ELSEVIER

Journal of Clinical Epidemiology 61 (2008) 41−51

Journal of Clinical Epidemiology

The binomial distribution of meta-analysis was preferred to model within-study variability

Taye H. Hamza[a,*], Hans C. van Houwelingen[b], Theo Stijnen[a]

[a]*Department of Epidemiology and Biostatistics, Erasmus MC—Erasmus University Medical Center, Rotterdam, P.O.Box 2040, 3000 CA Rotterdam, The Netherlands*

[b]*Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands*

Accepted 21 March 2007

Abstract

Objective: When studies report proportions such as sensitivity or specificity, it is customary to meta-analyze them using the DerSimonian and Laird random effects model. This method approximates the within-study variability of the proportion by a normal distribution, which may lead to bias for several reasons. Alternatively an exact likelihood approach based on the binomial within-study distribution can be used. This method can easily be performed in standard statistical packages. We investigate the performance of the standard method and the alternative approach.

Study Design and Setting: We compare the two approaches through a simulation study, in terms of bias, mean-squared error, and coverage probabilities. We varied the size of the overall sensitivity or specificity, the between-studies variance, the within-study sample sizes, and the number of studies. The methods are illustrated using a published meta-analysis data set.

Results: The exact likelihood approach performs always better than the approximate approach and gives unbiased estimates. The coverage probability, in particular for the profile likelihood, is also reasonably acceptable. In contrast, the approximate approach gives huge bias with very poor coverage probability in many cases.

Conclusion: The exact likelihood approach is the method of preference and should be used whenever feasible.

Keywords: Meta-analysis; Diagnostic tests; Random effects; Sensitivity; Specificity; Exact binomial

1. Introduction

In this paper we consider meta-analysis of proportions. Very frequently occurring examples of proportions being meta-analyzed are sensitivity or specificity of a diagnostic test. Therefore, this article is written from a diagnostic research perspective, though the results apply to meta-analysis of proportions in general, such as prevalences or incidences.

Meta-analytic methods for a diagnostic test depend on the type of data that are available from the different studies. In most medical articles, the commonly reported measures of diagnostic test accuracy are sensitivity and/or specificity. Alternatively, other measures such as diagnostic odds ratio (OR), predictive values, and area under the receiver operating characteristic (ROC) curve are reported.

Statistical methods to pool the results of diagnostic test measures from different studies lay on different assumptions. For example, it might be assumed that the observed differences between individual study results are only due to sampling variation, leading to what is called a fixed effect analysis. When an estimate of the sensitivity or specificity is reported in a single study, the simplest method to get a summary measure is to calculate the average sensitivity and/or specificity, possibly with weights depending on the within-study sample sizes or standard errors (SEs). However, this approach is usually inappropriate because it is likely that variability beyond chance can be attributed to between-study differences [1,2]. Some of the between-study variability could be accounted for using explanatory variables in a regression analysis. But mostly not all heterogeneity can be explained, and a random effects model is used in the statistical analysis that allows between-studies heterogeneity [3,4].

In the last decade, many random effects methods have been developed to relax the fixed effect assumptions in meta-analysis [5−8] of diagnostic tests [9,10]. Some of

* Corresponding author. Tel.: +31-10-40-87-481; fax: +31-10-40-89-382.

E-mail address: t.hussienhamza@erasmusmc.nl (T.H. Hamza).

0895-4356/08/$ − see front matter
doi: 10.1016/j.jclinepi.2007.03.016

these methods enable analyzing sensitivity and specificity jointly. However, in the medical literature numerous meta-analyses are published in which one is interested in meta-analyzing only sensitivity or specificity, and in this paper we concentrate on this situation. Then the standard way of analysis is with the DerSimonian and Laird [6] random effects model. It is not well known that this method can be heavily biased when it is applied to proportions, such as specificities or sensitivities, though some authors have mentioned this [5,11,12]. Chang et al. [11] have proposed a method that repairs the bias. However, this article has been cited only once since the year 2001, showing that in practice this method is not used. It might be due to the difficulty to perform the method easily in standard statistical packages. The reason for the standard method being biased is that the binomial within-study likelihood of the sensitivity or specificity is approximated using a normal likelihood. It is well known that this approximation can be bad if the proportion is close to one or zero, and/or the sample size is small. So bias can be expected if this is the case in a meta-analysis. However, even if the normal approximation would be good enough for ordinary applications, bias could be introduced because the use of the normal approximation in meta-analysis ignores the correlation between the estimated proportion and its variance. We come back to this point in the next section. Nowadays standard statistical packages allow for fitting generalized linear mixed models (GLMM). This makes it very easy to use the exact binomial within-study distribution of the estimated sensitivity or specificity instead of a normal approximation of it. In this article, we call the latter the approximate method and the former the exact method.

The purpose of this article is to compare the performance of the two modeling approaches, approximate and exact, through a simulation study. In Section 2, both methods are discussed. In Section 3 we describe the design of the simulation study, and in Section 4 we present the results. In Section 5, we apply the methods on real meta-analysis data. We end with a discussion in Section 6. We used SAS software (version 9; SAS Institute, Cary, NC) to simulate the data and to estimate the parameters for the models discussed in Section 2.

2. Random effects model

In a situation where the interest is to meta-analyze sensitivities or specificities separately, the commonly used method is the DerSimonian and Laird [6] random effects model. In the remainder of this paper we will talk about meta-analyzing sensitivities, but all the results apply to specificities as well. In fact, the results apply to any meta-analysis where the target parameter is a proportion or probability and each study contributes a sample size and a number of ''successes.'' Unlike a fixed effect model, a random effects model allows that sensitivities vary across studies beyond that expected from within-study sampling variability alone. More specifically, the true logit sensitivities, η_i, defined as ln(sensitivity/(1 − sensitivity)) are assumed to follow a normal distribution:

$$\eta_i \sim N(\eta, \tau^2)$$

Here i denotes the number of study, η the true mean logit sensitivity. The parameter τ^2 is called the between-studies variance, and it describes the variability between the true logit sensitivities of the different studies. The within-study sampling variability could be modeled by using the approximate normal likelihood or the exact binomial likelihood for the observed number of positive test results.

