that the performance of these tests deteriorates as odds ratios increase raises concern that they may not be appropriate [15–17].

Determinants of publication bias are likely to be different for investigations of test accuracy. The analysis of a study of test accuracy typically involves computation of estimates of sensitivity and specificity (or possibly likelihood ratios), together with 95% confidence intervals [18]. In contrast to reporting of randomized trials, there is no stated null hypothesis or computation of an associated *P*-value. Thus, publication bias is unlikely to be associated with statistical nonsignificance.

Funnel plots can detect any effect that is related to sample size. Publication bias is the most commonly cited sample-size-related effect, but other factors such as study quality or the type of population may also be related to sample size. Here we explore theoretical issues that underpin the investigation of any sample size effect for diagnostic tests and develop funnel plots that are appropriate for reviews of test accuracy. Section 2 reviews existing tests for funnel plot asymmetry and considers how their performance is likely to be affected by characteristics typical of studies of test accuracy. Section 3 introduces a new funnel plot and tests for asymmetry that we apply, together with existing tests, to a case study in section 4. Through simulation, described in sections 5 and 6, we evaluate the performance of new and existing funnel plot–based tests for detecting publication bias, and estimate

* Corresponding author. Tel.: +44-(0)1865-284403; fax: +44-(0)1865-284424.

[1] Present address: Centre for Statistics in Medicine, Wolfson College Annex, Linton Road, Oxford OX2 6UD, UK.

E-mail address: Jon.Deeks@cancer.org.uk (J.J. Deeks).

0895-4356/05/$ – see front matter © 2005 Elsevier Inc. All rights reserved.
doi: 10.1016/j.jclinepi.2005.01.016

the impact of publication bias on estimates of diagnostic accuracy. We base our investigations on the assumption that the probability of publication decreases with lower values of diagnostic accuracy, and investigate the impact of four possible selective publication mechanisms.

2. Theory and methods

2.1. Detection of publication bias and other sample size effects using funnel plots

The funnel plot has been recommended as a graphical device for investigating the possibility of publication bias or other sample size effects for reviews of randomized controlled trials [19]. By plotting estimates of study findings, usually the log odds ratio (lnOR), against their sample size or precision (estimated by the reciprocal of the standard error), indirect evidence for bias can be discerned from the shape of the plot. In the absence of a sample size effect, the points will form a symmetrical funnel shape around the overall estimate of effect, points from small or low-precision studies being more dispersed around the estimate of overall effect than points from large or high-precision studies. Nonpublication of small nonsignificant studies will cause a gap in the plot and introduce asymmetry if there is a treatment effect. Asymmetry may result from publication bias, but can also be caused by other so-called sample size effects, such as clinical heterogeneity and variation in study quality if they are also linked to sample size [20]. Various statistical tests, notably Begg's rank correlation [21], Egger's regression test [20], and Macaskill's regression test [16], have been devised to objectively assess asymmetry. If a funnel plot is asymmetric, it can be deduced that some mechanism that links study results with sample size is present—but identifying the mechanism is not straightforward.

Song et al. [12] proposed that the funnel plot can also be used for reviews of diagnostic test accuracy; they produced funnel plots of log diagnostic odds ratio (lnDOR) against standard errors for 28 meta-analyses and applied the Begg and Egger tests for asymmetry. Depending on the criteria used, between 6 and 12 of these meta-analyses demonstrated significant funnel plot asymmetry. Meta-analyses that included fewer studies and searched fewer databases were more likely to have asymmetrical plots.

2.2. Choice of horizontal axis for funnel plots of diagnostic test accuracy

Various funnel plots can be constructed for dichotomous data in a meta-analysis determined by the choice of the measure of effect and measure of precision [22,23]. Sterne and Egger [22] showed that plotting the lnOR against its standard error is optimal for meta-analyses of trials, because the expected funnel shape would be pyramidal rather than curvilinear and use of odds ratios or risk ratios minimizes unexplained heterogeneity.

For diagnostic test reviews, the DOR summarizes test accuracy as a single number and it is used routinely in summary receiver operating characteristic (ROC) meta-analyses [24,25]. Separate funnel plots for sensitivity and specificity (after logit transformation) are unlikely to be helpful for detecting sample size effects, because sensitivities and specificities will vary due to both variability of threshold between the studies and random variability. Simultaneous interpretation of two related funnel plots and two tests for funnel plot asymmetry also presents challenges. Hence, we restrict our investigation to funnel plots based on the lnDOR.

2.3. Existing tests for sample size effects

The performance of a statistical test is based on assessing both type I and type II error rates. In the present context, a type I error occurs when the test result is statistically significant but there is no sample size related effect. Type I errors should occur with the same probability as the *P*-value that defines statistical significance. Type I error rates that are lower give overly conservative hypothesis tests; those that are higher lead to false claims of sample size effects. Type II errors occur when the test is not statistically significant despite existence of a sample size effect. The lower the type II error rate, the higher the statistical power to detect sample size effects. Tests which have high power are preferred, provided their type I error rates are not inflated.

Begg and Mazumdar [21] proposed a test for publication bias based on assessing the significance of the correlation between the ranks of effect estimates and the ranks of their variances. The test involves standardizing the effect estimates to stabilize the variances and performing an adjusted rank correlation test based on Kendall's τ. It has been shown to have low power and a conservative type I error rate when used for dichotomous outcome data [15,16].

Egger et al. [20] proposed a test for funnel plot asymmetry based on a regression of standardized effect estimates against precision (standard error, or SE), to test whether the intercept deviates from zero. Sterne et al. [15] showed that the significance of the intercept in this model is equivalent to the significance of the slope of a simpler inverse variance-weighted regression of observed effect sizes against standard error, and demonstrated that Egger's approach may be more powerful than the Begg test for detecting publication bias.

Irwig et al. [26], however, expressed concern that Egger's regression approach is likely to be biased as the predictor (SE) in the regression model is measured with error, and Macaskill, Walter, and Irwig [16] later demonstrated by simulation the existence of a correlation that inappropriately inflated type I error rates when the OR differed from one. Macaskill et al. [16] proposed using study sample size (N) as a predictor variable in the inverse variance-weighted regression approach, and showed that it gave a more appropriate, if conservative, type I error rate. They also noted that computing regression weights as the inverse variance of the average prevalence (pooling samples and events across the two groups) gave appropriate type I error rates when treatment effects were large, but that all approaches that use total sample size as the explanatory variable have low statistical power.

The published evaluations of these three tests [15–17,21,22] have concentrated on randomized controlled trial scenarios where studies have equal numbers of treated and control participants and treatment effects are small (odds ratios are close to 1).

2.4. *Issues in applying existing tests to diagnostic accuracy meta-analyses*

There are three particular characteristics of studies of diagnostic test accuracy that can result in asymmetry for funnel plots that use standard error of the lnDOR or total sample size as a measure of precision, in the absence of a true underlying sample size related effect.

1. Values of DOR are typically very high, with the numbers of false positives or false negatives, or both, quite often being small. The asymptotic standard error is a biased estimate of the true standard error, with larger bias for smaller cell sizes, as occurs with larger DORs and smaller studies [27].
2. The standard error of the lnDOR depends on the proportion that is test positive. Individual studies of test evaluations often differ (either explicitly or implicitly) in the diagnostic threshold used to define test positives, leading to variability in the proportion that are test positive between studies.
3. Diagnostic studies commonly have unequal sample sizes in diseased and nondiseased groups, depending on (a) whether they use a case-control or clinical cohort design and (b) the prevalence of disease in the sample. Unequal numbers of nondiseased (n_1) and diseased (n_2) will reduce the precision of an estimate of test accuracy for a given total sample size. Sample size related precision when there are unequal group sizes is more appropriately summarized by the effective sample size ESS, where ESS $= (4n_1n_2)/(n_1 + n_2)$.

