Research Article

Statistics in Medicine

Received 13 September 2010, Accepted 31 October 2011 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4474

On the covariance of two correlated log-odds ratios

Pantelis G. Bagos[*†]

In many applications two correlated estimates of an effect size need to be considered simultaneously to be combined or compared. Apparently, there is a need for calculating their covariance, which however requires access to the individual data that may not be available to a researcher performing the analysis. We present a simple and efficient method for calculating the covariance of two correlated log-odds ratios. The method is very simple, is based on the well-known large sample approximations, can be applied using only data that are available in the published reports and more importantly, is very general, because it is shown to encompass several previously derived estimates (multiple outcomes, multiple treatments, dose–response models, mutually exclusive outcomes, genetic association studies) as special cases. By encompassing the previous approaches in a unified framework, the method allows easily deriving estimates for the covariance concerning problems that were not easy to be obtained otherwise. We show that the method can be used to derive the covariance of log-odds ratios from matched and unmatched case-control studies that use the same cases, a situation that has been addressed in the past only using individual data. Future applications of the method are discussed.

Keywords: odds ratio; covariance; epidemiology; delta method; contingency tables; correlated outcomes

1. Introduction

In many applications in epidemiological and clinical studies, two correlated estimates of the effect size, usually the odds ratio (OR), need to be considered simultaneously. These situations may occur when we simply want to perform a comparison of the two estimates and construct a confidence interval for the difference or some other function of the estimates [1], when we are interested in pooling the estimates to obtain a combined estimate usually in a meta-analysis setting [2], or when we are interested in performing a joint modeling in a form of multivariate meta-analysis [3]. In any case, if the estimates are uncorrelated as is usually the case, the situation is straightforward. However, it gets complicated when the estimates are stochastically dependent (correlated) in which case the within-study correlation and the associated covariance needs to be calculated. This however, in most situations, requires access to the individual data that may not be available to a researcher performing the analysis.

Several approaches have been proposed to deal with the problem of correlated estimates [4]. However, the general case includes a within-study correlation that needs to be computed using individual data [3, 5]. Special cases in which the correlation can be computed analytically using solely the observed summary counts, include studies with multiple treatment arms [6] or multiple outcomes [4], clinical trials with mutually exclusive outcomes [7], dose–response models [8–10], and multivariate modeling of log-odds ratios in genetic association studies [11]. Recently, the problem has also attracted increased interest in genetic association studies, notably in genome-wide association studies that share overlapping sets of controls [12, 13] and in multipoint meta-analysis of genetic association studies [14]. In other more general cases however, the correlation cannot be calculated and simulation techniques have been proposed [15, 16]. The problem of correlated estimates have been studied in detail in the context of meta-analysis and it is generally accepted that by ignoring or approximating the correlation leads to biased

Department of Computer Science and Biomedical Informatics, University of Central Greece, Papasiopoulou 2-4, Lamia GR35100, Greece

**Correspondence to: Pantelis G. Bagos, Department of Computer Science and Biomedical Informatics, University of Central Greece, Papasiopoulou 2-4, Lamia GR35100, Greece.*

†E-mail: pbagos@ucg.gr

estimates of the variance of the overall effect [17, 18]. As a last resort, an alternative model for bivariate meta-analysis was recently proposed requiring as input only the two (correlated) outcomes; this model does not separate the between-studies and the within studies correlation, but instead it computes a single parameter for the 'overall' correlation [19].

Similar problems concerning the correlation of two estimates arise when combining the results of matched and unmatched case-control studies that use the same group of cases [20]. The investigators choose to perform such analyses in an effort to control for different sources of confounding and simultaneously increase the power [21]. Although methods for combining such studies have been proposed for years, they all utilize individual data and require that the study populations are independent (i.e., the subjects are not overlapping) [22–24]. A characteristic example of this kind comes from the Multiple Environmental and Genetic Assessment of Risk Factors for Venous Thrombosis Study [25], where the primary interest was the effect of smoking. In this study, the cases were patients diagnosed with venous thrombosis between 1999 and 2004 (3986 in total). Of these, 2286 had an individually matched control, which was selected among their partners. A second population-based control group with the same age and sex distribution as the cases was acquired using random digit dialling (2612 subjects). Clearly, in such a situation the estimates are correlated because the same individuals (cases) are used in calculating both estimates.

A completely analogous situation is encountered in combining the results of family-based and population-based case-control association studies in which the same cases are used in both analyses. In genetics, the family-based design using the transmission disequilibrium test (TDT) is the direct analogue of the matched analysis, whereas the population-based allelic odds ratio corresponds to the unmatched analysis [26]. A typical example of studies of this kind, is a genetic association study that investigated the effects of CTLA4 polymorphisms in the development of autoimmune diseases [27]. The authors examined (among others) the effects of four CTLA4 polymorphisms (MH30*G/C, −1147*C/T, CT60*G/A and JO37_3*G/A) in Type 1 diabetes mellitus (T1DM), contrasting 350 cases against 900 unrelated controls in the population-based case-control analysis, and 218 out of the 350 cases against their parents (for which available data existed) in the family-based analysis. Similar situations are also encountered in genome-wide association studies (GWAS), in which the TDT design is also applicable [28]. Usually in GWAS the investigators are interested in replicating their initial significant findings (stage 1) and a combined analysis of the results of both stages (stage 1 and stage 2) has been shown to be more powerful than relying on the replication stage alone [29]. Because of increased cost of genotyping and the difficulty of obtaining large samples, several GWAS that use the family-based approach perform a kind of replication stage contrasting the already genotyped cases against a dataset of randomly chosen controls. A typical example is the GWAS for identification of susceptibility loci for multiple sclerosis [30], in which 931 family trios were initially genotyped for 334,923 single nucleotide polymorphisms (SNPs) and analysed using the TDT and subsequently, the same 931 cases were contrasted against 2431 population controls (1475 controls drawn from the Wellcome Trust Case Control Consortium and 956 controls provided by the National Institute of Mental Health) in a traditional case-control analysis.