2.1. Approximate method

This is the standard method in practice. Different transformations of the observed proportion, such as the probit, log(−log) or the arcsine could be used and approximated by a normal distribution. In this paper, we have chosen the logit transformation because it is the predominant choice in practice. If m_i is the total number of subjects with the disease of study i, and x_i is the observed number of true positive test results in the group with the disease, then the observed logit sensitivity, $\hat{\eta}_i = \ln(x_i/(m_i - x_i))$ is assumed to follow an approximate normal distribution with mean η_i, and within-study variance calculated from the observed data:

$$\hat{\eta}_i \sim N\left(\eta_i, \hat{\sigma}_i^2\right) \quad \text{with } \hat{\sigma}_i^2 = \frac{1}{x_i} + \frac{1}{m_i - x_i}$$

If x_i or $m_i - x_i$ is zero, the logit sensitivity and the within-study variance will be undefined. To avoid this problem 0.5 should be added to x_i and $m_i - x_i$ for all studies, including those with no zero [13,14]. The effect of adding 0.5 may bias the results [14]. Further, usually there is a high correlation between $\hat{\eta}_i$ and $\hat{\sigma}_i^2$. The correlation is positive if η is positive and negative if η is negative, leading to a bias toward zero in the estimate of the overall logit sensitivity, or a bias toward 0.5 in the estimate of the overall sensitivity. This is because both the mean and the variance of $\hat{\eta}_i$ are determined by the same parameter. The effect of this correlation in a random effects meta-analysis was discussed by several authors [5,11,12]. Though they suggested a correction to reduce the bias in the estimate of η due to this correlation, Chang et al. [11] mentioned that the estimated between-studies variance remained still biased even if we let grow the number of studies included. Using the approximate normal likelihood for the within logit sensitivity, the model is a linear random effects model, and the parameters can be estimated by standard likelihood procedures using a Linear Mixed Model program that is available in many standard statistical packages. For example, in SAS the procedure MIXED and in S-Plus/R the function lme can be used. As discussed by Turner et al. [15], three methods, maximum likelihood (ML), restricted (residual) maximum likelihood (REML), and the method of moments proposed by DerSimonian and Laird, are available to estimate the

random effects model. They differ mainly on the estimation of the between-studies variance, in which ML gives a downward bias. In the biomedical literature, usually the method of moments proposed by DerSimonian and Laird or the REML estimator is used for estimating the heterogeneity parameter [16]. The REML estimator is the iterative equivalent of the DerSimonian and Laird estimator and gives very similar results [6]. In this paper we used the REML method that can be specified in the method option of the MIXED or lme procedures.

2.2. Exact method

Here, we use the fact that the observed number of true positive test results x_i follows a binomial distribution:

$$x_i \sim \text{binomial}(\pi_i, m_i)$$

where $\pi_i = 1/\left(1 + e^{-\eta_i}\right)$ and m_i is the total number of subjects with the disease of study i. In this case, there is no need to add 0.5 even if a zero count is encountered, and there is no problem anymore of correlation between the observed values and their variance because there is no more need of calculating the $\hat{\eta}$ and its within-study variance, $\hat{\sigma}_\eta^2$ from the observed data and approximating by a normal distribution. Now the variance is inherent from the binomial distribution. In the approximate method, the distribution of x_i is basically approximated by a normal distribution with mean $m_i\pi$ and variance $m_i\pi(1-\pi)$ that is estimated by a normal distribution with mean $m_i\hat{\pi}$ and variance $m_i\hat{\pi}\left(1-\hat{\pi}\right)$. Clearly the estimated mean and variance are correlated, and, because the variance estimate is treated as a fixed and known number, this correlation is not modeled, which causes bias in the approximate method.

Now the model is a GLMM, and the parameters can be estimated by standard likelihood procedures. The practical disadvantage is that software is much more scarce and not yet available in all statistical packages. We used the NLMIXED procedure from the SAS package [17]. It is also possible to use the recently included GLIMMIX procedure in the SAS package, which is still experimental in SAS version 9.1. The GLIMMIX procedure allows more random effects, but it has the disadvantage that it uses an approximation instead of the true log likelihood. In contrast, although the number of random effects that can be practically managed is limited, NLMIXED uses very accurate integrating techniques to calculate the true likelihood. Unlike the MIXED procedure, the NLMIXED procedure only implements ML. This is because the analog to the REML method in NLMIXED would involve a high dimensional integral over all of the fixed effects parameters, and this integral is typically not available in closed form [17]. Hence, we used the ML method for the exact approach.

The exact binomial likelihood approach as used here leads to a logistic regression model with a random intercept, and is therefore analogous to the ''individual patient data methods'' as used by Turner et al. [15]. They considered meta-analysis of treatment effect log ORs and used the MLwiN software [18] to fit the model. In contrast to NLMIXED, this program uses an approximation of the likelihood instead of the true one. Turner et al. [15] also suggested the use of bootstrapping techniques to estimate τ^2, but this approach is not implemented in NLMIXED and was not incorporated in our simulation study.

In the Appendix, we have given the syntax needed to fit the models following the approximate and the exact methods.

3. Simulation study

A simulation study was carried out to compare the performance of the two methods, approximate and exact, discussed in Section 2. We investigated the effect of the number of studies included in the meta-analysis, the mean within-study sample size, the between-study variability, and the true median sensitivity. The data were simulated in two steps. First, the true logit sensitivity, η_i, was simulated from a normal distribution with a given mean logit sensitivity η and between-studies variance τ^2. Secondly, the within-study data were simulated from a binomial distribution with a probability $\pi_i = 1/\left(1 + e^{-\eta_i}\right)$ and within-study sample size m_i. In practice, the m_i's vary across studies included in the meta-analysis. In some meta-analyses the range of the size of studies is as big as 1,500 or more (e.g., [19,20]). To accommodate this variation, the m_i's were generated from a normal distribution and rounded to the nearest integer. Two different values were considered for the mean m_i (standard deviation [SD]): 40 (30) and 500 (450). The minimum study size was set to be 10, that is, if the generated m_i was less than 10 then it was taken to be 10. Consequently 40 and 500 are no more the means for the simulated sample sizes, but the medians and the realized SD become a bit smaller than 30 and 450, respectively. We considered 12 different situations, and for each situation we did the simulation assuming a different number of studies (10, 25, 50, 100) included in the meta-analysis at a time, that is, we simulated 48 scenarios in total. All values assigned to the parameters in the different scenarios were based on real data sets from the medical literature (e.g., [21–23]. An overview of the simulated scenarios is given in Table 1.