The algebraic relationship between the standard error of the lnDOR, effective sample size, proportion test positive and the estimated DOR is expounded in Appendix A. The Begg, Egger, and Macaskill tests all depend in some way on the standard error of lnDOR, and Macaskill's test also depends on total sample size.

3. A robust funnel plot and test for asymmetry suitable for use with meta-analyses of diagnostic test accuracy

A funnel plot for studies of diagnostic test accuracy should not display asymmetry if variation in the magnitude of the DOR is due solely to sampling error and/or there is variation in test thresholds. In Appendix A, we show that the SE of the lnDOR does not fulfill these criteria. The only term to behave appropriately was the sample size dependent term,

$$\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

or $(1/n_1 + 1/n_2)^{1/2}$, which is equal to 2/ESS$^{1/2}$. Consequently, we propose that funnel plots for diagnostic test accuracy plot the lnDOR against 1/ESS$^{1/2}$—or, equivalently, against $(1/n_1 + 1/n_2)^{1/2}$, which is proportional to 1/ESS$^{1/2}$.

Two obvious alternative tests for asymmetry follow as (a) an adaptation of Begg's rank correlation test, substituting 1/ESS for the variance of the log odds ratio; (b) a regression of lnDOR against 1/ESS$^{1/2}$, weighting by ESS. In section 5, we evaluate the performance of these new tests and the existing tests by simulation. First, however, we consider the application of the tests in a case study.

4. Case study

Kearon et al. [28] reviewed the diagnostic accuracy of noninvasive tests for detecting deep vein thrombosis. They located 14 suitable studies comparing venous ultrasonography in asymptomatic patients with venography (the reference standard).

Three alternative funnel plots are presented in Fig. 1 plotting lnDOR against (a) the standard error of the lnDOR, (b) the total sample size, and (c) the inverse of the square root of the effective sample size. For computation of the standard error, addition of 0.5 was made to all cell counts for all studies to avoid division by zero errors.

Regression lines are plotted as obtained from the Egger [E(SE)], Macaskill [M(N)], and the proposed effective sample size regression test [D(ESS)], respectively. Significance tests indicate that asymmetry is evident in the standard error plot (P = .006), borderline in the sample size plot (P = .10) and not present in the effective sample size plot (P = .89). Notably, the trend towards less precise studies giving higher values of diagnostic test accuracy evident in the Egger plot is reversed in the subsequent two plots.

Examination of the points reveals that the locations of studies 4 and 13 change between the three plots. These two studies have the highest estimated DOR (318 and 1,138) and standard errors considerably larger than the other 12 trials (1.48 and 1.52, the next largest being 0.98). These two points are influential in the E(SE) test of asymmetry; when they are deleted, the test is of only borderline significance (P = .09).

Plots of lnDOR against N and 1/ESS$^{1/2}$ reveal that, although these studies have the highest standard errors, their total sample sizes are above the median (ranked 4th and 6th out of 14) and effective sample sizes are ranked 6th and 9th. These changes in ranking render the regression tests nonsignificant or of only borderline significance. The high

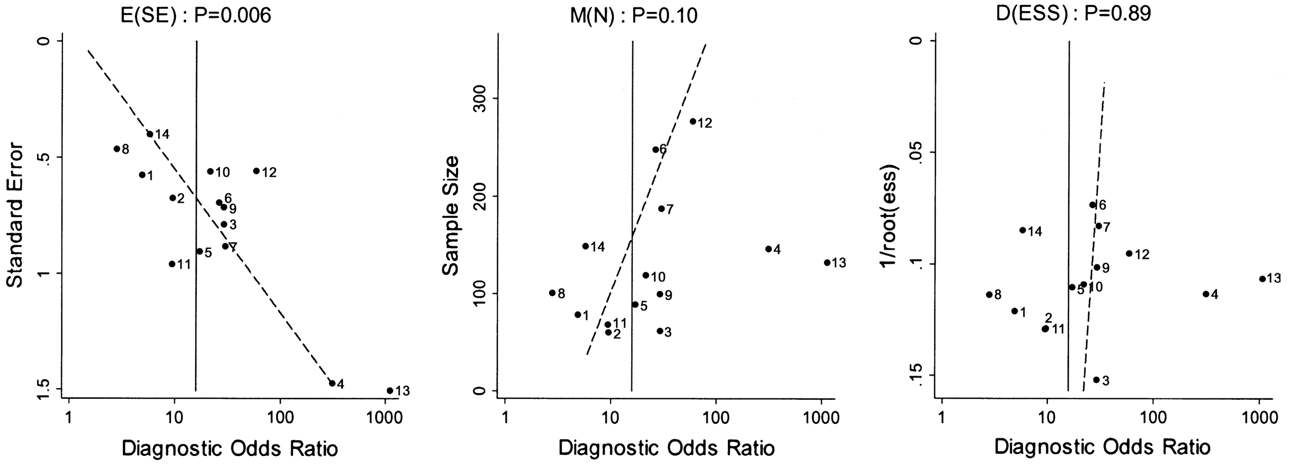

Fig. 1. Funnel plots for a meta-analysis of venous ultrasonography to detect deep vein thrombosis in asymptomatic patients [28] using Egger E(SE), Macaskill M(N), and the effective sample size D(ESS) weighted regression tests of funnel plot asymmetry.

standard errors of the lnDOR for studies 4 and 13 in the presence of average sample sizes are likely to be due to these two studies having (a) the highest observed specificity (100%), indicative of a high test threshold, and (b) the highest DOR.

Thus, the results of the Egger test may be explained by the estimates of standard error being overly influenced by the extreme diagnostic threshold and high test accuracy. Whether the nonsignificant result of the effective sample size regression test D(ESS) is likely to be a correct finding depends on the power of the test, which is evaluated in the simulation studies described in sections 5 and 6.

5. Evaluation by simulation

Because one of the best-known sample size effects is publication bias, we evaluated the performance of existing tests and the proposed new tests for sample size related effects through simulating meta-analyses of diagnostic tests with and without publication bias.

Simulations were undertaken in Stata version 8 (StataCorp, College Station, TX, USA). Each data set contained results from 20 studies ($k = 20$). Sample sizes were redefined for each simulation and varied between $n = 20$ and $n = 2{,}000$ (randomly sampled from a uniform distribution).

Using an underlying prevalence p, each study was randomly divided into diseased and nondiseased groups and a value of a continuous diagnostic measure, θ, randomly sampled for each individual from logistic distributions as shown in Fig. 2, with means and standard deviations of μ_1 and σ_1 for nondiseased and μ_2 and σ_2 for diseased (where $\mu_2 \geq \mu_1$). A diagnostic threshold t was defined for each study and test results declared positive if $\theta > t$ and negative if $\theta \leq t$. Participants in each study were classified as having true positive, false negative, false positive and true negative diagnoses as indicated in Fig. 2, with the DOR, sensitivity, and specificity computed as explained in Appendix B.