Each scenario was replicated 1,000 times, and the simulated data sets were analyzed according to the approximate and exact approach using the SAS procedures MIXED and NLMIXED, respectively. We concentrated on the estimation of the mean logit sensitivity η and between-study variance τ^2. The estimated results were compared using bias (difference between the mean estimate and the true value of the parameter), mean-squared error (MSE), and coverage probability of the 95% confidence interval (CI) (the frequency in which the true value falls in the CI). In meta-analysis, mostly Wald-type CIs are used. For η, the Wald-type CI is $\hat{\eta} \pm 1.96 \times \text{SE}\left(\hat{\eta}\right)$. A disadvantage of this CI might be that when the number

Table 1
The different scenarios used in the simulation study. Each subset of four scenarios corresponds to 10, 25, 50, and 100 studies in the meta-analysis. The corresponding true median sensitivity for η 0.41, 1.11, and 2.57 are 0.60, 0.75, and 0.93, respectively

Scenario	η	τ^2	m	Standard deviation of m
1–4	0.41	0.3	40	30
5–8	0.41	0.3	500	450
9–12	0.41	1.0	40	30
13–16	0.41	1.0	500	450
17–20	1.11	0.3	40	30
21–24	1.11	0.3	500	450
25–28	1.11	1.0	40	30
29–32	1.11	1.0	500	450
33–36	2.57	0.3	40	30
37–40	2.57	0.3	500	450
41–44	2.57	1.0	40	30
45–48	2.57	1.0	500	450

of studies is small, the SE of $\hat{\eta}$ is underestimated due to the fact that the uncertainty in the estimate of τ^2 is not accounted for. This problem may be solved using a profile likelihood-based CI [24], which is also discussed by several authors in the context of meta-analysis (e.g., [8,25–28]). Turner et al. [15] also discussed a bootstrapping technique to provide CIs for η and τ^2. This method is directly available in the MLwiN software [18]. Recently, Knapp et al. [16] proposed a new approach for a CI of the heterogeneity parameter. In this article, we restricted to the Wald and profile likelihood approaches.

The profile log ikelihood of η is defined as $\mathrm{pl}(\eta)=\max_{\tau^2} \mathrm{l}(\eta,\tau^2)$ where $\mathrm{l}(\eta,\tau^2)$ is the log likelihood for η and τ^2. The 95% profile likelihood CI for η is then given by all values that satisfy $\mathrm{pl}(\eta) > \mathrm{pl}(\hat{\eta}) - 1.92$ (1.92 is the 95% percentile of the χ_1^2 distribution (3.84) divided by two). The Wald-type CI for τ^2 was calculated through a logarithmic transformation as $\hat{\tau}^2 \exp\left(\pm 1.96 \mathrm{SE}(\hat{\tau}^2)/\hat{\tau}^2\right)$. In a similar way as for η, we also calculated a profile likelihood CI for τ^2. The profile likelihood for both approaches is based on ML only, because the likelihood ratio test statistics computed directly from REML for the fixed effect parameter may not be valid [29,30]. This type of CI cannot be automatically done in NLMIXED and needs some extra programming. For example, to find the profile likelihood CI of η, first the model is fitted in NLMIXED, and the ML value while estimating both η and τ^2 is saved. The value of the profile likelihood for a given η-value is calculated by running NLMIXED keeping η fixed to this value. Then the two η-values with profile likelihood equal to ML − 1.92 were found iteratively using the bisection method, using a self-written macro.

4. Simulation results

The results from the simulations are presented in Figs. 1 and 2, and Table 2. Fig. 1 shows the biases and MSEs for the mean logit sensitivity η. It can be seen from Fig. 1a that the exact likelihood approach yields estimates of η that are quite unbiased regardless of the different scenarios; that is, the expected value of the estimated η using the exact method is almost equal to the true value, and always closer to the true value than the approximate likelihood method. The bias in the approximate method varies considerably with the within-study sample size and true mean logit sensitivity. It increases dramatically when the within study sample size is smaller and the true median sensitivity is larger. For example, when the median within study sample size is 40 and the median sensitivity is 0.93, the estimated logit sensitivity is biased downward by about 35% or more regardless of the true between-studies variance and the number of studies included. Concerning the effect of between-studies variances, the larger the between-studies variance, the more the approximate method underestimates logit sensitivity. However, the number of studies included in the meta-analysis does not make much difference on the bias.

The MSEs in Fig. 1b are corrected for the number of studies included in the meta-analysis by $N/100$, that is, the vertical axis in Fig. 1b is MSE $\times$ $N/100$. Comparing the two approaches in terms of the MSEs, the approximate method tends to be worse (larger MSE) for the scenarios with small within study sample size and medium median sensitivity, or large median sensitivity, in which case the difference in the MSEs between the two methods increases with the number of studies included in the meta-analysis. The constant MSE $\times$ $N/100$ in the figure indicates the fact that the MSE, before multiplying by $N/100$, decreases with the number of studies included in the meta-analysis regardless of the methods used and the different scenarios.

The coverage probabilities based on the Wald-type CIs and the profile likelihood CIs are presented in Table 2. Regardless of the method used to construct the CIs, the exact method performs better than the approximate almost always. When $N = 10$, the coverage probabilities of the approximate method are particularly bad for larger sensitivity and small within-study sample size. The coverage probabilities of the exact method are quite reasonable except for the Wald-type CI. The intervals based on the profile likelihoods give a valuable improvement upon the Wald-type intervals. When $N \geqslant 25$, not much difference is observed between the two methods of constructing the CI: Wald and profile likelihood. In this case, the profile likelihood intervals improve the coverage probability only slightly. In all scenarios the profile likelihood CI of the exact method behaves very satisfactorily. Generally, the approximate likelihood method performed only slightly worse than the exact in the case of small sensitivity (0.60). For the scenarios with large sensitivity (0.93), the coverage percentages are dramatically bad. For some scenarios the coverage dropped down to even 0%. For the intermediate value of the sensitivity, the approximate method is considerably worse than the exact method when the within study sample size is small.