5.1. Parameters varied in simulation

The base scenario considered distributions for diseased and nondiseased created using the same standard deviation ($\sigma_1 = \sigma_2 = \sigma$), and fixing the threshold parameter halfway between the means of the distributions—that is, with $t = (\mu_1 + \mu_2)/2$—such that sensitivity = specificity. The prevalence p was set at 0.5. From this base scenario, variations to parameter values were made as follows.

1. The threshold t was increased in steps of $0.5\sigma_1$ from the average of the means up to $2\sigma_1$ above the average of the means.
2. The threshold t was randomly chosen from a uniform distribution for each study with ranges of between $0.5\sigma_1$ and $2\sigma_1$. Symmetry of the threshold around the average of the means was relaxed.
3. Prevalence p of disease took values of 50%, 40%, 30%, 20%, 10%, and 5%.
4. The prevalence p was randomly chosen from a uniform distribution for each study. Values were chosen from the ranges 40%–50%, 30%–50%, 20%–50%, 10%–50%, and 5%–50%.
5. Heterogeneity in diagnostic accuracy was introduced by adding a value τ to the difference between the means $\mu_2 - \mu_1$ for each study. The value of τ was sampled from a normal distribution with zero mean and standard deviation $0.1\sigma_1$, $0.2\sigma_1$, or $0.3\sigma_1$.
6. The variability of the diagnostic measure in the diseased was increased to $2\sigma_1$ introducing asymmetry into the shape of the ROC curve.

Results are reported only for DOR of 1, 38, and 231, and only for the selection of the parameter combinations necessary to demonstrate key findings. Uniform distributions were used to introduce random specifications for design features (threshold, prevalence, sample size) for which the investigator has control, and normal distributions used otherwise (for heterogeneity in diagnostic accuracy).

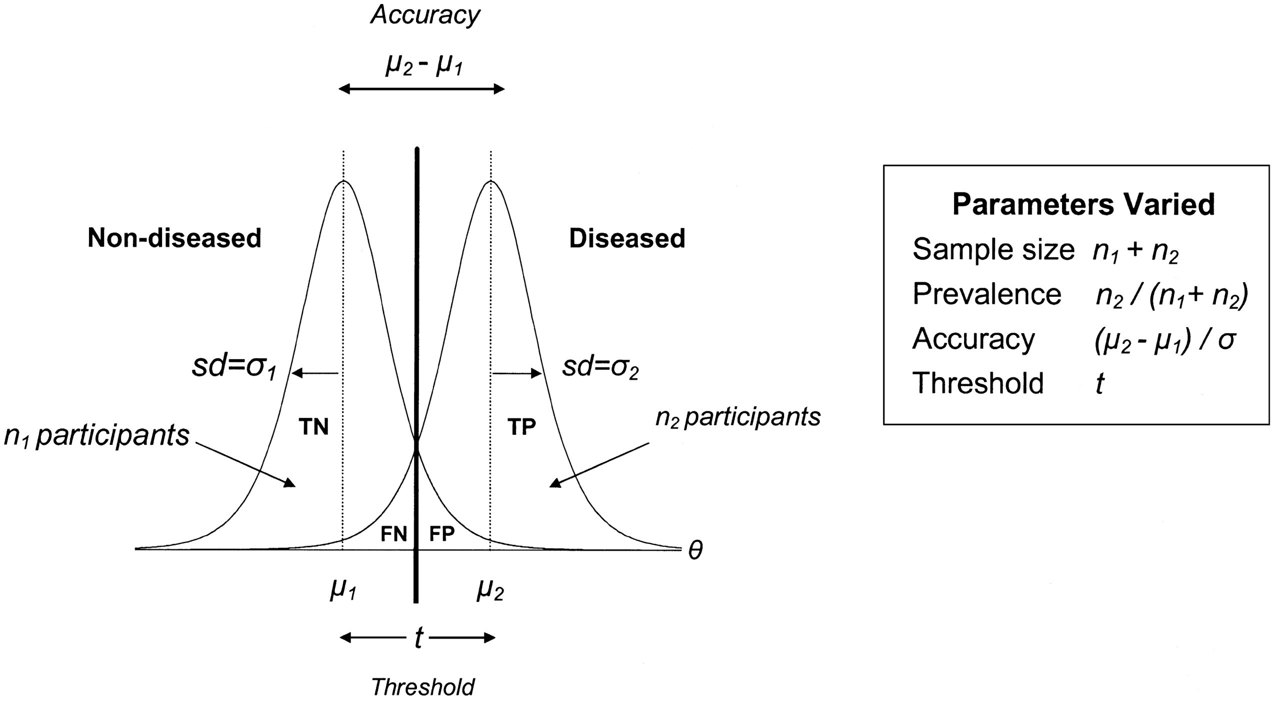

Fig. 2. Underlying bilogistic distribution model used in the simulations. *Abbreviations:* FN, false negative; FP, false positive; TN, true negative; TP, true positive.

5.2. Methods used to introduce publication bias

A one-sided censoring mechanism, adapted from the function used by Begg and Mazumdar [21], was used to introduce differing degrees of publication bias. The probability of selection of a study for inclusion in a meta-analysis is given by a weight function $w(\lambda_i) = \exp[-\beta(1 - \lambda_i)^{\alpha}]$, where λ_i is a measure of diagnostic accuracy (Fig. 3). Studies were included if a random number drawn from a [0,1) uniform distribution was less than $w(\lambda_i)$. Studies continued to be sampled until 20 studies had been included in each meta-analysis, the number censored in the process being noted.

Four alternative measures were considered for λ_i: sensitivity, specificity, square root of Youden's index (sensitivity + specificity − 1), and the square of the area under the ROC curve (AUC). Transformations for the last two parameters were chosen empirically to achieve similar proportions being censored as with comparable values of sensitivity and specificity. The AUC was used as a measure of overall diagnostic accuracy, rather than the DOR,

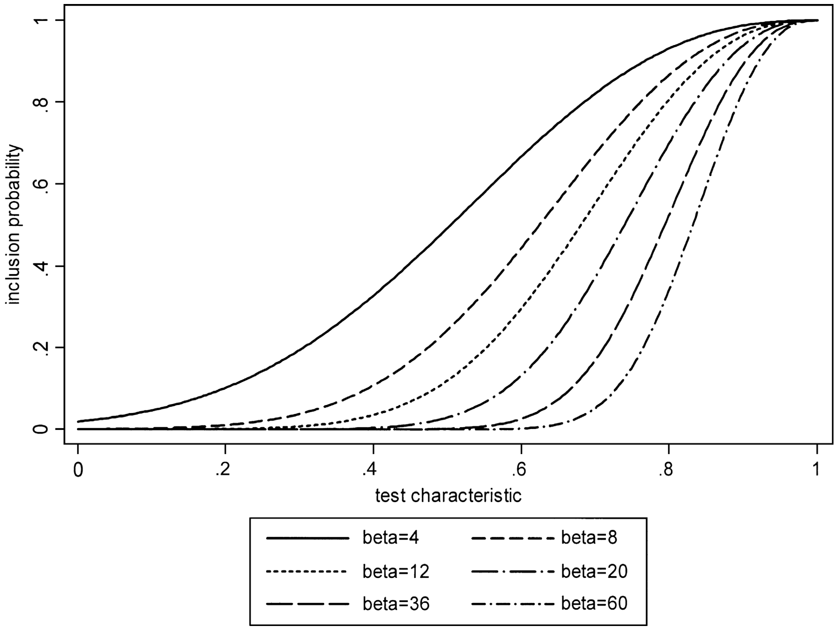

Fig. 3. Publication bias censoring functions based on the Begg and Mazumdar weight function [21].

because of the convenience of it taking values between 0.5 and 1, similar to the values of the other three measures of diagnostic accuracy. The value α was fixed at 2.5. Values of β that censored 10%, 25%, and 50% for each of the four accuracy parameters were identified empirically.