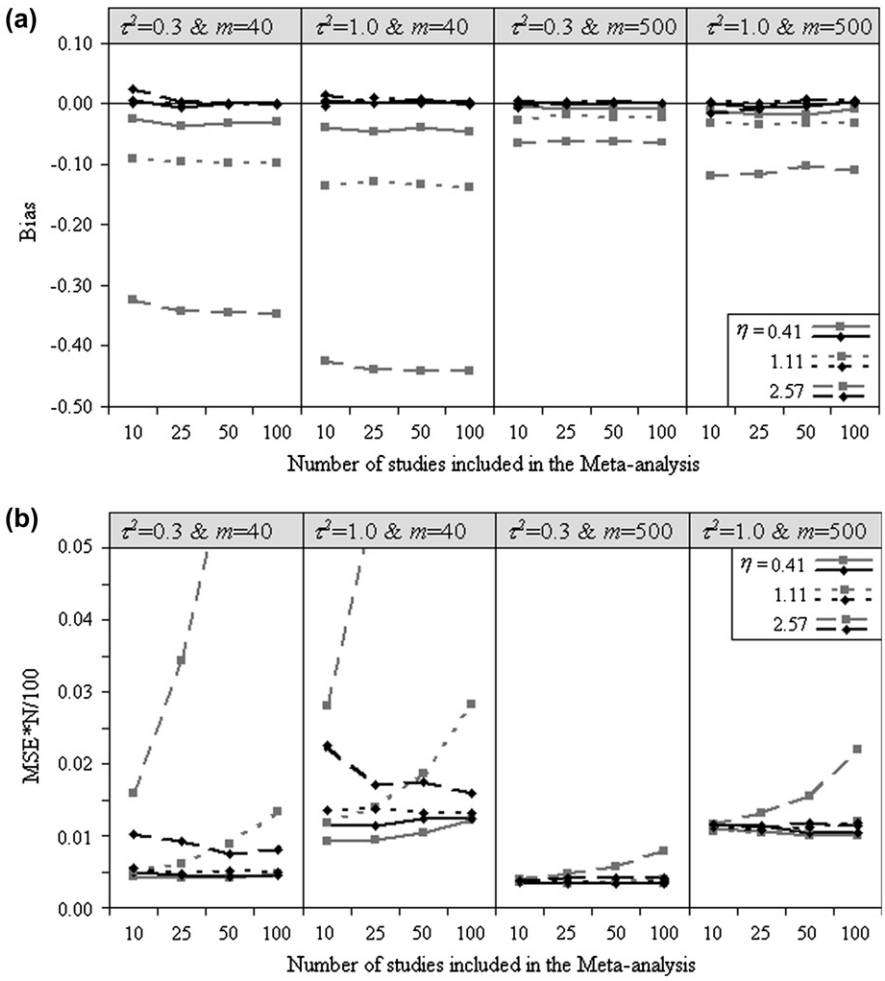

Fig. 1. Simulation results for η. The bias (a) and MSE (b) are given for the approximate likelihood method (gray lines) and the exact likelihood method (black lines). The true between-studies variance and median within-study sample sizes are given at the top of each plot.

Similarly, the estimation of the between-studies variance is investigated using the bias, MSE, and coverage probabilities. The simulation results are presented in Fig. 2 and Table 2. Broadly speaking, the results are analogous to the estimates of the mean logit sensitivity. From Fig. 2a it appears that the exact method always performs better than the approximate method in terms of bias, especially in the scenarios with intermediate and high values of the sensitivity. In scenarios with a larger number of studies (50−100), the exact method is practically unbiased, but when the number of studies included is small and the true between-study variance is large, the exact method underestimates the between-study variance by about 10%. The pattern of the simulation results for the MSEs (Fig. 2b) is also similar to that of the mean logit sensitivity.

The coverage probabilities are presented in Table 2. In some scenarios, the Wald-based coverage probabilities are greater than the nominal level. For $N = 10$, the difference between the exact and approximate methods is not large. In most scenarios, the profile likelihood-based intervals have better coverage than the Wald-based intervals. Overall, the profile likelihood-based interval of the exact method behaves the best. For $N \geqslant 25$, the Wald-based CIs of the approximate method behave worse than those of the exact method, in particular if the within study sample size is small and the value of τ^2 is large. The profile likelihood method improves the coverage probabilities a little bit for the approximate method, but still they are very bad in many scenarios. For the exact method, the Wald-based CIs are already quite good, and they are improved by the profile likelihood method.

In summary, the approximate likelihood method gives a biased estimate for η and τ^2, even a considerably large bias when the true median within study sample size is smaller and median sensitivity is larger. This result is in consistent with the findings of Sidik and Jonkman [31] who reported that τ^2 is underestimated when the approximate method using ML or REML estimation technique is applied. Other authors [5,32] also noted the downward bias in the estimation of the between-studies variance. The coverage probabilities for η and τ^2 of the approximate method are also far from the nominal value in the same region. On the contrary, the exact method gives unbiased estimates for η with a reasonable coverage probability, in particular for

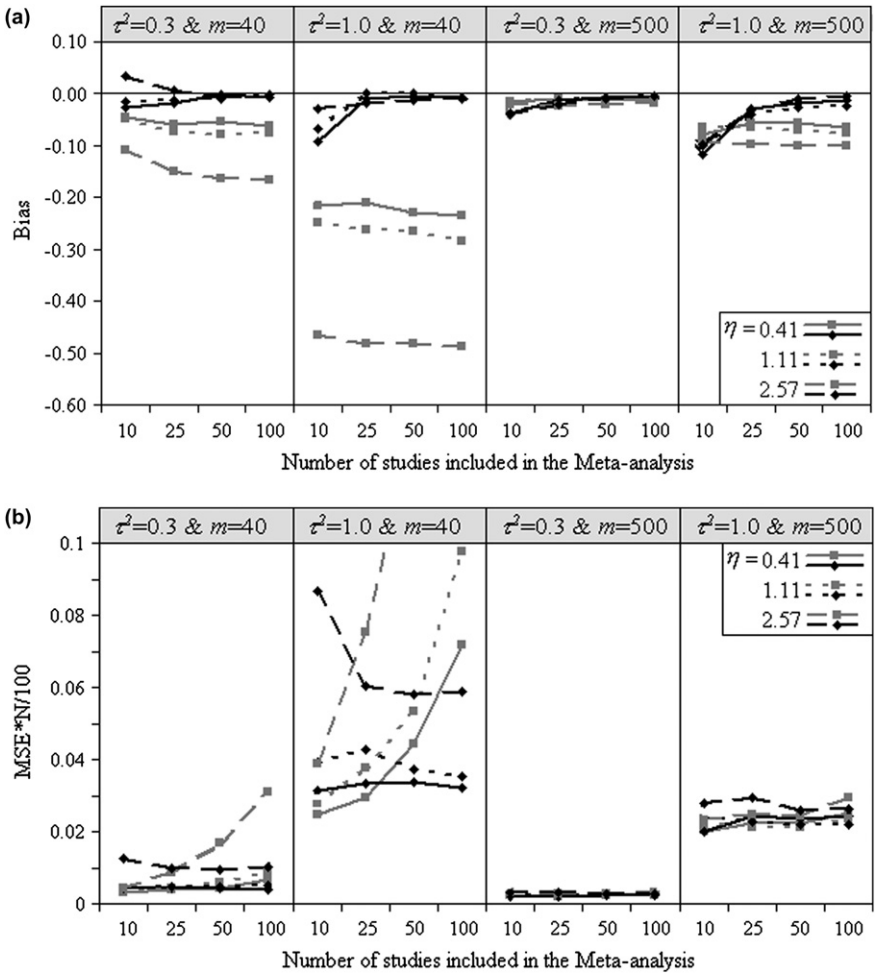

Fig. 2. Simulation results for τ^2. The bias (a) and MSE (b) are given for the approximate likelihood method (gray lines) and the exact likelihood method (black lines). The true between-studies variance and median within-study sample sizes are given at the top of each plot.

the profile likelihood method. It also gives unbiased estimates for τ^2 except when the true between-studies variance is large and the number of studies is small, in that case there is a slight downward bias. This might be due to the fact that the maximum likelihood estimate of the variance parameter is biased for small sample sizes even in the simple independently and identically distributed observations. The coverage probabilities, in particular the ones based on the profile likelihood, are reasonably acceptable.

5. Data example

To illustrate the methods discussed in this article, we reanalyzed the data of a published meta-analysis [33]. Patwardhan et al. [33] present data from 15 studies to assess the operating characteristics of positron emission tomography (PET) by using fluorine 18 fluorodexyglucose (FDG). They performed a literature search in the MEDLINE, CINAHL, and HealthSTAR databases published between 1989 and 2003. Articles were selected if FDG PET was performed with a dedicated scanner and the resolution was specified, if standard criteria were used for the diagnosis of Alzheimer disease, if at least 12 human subjects with Alzheimer disease were enrolled in the study, if clinical diagnosis or histopathologic findings were used as the reference standard, and if sufficient data were provided to construct a 2×2 table. Out of the 15 studies, only nine allowed to construct a 2×2 table with reasonable certainty, and they believed these studies were suitable for meta-analysis. The authors pooled sensitivity and specificity separately using a random effects model in Meta-Test (version 6.0) software. We also reanalyzed only the nine studies. The data and the SAS code are given in the Appendix.

Sensitivity and specificity were pooled separately using both the approximate and exact likelihood methods. Some studies had zero counts and therefore, we added 0.5 in each of the 2×2 tables to avoid undefined values for the approximate likelihood method. The parameter estimates (SE) and the corresponding 95% Wald-type and profile likelihood-based CIs are tabulated in Table 3.

The magnitudes of the parameter estimates from the exact likelihood are larger than from the approximate likelihood method. Back transforming the estimates, the median

Table 2
Coverage probabilities for η and τ^2 using approximate and exact methods

True parameter values				η				τ^2			
				Approximate		Exact		Approximate		Exact	
N[a]	η	τ^2	m[b]	Wald	PL[c]	Wald	PL	Wald	PL	Wald	PL
10	0.41	0.30	40	0.91	0.93	0.91	0.93	1.00	0.94	0.99	0.96
			500	0.91	0.92	0.89	0.92	0.93	0.92	0.90	0.92
		1.00	40	0.93	0.94	0.92	0.94	0.94	0.88	0.94	0.92
			500	0.91	0.93	0.90	0.93	0.93	0.93	0.92	0.94
	1.11	0.30	40	0.90	0.92	0.91	0.93	0.99	0.95	0.98	0.96
			500	0.89	0.91	0.88	0.91	0.92	0.91	0.89	0.91
		1.00	40	0.90	0.92	0.92	0.93	0.96	0.88	0.94	0.91
			500	0.92	0.93	0.91	0.93	0.93	0.92	0.91	0.94
	2.57	0.30	40	0.70	0.79	0.93	0.94	0.84	0.96	0.88	0.96
			500	0.90	0.91	0.90	0.93	0.95	0.92	0.92	0.92
		1.00	40	0.67	0.75	0.90	0.92	0.96	0.80	0.98	0.95
			500	0.90	0.92	0.90	0.92	0.93	0.91	0.90	0.92
25	0.41	0.30	40	0.92	0.93	0.93	0.93	0.97	0.89	0.96	0.93
			500	0.94	0.94	0.94	0.94	0.95	0.94	0.93	0.95
		1.00	40	0.94	0.95	0.95	0.96	0.92	0.89	0.95	0.95
			500	0.93	0.93	0.92	0.93	0.94	0.94	0.93	0.94
	1.11	0.30	40	0.87	0.90	0.94	0.95	0.99	0.87	0.98	0.93
			500	0.92	0.93	0.91	0.92	0.93	0.93	0.93	0.93
		1.00	40	0.88	0.90	0.92	0.93	0.87	0.84	0.92	0.92
			500	0.93	0.94	0.95	0.95	0.93	0.93	0.93	0.94
	2.57	0.30	40	0.34	0.43	0.93	0.94	0.93	0.82	0.97	0.96
			500	0.90	0.91	0.92	0.93	0.94	0.93	0.93	0.94
		1.00	40	0.39	0.46	0.94	0.94	0.80	0.64	0.96	0.94
			500	0.91	0.91	0.93	0.94	0.93	0.93	0.93	0.93
50	0.41	0.30	40	0.94	0.94	0.95	0.95	0.93	0.88	0.96	0.94
			500	0.94	0.94	0.94	0.94	0.93	0.93	0.93	0.94
		1.00	40	0.94	0.94	0.95	0.95	0.82	0.80	0.95	0.95
			500	0.95	0.95	0.94	0.94	0.93	0.93	0.94	0.94
	1.11	0.30	40	0.79	0.81	0.94	0.94	0.91	0.84	0.95	0.94
			500	0.94	0.95	0.94	0.95	0.94	0.94	0.94	0.95
		1.00	40	0.82	0.83	0.95	0.95	0.78	0.74	0.95	0.95
			500	0.94	0.95	0.94	0.95	0.94	0.93	0.95	0.96
	2.57	0.30	40	0.06	0.09	0.96	0.96	0.97	0.62	0.97	0.94
			500	0.87	0.87	0.93	0.93	0.93	0.93	0.94	0.95
		1.00	40	0.10	0.14	0.94	0.94	0.46	0.40	0.95	0.94
			500	0.88	0.88	0.95	0.95	0.92	0.92	0.94	0.94
100	0.41	0.30	40	0.90	0.91	0.94	0.95	0.84	0.81	0.95	0.94
			500	0.95	0.95	0.95	0.95	0.93	0.93	0.94	0.94
		1.00	40	0.92	0.92	0.94	0.94	0.71	0.68	0.95	0.95
			500	0.95	0.95	0.95	0.95	0.92	0.92	0.94	0.94
	1.11	0.30	40	0.66	0.67	0.95	0.95	0.80	0.74	0.95	0.94
			500	0.92	0.92	0.95	0.95	0.93	0.93	0.94	0.94
		1.00	40	0.70	0.71	0.94	0.94	0.58	0.55	0.95	0.95
			500	0.93	0.93	0.93	0.93	0.92	0.92	0.94	0.95
	2.57	0.30	40	0.00	0.00	0.95	0.95	0.64	0.36	0.97	0.95
			500	0.79	0.80	0.95	0.95	0.93	0.92	0.94	0.94
		1.00	40	0.00	0.01	0.95	0.95	0.15	0.13	0.94	0.94
			500	0.82	0.82	0.95	0.95	0.89	0.89	0.96	0.96