5.3. Number of simulations

To give adequate precision for estimating empirical type I error rates, 10,000 simulations were undertaken for each combination of parameters The standard errors for estimates of event rates of 2.5%, 5%, and 10% are 0.16%, 0.22%, and 0.30%, respectively.

5.4. Assessment of the impact of publication bias and performance of tests for funnel plot asymmetry

Type I error rates were assessed from simulations without study censoring for the following five tests of funnel plot asymmetry:

1. B(SE), rank correlation of lnDOR with var(lnDOR) (Begg and Mazumdar [21]);
2. E(SE), regression of lnDOR with SE(lnDOR) weighted by inverse variance lnDOR (Egger et al. [20]);
3. M(N), regression of lnDOR with *n* weighted by inverse variance lnDOR (Macaskill et al. [16]);
4. D(ESS), regression of lnDOR with $1/\text{ESS}^{1/2}$ weighted by effective sample size; and
5. B/D(ESS), rank correlation of lnDOR with 1/ESS.

For diagnostic accuracy reviews, we would expect the probability of publication to be higher for higher diagnostic accuracies. Thus, we have presented the performance of tests for asymmetry using one-sided 2.5% and 5% significance tests, as well as the more conventional two-sided 5% and 10% tests.

Type I error rates were estimated in simulations where no censoring was present. The proportions statistically significant at 2.5% and 5% levels in each tail were compared with nominal 2.5% and 5% significance levels. Statistical power was measured in simulations where censoring did occur. The proportion of tests for funnel plot asymmetry significant at 5% and 10% two-tailed levels were noted.

Two approaches were used for meta-analysis. Separate estimates of sensitivity and specificity were obtained by computing weighted averages of logit sensitivity and logit specificity using inverse variance weighting. An estimate of the average DOR was obtained from the Moses SROC regression model [24]. An unweighted analysis was used, as recommended by Irwig et al. [29], and was noted to give estimates close to those predicted from the chosen parameter values when no publication bias was present. The impact of publication bias was assessed by comparing results of meta-analyses where censoring did occur with results without study censoring, as well as with the theoretical result expected from the specified parameter values. The impact of publication bias on statistical power was assessed in simulations with parameter values chosen to be characteristic of a meta-analysis with underlying variation in diagnostic threshold. Studies were generated with an average DOR of 38 ($\mu_2 - \mu_1 = 2\sigma$), with diagnostic thresholds varying uniformly over 2σ between μ_1 and μ_2, with the proportion diseased varying uniformly between 10% and 50% and the variance of the diagnostic marker equal in diseased and nondiseased. Simulations were first undertaken with no heterogeneity in test accuracy, but then with increasing degrees of heterogeneity generated by introducing a random effect with standard deviations up to 0.3σ.

6. Results

6.1. Type I error rates

Empirical type I error rates for the base scenario and a selection of parameter combinations are shown in Fig. 4. In the base scenario with a DOR of one, a diagnostic threshold set so that sensitivity = specificity, and with equal numbers of diseased and nondiseased (Fig. 4, row 1, column 1), all tests achieve empirical type I error rates close to the nominal 2.5% and 5% values in both tails, although rates for the rank correlation tests B(SE) and B/D(ESS) are a little low. The percentage significant at the two-tailed 5% (10%) levels are: E(SE) 4.7% (10.0%), M(N) 5.3% (10.0%), D(ESS) 4.8% (10.1%), B(SE) 3.8% (8.9%), and B/D(ESS) 3.7% (8.9%).

6.2. Impact of increasing diagnostic accuracy

Increasing diagnostic accuracy adversely affected the performance of B(SE), E(SE), and M(N), but had little impact on the D(ESS) and B/D(ESS) tests (Fig. 4, column 1). Two-tailed type I error rates were reasonable for all tests, but type I error rates for one-tailed tests were not. At a DOR of 231, the proportions in the left tail at the 2.5% (5%) levels were nearly twice their nominal level: (E(SE) 5.3% (10.2%), M(N) 5.5% (9.9%), and B(SE) 4.8% (9.5%). The ESS-based regression test D(ESS) had more appropriate proportions significant of 2.1% (4.2%); the Begg ESS-based test B/D(ESS) behaved conservatively, with only 1.5% (3.6%) significant.

Where variation in other simulation parameters introduced poor performance for particular tests, the problems almost always were magnified with increasing diagnostic accuracy.

6.3. Impact of threshold selection

Type I error rates were increased for E(SE), M(N), and B(SE) when the threshold differed from the sensitivity = specificity value and when the DOR was greater than 1. The values in Fig. 4, column 2 are based on positioning the threshold a distance σ_1 above the average of the means