[a] N, Number of studies included in the meta-analysis.
[b] m, median within-study sensitivity.
[c] PL, profile likelihood.

sensitivities are 85.1% and 90.0%, and the median specificities are 84.0% and 90.6% for the approximate and exact likelihood methods, respectively. That is, the estimates from the approximate method are lower for the median sensitivity by 4.9% and for the median specificity by 6.6% compared to those from the exact method. The differences between the two methods for the estimates of the between-studies variances are considerable as well. The estimated between-studies variances for the logit sensitivity are 0.48 and 0.97 using the approximate and exact methods, respectively. For

Table 3
Parameter estimates (SE) and Wald-type confidence intervals (CIs) based on the normal approximation and profile likelihood-based confidence interval

		95% CIs	
Parameter	Estimate (SE)	Wald	Profile likelihood
Approximate likelihood method			
logit(sensitivity)	1.74(0.32)	[1.12, 2.37]	[1.12, 2.52]
Sensitivity	0.85(0.04)	[0.75, 0.91]	[0.76, 0.93]
τ^2 logit(sensitivity)	0.48(0.45)	[0.08, 3.00]	[0.01, 2.08]
logit(specificity)	1.66(0.32)	[1.03, 2.29]	[1.03, 2.46]
Specificity	0.84(0.04)	[0.74, 0.91]	[0.74, 0.92]
τ^2 logit(specificity)	0.42(0.47)	[0.05, 3.84]	[0.10, 2.13]
Exact likelihood method			
logit(sensitivity)	2.20(0.44)	[1.34, 3.06]	[1.40, 3.40]
Sensitivity	0.90(0.04)	[0.79, 0.96]	[0.80, 0.97]
τ^2logit(sensitivity)	0.97(0.80)	[0.19, 4.93]	[0.17, 5.28]
logit(specificity)	2.27(0.49)	[1.31, 3.23]	[1.35, 3.59]
Specificity	0.91(0.04)	[0.79, 0.96]	[0.79, 0.97]
τ^2logit(specificity)	1.23(1.01)	[0.24, 6.17]	[0.21, 6.59]

the logit specificity, the estimates are 0.42 and 1.23, respectively. Comparing the CIs based on Wald and profile likelihood for logit sensitivity, η and logit specificity, ξ, the profile likelihood intervals are wider, which is expected because the profile likelihood takes into account the uncertainty in the between-studies variance. Also the CIs from the exact method are wider compared to the approximate methods for the variance parameters. In summary, the parameter estimates from the practical data example follow the same pattern as the simulation study, that is, the approximate method gives lower estimates compared to the exact method for a given parameter. Furthermore, the example shows that the differences between the approximate and exact methods in practice are not negligible.

6. Discussion

In numerous medical articles sensitivities or specificities, or more generally proportions are analyzed, nowadays almost invariably with the DerSimonian and Laird [6] random effects model. This model uses a normal distribution for the logit transformed true probabilities. Alternatively, one could assume a beta distribution for the true probabilities. Then the model can be fitted in a statistical package such as EGRET. However, this model is not used in practice, may be due to the fact that many statistical packages allow only a normal distribution for the random effects. We restricted our study to the standard method of DerSimonian and Laird. Instead of the usual logit transformation of the observed data, other transformations such as the probit, log(−log), arcsine could be used and implemented in the same program. We do not expect that the results of this paper would change substantially if another transformation was used and approximated by a normal distribution. The reason is that there will always be a correlation between the estimate and the within-study variance as they are determined by the same parameter that, if not accounted for in the model, may lead to biased parameter estimates [11]. Hence, we restricted the study to the standard logit transformation.

In this paper, we compared the use of the approximate normal within-study likelihood that is used in practice with the alternative exact binomial likelihood. Calculation of the exact binomial likelihood involves an approximation of the integral. In NLMIXED, the method of Gaussian quadrature is used, with the number of quadrature points to be specified by the user or automatically by SAS. The larger that number is chosen, the better the approximation, but at the cost of more computational time. For example, Carlin et al. [34] have shown that for binary outcome longitudinal data, a reasonably large number of quadrature points (i.e., 20) is required to ensure convergence on model parameter estimates. In our data example, to study the impact of the number of quadrature points we fitted the model for varying number of quadrature points. It turned out the estimates (SE) of sensitivity and specificity did not change for a number of quadrature points greater than or equal to 10 and 15, respectively. We used 20 quadrature points for our simulation study.

Our simulations have shown that the approximate method yields biased estimates for the overall sensitivity or specificity as well as for the between-studies variance, the bias being especially considerable in cases with smaller within study sample sizes, larger between-studies variance, and larger values of the overall sensitivity, as is frequently reported in diagnostic tests. In these cases also the coverage percentages of the CIs are far off the nominal values. Considering possible bias in analyzing log ORs is beyond the scope of this paper. However, we expect that in analyzing log ORs the bias might be less of a problem, because it might be at least partly canceled out. The bad performance of the approximate method is mainly caused by the fact that it does not adjust for the correlation between the estimate of the sensitivity or specificity and its SE. Also the addition of 0.5 when there is a zero count adds to the bias [14,35,36]. Although it was mentioned in the literature that the standard random effects method could be biased when the parameter to be estimated is a proportion [5,11,12], it is still generally used. A possible explanation is that there were no practically feasible methods available that address this bias. However, the implementation of procedures for the GLMM in standard packages has made it practically feasible nowadays to use the exact within-study distribution of the estimated sensitivity or specificity. To carry out the exact method, the sample size and the number of positive test results are needed. In practice, these quantities will always be available, though sometimes indirectly. For instance, if only the estimated proportion and its SE are given, the sample size and number of positive results can be easily calculated. In this paper, we have compared the exact method with the standard method through an extensive simulation study. We accounted for possibly important

factors such as the number of studies included in the meta-analysis, the magnitude of the mean logit sensitivity, the between-studies variance, and the within-studies sample size. We have shown that in all scenarios studied the exact method outperformed the approximate method with respect to bias of the estimated mean logit sensitivity and coverage percentages of the corresponding CI. The exact method yielded unbiased estimates of the logit sensitivity in all scenarios with reasonable coverage percentages of the Wald CI, with the exception of the scenarios where the number of studies in the meta-analysis is small. Mostly the coverage probabilities were slightly lower than the nominal value. This could be due to the fact that the standard Wald method does not adjust for the between-studies variance being estimated. A profile likelihood-based CI that allows for the uncertainty in the estimated between-studies variance appeared to improve the coverage percentage to an acceptable level close enough to the nominal level. The Wald-type CI can be automatically done in SAS. For the profile likelihood method, some extra programming is needed. An SAS macro is available from the authors on request.

Concerning estimation of the between-studies variance, we have shown that the approximate method yielded underestimates in all scenarios studied, with bias of considerable magnitude and coverage probabilities far from the nominal level in many cases. The exact method always outperformed the approximate method, although there was some bias left for cases where the number of studies in the meta-analysis was small and a coverage percentage of the Wald-type CI a bit less than 0.90 for some scenarios that could be corrected using profile likelihood. In the spirit of REML estimation, a possible improvement might be to multiply the estimate by $k/(k-1)$, where k is the number of studies in the meta-analysis. Another possibility might be the use of a bootstrapping technique similar to that used by Turner et al. [15], which is directly available in the MLwiN software [18].

We did some further simulations when the true parameters were simulated from a skewed distribution where most of the proportions are close to 1. We used Fernandez and Steel's [37] approach to introduce skewness into a unimodal normal distribution. In most cases, the exact method outperformed the approximate method, with only slight bias ($\cong 0.10$) for the exact method left for scenarios with highly skewed true distribution and small within-study size. The coverage probability from the exact method was also better than that from the approximate method in most cases and close to the nominal level, especially when we used profile likelihood-based CI (results not presented and available from the authors on request).

Though we focused in this paper on sensitivities or specificities, the results apply more generally to meta-analyzing proportions, such as prevalences or incidences of a disease. The proportions can also be corrected for important covariates, using the same GLMM programs.

In a situation when two end points, for example, sensitivity and specificity, are presented in a study and when there is a need to incorporate the correlation that might exist between these two measures, a bivariate meta-analysis approach can be used [8,10,25]. Reitsma et al. [10] assume the within-study error distribution of sensitivity and specificity to follow a normal distribution. However, an exact binomial likelihood can also be implemented in GLMM programs [38], for example, SAS NLMIXED or S-Plus/R nlme. Although we did no simulations for the bivariate model, it is very likely that the results of the univariate meta-analysis carry over to the bivariate case.

In this article, the models were fitted using classical likelihood methods. An alternative would be to use a Bayesian hierarchical modeling approach [34,39] that can be carried out using the publicly available software Win BUGS [40].

Our overall conclusion from this paper is that in many cases the standard approximate method falls short and that the exact method should be used, preferably accompanied by profile likelihood-based CIs, whenever that is feasible in practice.

Appendix

In this appendix the SAS syntax is given to estimate the parameters using approximate and exact likelihood methods for the meta-analysis data used in Section 5. The data are given in the table below. The SAS code for the MIXED and NLMIXED procedures is given to pool sensitivity, but it can also be used to pool specificity

Study	TP(x_i)	FN	FP	TN	m_i	logitsens	est
15	33	6	5	35	39	1.640	0.184
17	18	6	5	10	24	1.046	0.208
19	20	13	0	41	33	0.418	0.123
20	19	0	0	19	19	3.664	2.051
22	44	6	10	19	50	1.924	0.176
24	18	3	1	9	21	1.665	0.340
25	27	1	4	21	28	2.909	0.703
29	21	0	1	9	21	3.761	2.047
30	18	1	1	20	19	2.512	0.721

The meaning of the variables that are used in the table and SAS code below are the following:

Study, a number given for a study.
TP, true positive; FN, false negative; FP, false positive; TN, true negative.
x_i, number of patients with true positive test result.
m_i, within-study sample size in the diseased group.
logitsens, observed logit sensitivity $(=\ln((x_i+0.5)/(m_i-x_i+0.5)))$.
est[1], estimated within-study variance of logit sensitivity $= 1/(x_i+0.5)+1/(m_i-x_i+0.5)$.

0.5 is added in each of the 2×2 tables when we calculate logitsens and est to avoid undefined values.