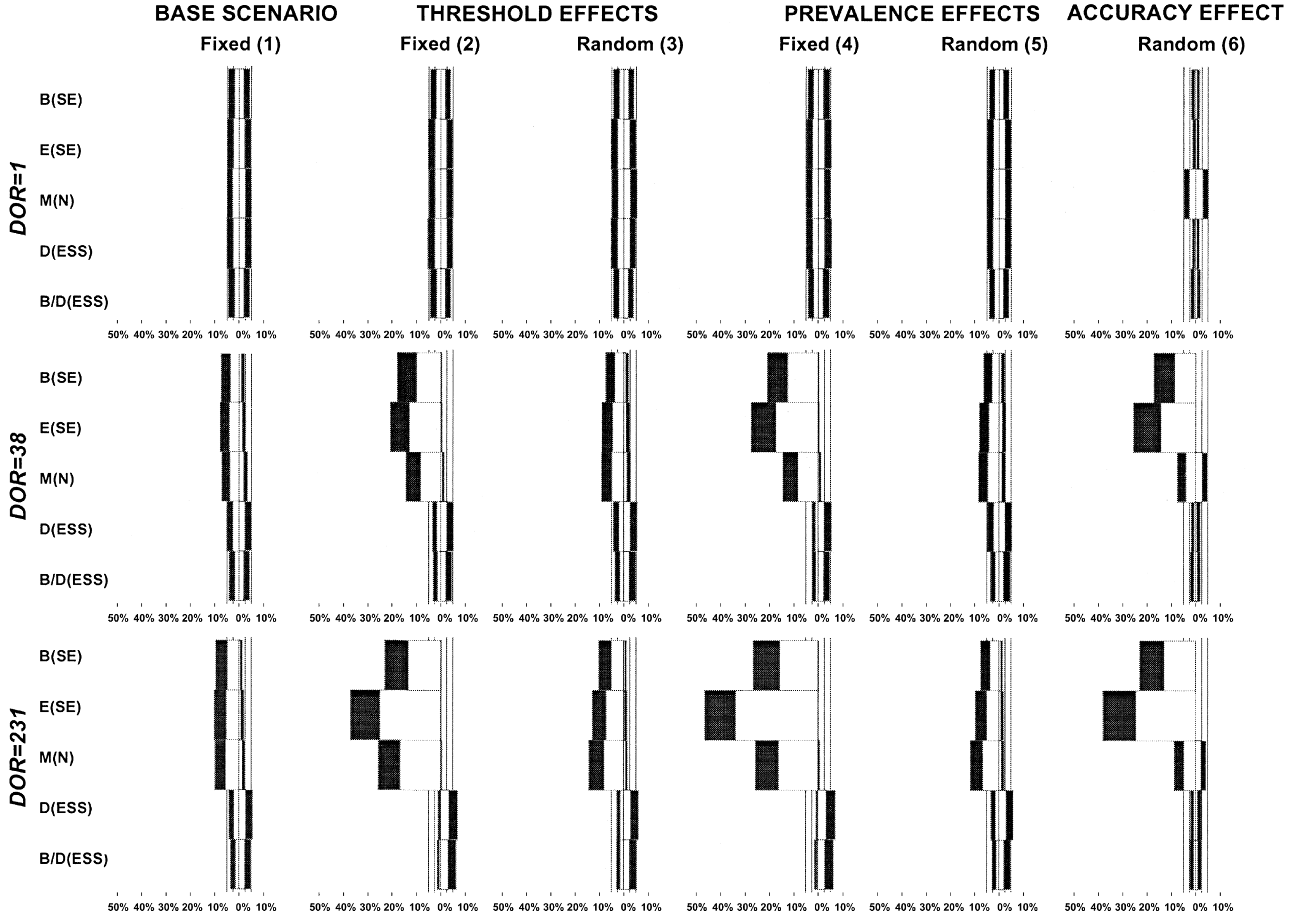

Fig. 4. Empirical type I error rates for five tests of funnel plot asymmetry: B(SE), E(SE), M(N), D(ESS), and B/D(ESS). White boxes indicate 2.5% tails; black boxes indicate 5% tails; vertical lines indicate nominal positions of tails. The five tests are explained in section 5.4.

of the distributions, equivalent to fixing sensitivities at 50% and 71% and specificities at 97% and 99%, for DORs of 38 and 231, respectively.

Proportions significant at the two-tailed 5% (10%) type I error rates for Egger's test E(SE) were highly inflated to 13.1% (21.4%) and 25.0% (37.2%) for DORs of 38 and 231. In contrast, the ESS-based regression test D(ESS) maintained 5% (10%) type I error rates of 4.1% (8.8%) and 3.9% (8.1%), close to the nominal values, although the distribution of results became slightly asymmetric in the two tails as DORs increased.

Values in Fig. 4, column 3, display empirical type I error rates for simulations where a different threshold was selected for each study from a range between the average of the means and σ_1 above the average of the means. The same pattern of inappropriately increased type I error rates for E(SE), B(SE), and M(N) was evident, although error rates are of smaller magnitude: the average threshold is only $\sigma_1/2$ above the average of the means.

6.4. Impact of disease prevalence

Decreasing the proportion of diseased in the sample caused problems for all non-ESS-based methods, with exceptionally high and asymmetric empirical type I probabilities (Fig. 4, column 4). The problems increased when the percentage diseased decreased and DOR increased. When only 5% of study participants were diseased, proportions significant at the 5% (10%) type I error rates for a DOR of 231 were 33.9% (46.5%) for E(SE), 16.5% (26.3%) for M(N), and 15.8% (26.7%) for B(SE). Although the effective sample size tests had more appropriate type I error rates of 3.6% (7.8%) for D(ESS) and 3.2% (7.6%) for B/D(ESS), the distributions were asymmetric, the proportions significant at the 2.5% (5%) level in the left tail being only 0.3% (0.8%) for the D(ESS) and 0.5% (1.3%) for the B/D(ESS).

Values in Fig. 4, column 5, represent type I error rates for simulations where a different disease prevalence of between 5% and 50% was selected for each study in the meta-analysis, the average prevalence being 22.5%. The same pattern of inflated type I error rates for E(SE), B(SE), and M(N) was evident, although type I error rates were only approximately double their nominal significance levels. The asymmetry for D(ESS) and B/D(ESS) was evident at more extreme disease prevalences.

6.5. Impact of heterogeneity in diagnostic accuracy

Underlying diagnostic accuracy was varied between studies in a meta-analysis by introducing a random effect with a normal distribution. Introduction of the random effect affected E(SE), B(SE), D(ESS), and B/D(ESS) even when the average DOR was equal to one. Type I errors for the Egger and Begg methods became inflated, and ESS-based methods became conservative.

Column 6 of Fig. 4 shows results when a large random effect with a standard deviation of $0.3\sigma_1$ is introduced. For a DOR of 38, a decrease of $0.3\sigma_1$ in the difference between the means causes the DOR to drop to 13, and an increase of $0.3\sigma_1$ gives a DOR of 112. For a DOR of 231, the equivalent figures are 78 to 685. Type I error rates are excessive for E(SE) and B(SE), whereas the ESS-based tests become conservative. For a DOR of 231, the proportion significant at 5% (10%) significance levels are 24.4% (38.1%) for E(SE), 13.0% (23.0%) for B(SE), 1.5% (4.3%) for D(ESS), and 1.7% (4.9%) for B/D(ESS). For Macaskill's sample size–based test M(N), the empirical type I error rate was 7.1% (13.0%).

6.6. Impact of asymmetry in the SROC (DOR related to threshold)

Doubling the standard deviation of the distribution of diseased participants has the effect of introducing asymmetry into the shape of the underlying ROC curve such that the DOR changes with threshold. There was no additional impact of having an underlying asymmetric ROC when the threshold was fixed or varied between studies (data not shown).

6.7. Impact of publication bias on estimates of diagnostic test accuracy

Figure 5 depicts four illustrative meta-analysis datasets created by simulation with an underlying DOR of 38 with variation in threshold, disease prevalence and accuracy (introduced by a random effect with standard deviation of $0.3\sigma_1$). Studies have been censored using the four alternative publication bias mechanisms. Open circles indicate missing studies; these are distributed as would be expected according to the censoring mechanism. Censoring removed between 5 and 12 studies in each plot; censoring on sensitivity removed studies with low sensitivity, censoring on specificity removed studies with low specificity, censoring using Youden's index removed studies with either low sensitivity or low specificity, and AUC-based censoring removed studies beneath the ROC curve with lower DOR. Despite the censoring, changes in the estimated summary ROC curve superficially appear small.

Average meta-analytical estimates of sensitivity, specificity, and DOR according to degree of censoring are given in Table 1. These are based on simulations with the same parameter values as in Fig. 5, with and without heterogeneity in DOR. For the Moses SROC regression model estimates of DOR, censoring of ≤50% of studies appeared to make only a modest difference in test accuracy, regardless of the censoring mechanism used. Censoring on the DOR (through values of the AUC) introduced the largest bias, increasing the DOR from 37 to 53 when 50% of studies were excluded. In practical application, this change is small: the summary ROC curve for a DOR of 53 passes through the sensitivity = specificity = 88% point compared to sensitivity = specificity = 86% point for a DOR of 37. Notable bias is introduced, however, when separate meta-analytical estimates are computed for sensitivity and specificity if the censoring mechanism acts unequally on sensitivity and specificity.

6.8. Statistical power

The statistical power of 10% two-sided tests for funnel plot asymmetry is shown in Fig. 6 for simulations created using the same parameters as in Fig. 5, without and with heterogeneity in DOR (including a random effect of $0.2\sigma_1$).