[1] This is the prescribed name by SAS of the variable that contains the variances.

```
/* Approximate likelihood method using SAS procedure MIXED*/
proc mixed data = d method = REML;
            class  study;
            model logitsens = / intercept Solution cl df=1000;
/* df=1000 is specified to get Wald type confidence interval instead of
the t */
            random intercept /  subject = study ;
            repeated /group = study;
/*dataest is the name of the data set that contains only the variable
called est and 10 lines. The first value is a starting value for the
between studies variance. The next nine values are the estimated within
study variances (=1/(xi+0.5)+ 1/(mi-xi+0.5)). eqcons is used to specify
that the within study variance are assumed to be known */
            parms / parmsdata = dataest eqcons= 2 to 10;
run;
/* The Exact approach using the SAS procedure NLMIXED*/
proc nlmixed data=d df = 1000;
      parms mtlnsens=2.0   vtlnsens=0.8;        /*Initial values*/
      pi = 1/(1+exp(-tlnsens));     /*tlnsens  =  is  the  unknown  true
logit sensitivity*/
      model xi~binomial(mi,pi);
      random tlnsens ~ normal(mtlnsens , vtlnsens) subject=study;
run;
```

References

[1] Irwig L, Tosteson ANA, Gatsonis C, Lau J, Colditz G, Chalmers TC, et al. Guidelines for meta-analyses evaluating diagnostic tests. Ann Intern Med 1994;120:667–76.

[2] Lijmer JG, Bossuyt PMM, Heisterkamp SH. Exploring sources of heterogeneity in systematic reviews of diagnostic tests. Stat Med 2002;21:1525–37.

[3] Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted. Stat Med 2002;21:1559–73.

[4] Hardy RJ, Thompson SG. Detecting and describing heterogeneity in meta-analysis. Stat Med 1998;17:841–56.

[5] Berkey CS, Hoaglin DC, Mosteller F, Colditz GA. A random-effect regression model for meta-analysis. Stat Med 1995;4:395–411.

[6] DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986;7:177–88.

[7] Arends LR, Voko Z, Stijnen T. Combining multiple outcome measures in meta-analysis: an application. Stat Med 2003;22:1335–53.

[8] Van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. Stat Med 2002;21:589–624.

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

[10] Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol 2005;58:982–90.

[11] Chang BH, Waternaux C, Lipsitz S. Meta-analysis of binary data: which study variance estimate to use? Stat Med 2001;20:1947–56.

[12] Platt RW, Leroux BG, Breslow N. Generalized linear mixed models for meta-analysis. Stat med 1999;14:395–411.

[13] Cox DR. The analysis of binary data. London: Methuen; 1970.

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

[15] Turner RM, Omar RZ, Yang M, Goldstein H, Thompson SG. A multilevel model framework for meta-analysis of clinical trials with binary outcomes. Stat Med 2000;19:3417–32.

[16] Knapp G, Biggerstaff BJ, Hartung J. Assessing the amount of heterogeneity in random-effects meta-analysis. Biom J 2006;48:271–85.

[17] SAS Institute Inc.. SAS/STAT 9.1 user's guide. Cary, NC: SAS Institute Inc.; 2004.

[18] Rasbash J, Steele F, Browne W, Prosse B. A user's guide to MLwiN. Centre for Multilevel Modelling, version 2.0. London: Institute of Education; 2004.

19 Mol BWJ, Lijmer JG, Ankum WM, van der Veen F, Bossuyt PMM. The accuracy of single serum progesterone measurement in the diagnosis of ectopic pregnancy: a meta-analysis. Human Reproduction 1998;13:3220−32.
[20] Cruciani M, Nardi S, Malena M, Bosco O, Serpelloni G, Mengoli C. Systematic review of the accuracy of the ParaSightTM-F test in the diagnosis of Plasmodium falciparum malaria. Med Sci Monit 2004;10:MT81−8.
[21] Bipat S, Glas AS, Slors FJM, Zwinderman AH, Bossuyt PMM, Stoker J. Rectal cancer: local staging and assessment of lymph node involvement with endoluminal US, CT, and MR imaging—a meta-analysis. Radiology 2004;232:773−83.
[22] Stengel D, Bauwens K, Rademacher G, Mutze S, Ekkernkamp A. Association between compliance with methodological standards of diagnostic research and reported test accuracy: meta-analysis of focused assessment of US for trauma. Radiology 2005;236:102−11.
[23] Cruciani M, Marcati P, Malena M, Bosco O, Serpelloni G, Mengoli C. Meta-analysis of diagnostic procedures for Pneumocystis carinii Pneumonia in HIV-1-infected patients. Eur Respir J 2002;20: 982−9.
[24] Cox DR, Hinkley DV. Theoretical statistics. London: Chapman and Hall; 1974.
[25] van Houwelingen HC, Zwinderman KH, Stijnen T. A bivariate approach to meta-analysis. Stat Med 1993;12:2273−84.
[26] Brockwell SE, Gordon R. A comparison of statistical methods for meta-analysis. Stat Med 2001;20:825−40.
[27] Sidik K. Simple heterogeneity variance estimation for meta-analysis. Appl Stat 2005;54:367−84.
[28] Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects. Stat Med 1996;15:619−29.
[29] Roger JH, Kenward MG. Repeated measures using proc mixed instead of proc glm. In: Proceedings of the First Annual South-East SAS Users Group conference. Cary NC: SAS Institute; 1993. p. 199−208.
[30] Welham SJ, Thompson R. Likelihood ratio tests for fixed model terms using residual maximum likelihood. J R Statist Soc B 1997;59:701−14.
[31] Sidik K, Jonkman JN. A comparison of heterogeneity variance estimators in combining results of studies. Stat Med. 2007;26:1964−81.
[32] Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. Stat Med 1999;18:2693−708.
[33] Patwardhan MB, McCrory DC, Matchar DB, Samsa GP, Rutschmann OT. Alzheimer disease: operating characteristics of PET a meta-analysis. Radiology 2004;231:73−80.
[34] Carlin JB, Wolfe R. A case study on the choice, interpretation and checking of multilevel models for longitudinal binary outcomes. Biostatistics 2001;2:397−416.
[35] Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. Stat Med 2004;23:1351−75.
[36] Bradburn MJ, Deeks JJ, Berlin JA, Localio AR. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. Stat Med 2007;26:53−77.
[37] Fernandez C, Steel MFG. On Bayesian modeling of fat tails and skewness. J Am Statist Assoc 1998;93:359−71.
[38] Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JAC. A unification of models for meta-analysis of diagnostic accuracy studies. Biostatistics 2006;1:1−21.
[39] Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approach to random-effects meta-analysis: a comparative study. Stat Med 1995;14: 2685−99.
[40] Spiegelhalter D, Thomas A, Best N, Lunn D. WinBUGS user manual, version 1.4.1. Cambridge: MRC Biostatistics unit; 2004. Available at http://www.mrc-bsu.cam.ac.uk/bugs.