Although the Egger E(SE) and Begg B(SE) tests show the greatest power for detecting publication bias, the inflated type I error rate of these tests is evident in the values of power being above the nominal two-sided 10% level where they intersect the true value of the DOR. The proposed ESS-based regression D(ESS) and rank correlation B/D(ESS) tests show reasonable power when there is no heterogeneity in DOR, the regression test outperforming the correlation test. When test accuracy varies with a random effect as well as through sampling variability, the power of all tests rapidly dissipates.

7. Discussion

We found that a funnel plot can be used to identify a sample size related effect such as that caused by publication bias in reviews of diagnostic test accuracy. The Begg, Egger, and Macaskill tests of funnel plot asymmetry used for RCTs are, however, likely to be seriously misleading if applied in typical diagnostic test scenarios. DORs usually take values well above one, test thresholds often preferentially favor sensitivity over specificity (or vice versa), there are usually fewer diseased than nondiseased, and heterogeneity in test accuracy is common. We have shown algebraically and by simulation how the approximate asymptotic standard error of the log odds ratio is affected by these phenomena and how they cause the Begg, Egger, and Macaskill tests to overestimate the frequency of sample size related effects.

We propose that systematic reviewers should undertake funnel plot investigations to examine the possibility of publication and other sample size related effects using plots of lnDOR against $1/\text{ESS}^{1/2}$, and test for asymmetry using related regression or rank correlation tests. We have shown that these tests are robust to features characteristic of studies of diagnostic test accuracy, and that the regression test has greater power to detect publication bias than the rank correlation test. Our observation of lower power for the correlation test is consistent with findings from previous studies [14,16].

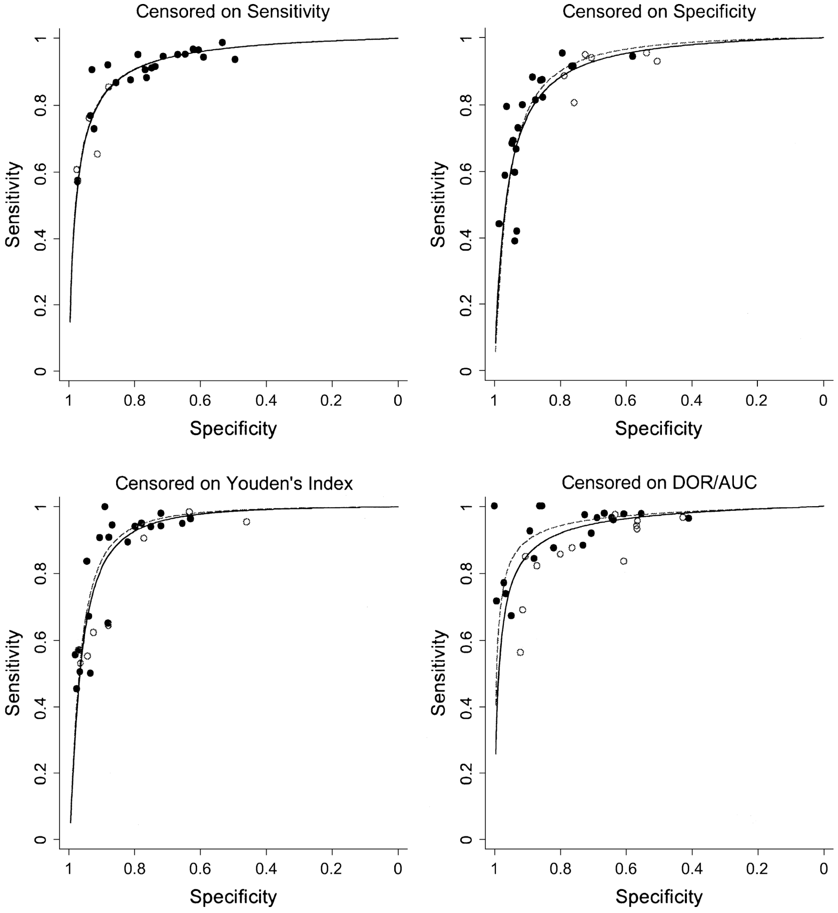

Fig. 5. Impact of publication bias on an example simulated meta-analyses. Censored studies indicated by open circles. The dashed line indicates the SROC curve estimated from the published studies, the solid line indicates the summary ROC curve estimated from all studies. DOR/AUC indicates censoring on the diagnostic odds ratio through values of the area under the ROC curve.

Table 1
Estimates of the effects of publication bias on meta-analytical estimates of test accuracy

	Censored on			
Measure	Sensitivity	Specificity	Youden's index	DOR/AUC
0% censored				
DOR	37	37	37	37
Sensitivity, %	87	87	87	87
Specificity, %	86	86	86	86
10% censored				
DOR	38	38	39	40
Sensitivity, %	89	85	87	87
Specificity, %	84	88	86	87
25% censored				
DOR	40	40	41	44
Sensitivity, %	91	82	88	88
Specificity, %	81	90	87	87
50% censored				
DOR	44	46	46	53
Sensitivity, %	94	75	88	90
Specificity, %	75	94	87	88

Diagnostic odds ratio (DOR) estimated from an unweighted Moses summary receiver operating characteristics (SROC) regression model. Sensitivity estimated as the inverse variance-weighted average of logit sensitivities. Specificity estimated as the inverse variance-weighted average of logit specificities. DOR/AUC indicates censoring on the diagnostic odds ratio through values of the area under the ROC curve.

The problems we have identified with the approximate asymptotic standard error of the log odds ratio also affect the use of inverse variance study weights for meta-analytical pooling and meta-regression investigations when DORs are very much greater than one. Effective sample size–based weights may offer a preferable alternative.

Our simulations also suggest that, if meta-analysis is based on estimation of a DOR, the impact of publication bias on estimates of DOR is unlikely to be large; however, selective publication may seriously affect estimates of sensitivity and specificity. These findings were shown to be valid for varying degrees of publication bias and for different selection mechanisms. The evaluation of the impact of censoring 50% of studies was based on simulations where study results were very variable and the censoring function took a particularly steep form, which are conditions for the impact of publication bias to be strong. Further empirical work needs to be undertaken to better understand the determinants and magnitude of publication bias for diagnostic accuracy studies.

Notably, we have shown that the power of all statistical tests of funnel plot asymmetry decreases when the DOR varies more than expected by chance in a way that is not associated with sample size. The proportion of the total

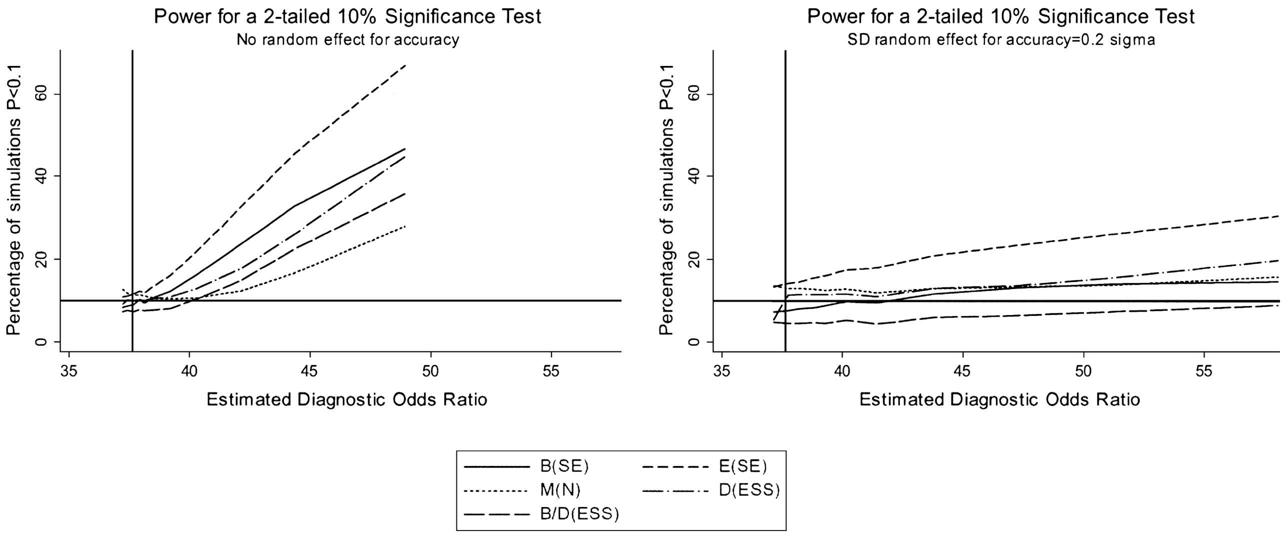

Fig. 6. Empirical power of tests for funnel plot asymmetry.

variability that is attributable to sampling variability decreases as heterogeneity increases. In situations where studies have highly variable results, the chances of detecting variability in accuracy associated with sample size are likely to be low.

Funnel plot analyses show asymmetry for a variety of reasons other than publication bias, including the type of population studied and poor study quality, if they are linked both to sample size and observed diagnostic accuracy [30]; however, the likely direction of the relationships is not always clear. Whereas for trials smaller studies are more likely to have larger effects due to poor methodological quality, in diagnostic research it is possible that the larger studies are of poorer quality. For, example large retrospective studies in which investigators obtain test results from clinical databases may be more biased than smaller prospective studies in which clinicians carefully recruit patients presenting with a specific clinical problem. Larger studies may also be more prone to verification bias, if adequate resources are not available to correctly ascertain the gold standard reference diagnosis on all participants. Real differences in test accuracy between participant groups will also induce asymmetry if they are linked to sample size. For RCTs, sample sizes are usually smaller for groups in which treatment effects are expected to be large [31]. Because the design of diagnostic test accuracy studies does not usually involve power calculations for testing hypotheses, it is not clear that the same relationship would be as evident. Constructing funnel plots using different symbols and colors to denote key study characteristics may assist in eliciting likely causes of asymmetry.

In summary, we have shown that funnel plot investigations based on the standard error of the lnDOR are seriously misleading. We recommend instead using effective sample size–based funnel plots and associated regression tests of asymmetry. The impact of this change in funnel plot structure is likely to be high. Applying the proposed regression test to the 28 reviews identified by Song et al. [12], only 3 show significant ($P < .10$) funnel plot asymmetry, compared to 12 using the Egger test, 5 using the Begg test, and 8 using the Macaskill test.

Acknowledgments

We are grateful to Fujian Song for providing datasets from his study [12], and to Patrick Bossuyt for comments on a previous draft. The work was supported by National Health and Medical Research Council (NHMRC) grant 211205 to the Screening and Test Evaluation Program. Jon Deeks is supported by a U.K. Department of Health Senior Research Fellowship in Evidence Synthesis.

Appendix A

Relationship of standard error with diagnostic odds ratio, threshold, and sample size

To assess the mechanism by which the standard error of the log-diagnostic odds ratio, or SE(lnDOR), operates, consider the re-parameterization of the asymptotic estimator (where TP is true positive, FN is false negative, FP is false positive, and TN is true negative)

$$\mathrm{SE}(\ln\mathrm{DOR}) = \sqrt{\frac{1}{\mathrm{TP}} + \frac{1}{\mathrm{FN}} + \frac{1}{\mathrm{FP}} + \frac{1}{\mathrm{TN}}} \qquad \text{(A1)}$$

using the terms $\phi = \mathrm{DOR} = (\mathrm{TP} \times \mathrm{TN})/(\mathrm{FP} \times \mathrm{FN})$; n_1 = number not diseased = TN + FP; n_2 = number diseased = TP+ FN; and r = odds of testing negative in the nondiseased = TN/FP. Recall that variation in r between studies in a meta-analysis reflects variation in the number testing positive due to differences in the threshold used to define test positive. This yields the following equation:

$$\text{SE}(\ln \text{DOR}) = \sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\left(r+\frac{1}{r}+2\right)+\left(\frac{\phi-1}{n_2}\right)\left(\frac{1}{r}-\frac{r}{\phi}\right)} \quad \text{(A2)}$$

The three functions contained in this equation have the following three properties.

First, the sample size dependent term:

$$f(n_1,n_2) = \frac{1}{n_1}+\frac{1}{n_2}=\frac{n_1+n_2}{n_1 n_2} \quad \text{(A3)}$$

The standard error inversely relates to effective sample size $(4n_1n_2)/(n_1+n_2)$, appropriately reflecting unequal numbers in diseased and nondiseased groups.

Second, the proportion testing positive dependent term:

$$g(r) = r + (1/r) + 2 \quad \text{(A4)}$$

The standard error is minimized when the numbers of true negatives and false positives are equal ($r = 1$). For fixed values of n_1 and n_2, shifting the threshold changes r and alters the standard error in a multiplicative manner.

Third, the DOR dependent term:

$$h(\phi,r,n_2) = \left(\frac{\phi-1}{n_2}\right)\left(\frac{1}{r}-\frac{r}{\phi}\right) = \left(\frac{\phi-1}{n_2}\right)\left(\frac{\text{FP}}{\text{TN}}-\frac{\text{FN}}{\text{TP}}\right) \quad \text{(A5)}$$

The standard error is increased or decreased according to an additive term dependent on the DOR. The term is zero when DOR = 1 (i.e., for a test with no diagnostic value) and in the special case when sensitivity and specificity are equal. For a fixed value of r, the term is positive if sensitivity is greater than specificity and negative otherwise. The magnitude of the term decreases with increasing numbers of diseased.

Thus, of the three functions, only the first, $f(n_1, n_2)$, operates appropriately under the three characteristics of meta-analyses of diagnostic test accuracy noted above. The second function, $g(r)$, will introduce variation in precision solely due to differences in threshold, and the third $h(\phi, r, n_2)$ will introduce a direct correlation between precision and observed diagnostic accuracy. Sampling variability (measurement error) in the standard error is a problem for regression tests of asymmetry. Both of the functions $g(r)$ and $h(\phi, r, n_2)$ are also observed with measurement error. The $f(n_1, n_2)$ is observed with measurement error if diseased and nondiseased groups are recruited as part of a cohort (but not if they are recruited separately), but will be small in most studies.

For the example case study in section 4, we highlighted studies 4 and 13, whose location on the funnel plot changed substantially with choice of axes. We compute the standard error and its terms $f(n_1, n_2)$, $g(r)$, and $h(\phi, r, n_2)$ are

$$\text{SE}(\ln \text{DOR}) = \sqrt{f(n_1, n_2)g(r) + h(\phi, r, n_2)} \quad \text{(A6)}$$

This calculates as $[(0.052 \times 249.00) - 10.65]^{1/2} = 1.48$ for study 4 and as $[(0.045 \times 211.00) - 7.26]^{1/2} = 1.52$ for study 13.

It therefore appears that for these two studies the observed high threshold dominates the computation of SE (from comparison of the values of $g(r)$ of 249 and 211 with the minimal value of $g(r)$ of 4). The DOR dependent term $h(\phi, r, n_2)$ is also large, but is negative because the specificity is higher than sensitivity.

Appendix B

Computation of diagnostic odds ratio, area under the curve, sensitivity, and specificity

For fixed values of μ_1, μ_2, σ_1, σ_2, and t, the underlying value of the diagnostic odds ratio (DOR) for the simulation can be calculated as

$$\text{DOR} = \exp\left[\sqrt{\frac{\pi^2}{3}}\left(\frac{\mu_2-t}{\sigma_2}-\frac{\mu_1-t}{\sigma_1}\right)\right] \quad \text{(B1)}$$

which simplifies to

$$\text{DOR} = \exp\left[\sqrt{\frac{\pi^2}{3}}\left(\frac{\mu_2-\mu_1}{\sigma}\right)\right] \quad \text{(B2)}$$

when $\sigma_1 = \sigma_2 = \sigma$, and is independent of the threshold, t. For this simplified situation, the equivalent area under the curve (AUC) can be computed using a formula from Walter [32]:

$$\text{AUC} = \frac{\text{DOR}}{(\text{DOR}-1)^2}\left[(\text{DOR}-1)-\ln \text{DOR}\right] \quad \text{(B3)}$$

The underlying sensitivity and specificity can be obtained as follows.

$$\text{Sensitivity} = \frac{\exp\left[\sqrt{\frac{\pi^2}{3}}\left(\frac{\mu_2-t}{\sigma_2}\right)\right]}{1-\exp\left[\sqrt{\frac{\pi^2}{3}}\left(\frac{\mu_2-t}{\sigma_2}\right)\right]} \quad \text{(B4)}$$

and

$$\text{Specificity} = 1-\frac{\exp\left[\sqrt{\frac{\pi^2}{3}}\left(\frac{\mu_1-t}{\sigma_1}\right)\right]}{1+\exp\left[\sqrt{\frac{\pi^2}{3}}\left(\frac{\mu_1-t}{\sigma_1}\right)\right]} \quad \text{(B5)}$$

The standardized distance between the means determines the diagnostic accuracy of the test. Simulations were undertaken for five different levels of accuracy: $(\mu_2 - \mu_1)/\sigma_1 = 0$

(DOR = 1, AUC = 0.50); $(\mu_2 - \mu_1)/\sigma_1 = 1$ (DOR = 6, AUC = 0.77); $(\mu_2 - \mu_1)/\sigma_1 = 2$ (DOR = 38, AUC = 0.93); $(\mu_2 - \mu_1)/\sigma_1 = 3$ (DOR = 231, AUC = 0.98); and $(\mu_2 - \mu_1)/\sigma_1 = 4$ (DOR = 1415, AUC = 0.996).

References

[1] Begg CB, Berlin JA. Publication bias: a problem in interpreting medical data. J R Stat Soc A 1988;151:419–63.

[2] Begg CB. Publication bias. In: Cooper H, Hedges LV, editors. The handbook of research synthesis. New York: Sage Foundation, 1994; 399–410.

[3] Clarke M, Oxman AD, editors. The Cochrane Reviewers' Handbook 4.2.0. The Cochrane Library. Oxford: Update Software; Issue 2; 2003 [current version available at http://www.cochrane.org/resources/handbook/].

[4] Khan KS, ter Riet G, Glanville J, Sowden AJ, Kleijnen J, editors. Undertaking systematic reviews of research on effectiveness: CRD's guidance for those carrying out or commissioning reviews. Report 4. 2nd ed. York, UK: NHS Centre for Reviews and Dissemination, University of York; 2001.

[5] Dickersin K, Scherer R, Lefebvre C. Identifying relevant studies for systematic reviews. BMJ 1994;309:1286–91.

[6] Egger M, Juni P, Bartlett C, Holenstein F, Sterne J. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. Health Technol Assess 2003;7(1):1–76.

[7] Song F, Eastwood AL, Gilbody S, Duley L, Sutton AJ. Publication and related biases. Health Technol Assess 2000;4(10):1–115.

[8] Dickersin K. How important is publication bias? A synthesis of available data. AIDS Educ Prev 1997;9:15–21.

[9] Dickersin K. The existence of publication bias and risk factors for its occurrence. JAMA 1990;263:1385–9.

[10] Ioannidis JPA. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. JAMA 1998;279:281–6.

[11] Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. BMJ 1997;315:640–5.

[12] Song F, Khan K, Dinnes J, Sutton AJ. Asymmetric funnel plots and publication bias in meta-analyses of diagnostic accuracy. Int J Epidemiol 2002;31:88–95.

[13] Clark TJ, Voit D, Gupta JK, Hyde C, Song F, Khan KS. Accuracy of hysteroscopy in the diagnosis of endometrial cancer and hyperplasia: a systematic quantitative review. JAMA 2002;288:1610–21.

[14] Clark TJ, Mann CH, Shah N, Khan KS, Song F, Gupta JK. Accuracy of outpatient endometrial biopsy in the diagnosis of endometrial cancer: a systematic quantitative review. BJOG 2002;109:313–21.

[15] Sterne JAC, Gavaghan DJ, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. J Clin Epidemiol 2000;53:1119–29.

[16] Macaskill P, Walter SD, Irwig L. A comparison of methods to detect publication bias in meta-analysis. Stat Med 2001;20:641–54.

[17] 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. Stat Med 2002;21:2465–77.

[18] Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, de Vet HCW. Standards for Reporting of Diagnostic Accuracy (STARD) group. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Ann Intern Med 2003;138:40–4.

[19] Light RJ, Pillemer DB. Summing up: the science of reviewing research. Cambridge, MA: Harvard University Press; 1984.

[20] Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997;315:629–34.

[21] Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. Biometrics 1994;50:1088–101.

[22] Sterne JAC, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. J Clin Epidemiol 2001;54:1046–55.

[23] Tang J-L, Liu JLY. Misleading funnel plot for detection of bias in meta-analysis. J Clin Epidemiol 2000;53:477–84.

[24] Moses LE, Shapiro DE, Littenberg B. Combining independent studies of a diagnostic tests into a summary ROC curve: data-analytic approaches and some additional considerations. Stat Med 1993;12: 1293–316.

[25] Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med 2001;20: 2865–84.

[26] Irwig L, Macaskill P, Glasziou P, Berry G. Bias in meta-analysis detected by a simple graphical test. Graphical test is itself biased. [Letter]. BMJ 1998;316:470; author reply 470–1.

[27] Armitage P, Berry G, Matthews JNS. Statistical methods in medical research. 4th ed. Oxford: Blackwell Science; 2002.

[28] Kearon C, Julian JA, Newman TE, Ginsberg JS. Noninvasive diagnosis of deep vein thrombosis. McMaster Diagnostic Imaging Practice Guidelines Initiative. Ann Intern Med 1998;128:663–7 [Erratum in: Ann Intern Med 1998;129:425].

[29] Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. J Clin Epidemiol 1993;48:119–30.

[30] Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, Bossuyt PM. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999;282:1061–6 [Erratum in: JAMA 2000;283:1963].

[31] Sterne JAC, Egger M, Davey Smith G. Systematic reviews in health care: investigating and dealing with publication and other biases in meta-analysis. BMJ 2001;323:101–5.

[32] Walter SD. Properties of the summary receiver operating characteristic (ROC) curve for diagnostic test data. Stat Med 2002;21:1237–56.