For the aforementioned study designs (epidemiological studies and genetic association studies) in which cases are overlapping, methods for performing a combined analysis have already been proposed in the literature [20, 31–37]. However, they all require access to individual data and thus, a combined estimate cannot be obtained using only data that are available in the published reports. Consequently, the data cannot be pooled in a meta-analysis of the literature, because this would require an estimate for the covariance that should be calculated using only summary data.

In this work we propose a general framework that can be used for computing the covariance of two correlated log-odds ratios. The method is very simple and straightforward and it can be applied using only data that are available in the published reports. More importantly, the method is very general because it is shown to encompass all the previously mentioned situations as special cases. The method is based on a contingency table formulation and uses standard asymptotic theory concerning the large sample properties of the log-odds ratio.

The manuscript is organized as follows: in Section 2 the method is introduced and a general expression for the covariance is given. In Section 3, we derive the covariance between log-odds ratios obtained from matched and unmatched case-control studies in which the same cases are used in both analyses. In Section 4, we show that the general method encompasses all the previously mentioned situations (multiple outcomes, multiple treatments, genetic association studies) as special cases. In Section 5, the method is applied to the three datasets from the motivating examples outlined in the previous paragraphs. Finally, in Section 6 the findings of this study and the implications for future research are discussed.

Statistics in Medicine

2. The general framework

Let Y, X_1, and X_2 denote three categorical random variables with two levels (i.e., 0, 1) that are used to classify n individuals. Usually, in a retrospective case-control study, Y denotes the case-control status and X_1, X_2 denote the two different exposures. In clinical trials (or in prospective observational studies), Y will denote the disease/nondisease outcome whereas X_1, X_2 may denote two different treatments (or exposures). Alternatively, Y may denote the single treatment whereas X_1, X_2 will denote the two alternative outcomes.

In the general case, the data would be presented in the form of a three-dimensional ($2 \times 2 \times 2$) contingency table (Table I), where we denote the counts as n_{ijk} with $i, j, k \in \{0, 1\}$. We further denote $n_{ij+} = \sum_k n_{ijk} = n_{ij1} + n_{ij0}$ and $n_{i+k} = \sum_j n_{ijk} = n_{i0k} + n_{i1k}$. Table I, which contains the full distribution of Y, X_1, and X_2, is usually referred to as a *partial contingency table* and thus, the two tables displayed in Table II can be considered as *marginal contingency tables* [38]. Such tables can arise in many practical applications in observational and clinical studies, depending on the assumed sampling scheme [38]. For instance, the Poisson sampling scheme assumes that the sample counts n_{ijk} in the contingency table are independent random variables and so is their sum $n = \sum_{ijk} n_{ijk}$, the multinomial sampling scheme conditions on the total sample size n, which is now considered fixed, whereas the independent binomial sampling scheme assumes that one of the marginal totals are fixed in advance (n_{ij+}, n_{i+k} or n_{+jk}) [38]. Finally, the hypergeometric sampling scheme assumes that all marginals are fixed by design [38].

Intuitively, the independent binomial sampling scheme is suitable for the majority of the situations that we encounter here. However, it is well known in the literature that all sampling schemes lead to the same inferences for the measures of association (such as the odds-ratio), because those inferences condition on marginal totals that contain as a subset the naturally fixed totals and additionally, the parameters of interest are not quantities that are fixed under some sampling designs but not under others [38, 39]. Thus, the assumed sampling scheme is usually chosen on the grounds of convenience regarding the calculations. In this work, following others (i.e., [8, 39]), we assume a Poisson sampling scheme, even though the same results could have been derived using the independent binomial sampling scheme.

Table I. The full table describing the association of binary variables Y, X_1, and X_2. The interpretation of the variables in terms of the sampling design (i.e., prospective, retrospective) is discussed in the text. We denote the counts as n_{ijk}, with i being the indicator for Y, j the indicator for X_1 and k the indicator for X_2. In the general case we assume that $n_{ij+} = n_{ij1} + n_{ij0}$ and $n_{i+k} = n_{i0k} + n_{i1k}$. Special cases are discussed in the text.

		X_1			
		$X_1 = 1$		$X_1 = 0$	
		X_2			
		$X_2 = 1$	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$
Y	$Y = 1$	n_{111}	n_{110}	n_{101}	n_{100}
	$Y = 0$	n_{011}	n_{010}	n_{001}	n_{000}

Table II. The two marginal 2×2 contingency tables used for classifying individuals according to Y, X_1 status and Y, X_2 respectively. The interpretation of the variables in terms of the study design (i.e., prospective, retrospective) is discussed in the text. In some of the special cases discussed in the text, some pairs of cells coincide. We denote the counts as n_{ijk}, with i being the indicator for Y, j the indicator for X_1 and k the indicator for X_2. Thus, the interior cells of the table consist of the marginals of Table I. (i.e., $n_{ij+} = n_{ij1} + n_{ij0}$ and $n_{i+k} = n_{i0k} + n_{i1k}$).

		X_1		X_2	
		$X_1 = 1$	$X_1 = 0$	$X_2 = 1$	$X_2 = 0$
Y	$Y = 1$	n_{11+}	n_{10+}	n_{1+1}	n_{1+0}
	$Y = 0$	n_{01+}	n_{00+}	n_{0+1}	n_{0+0}

By definition the two marginal odds ratios from Table II, the OR for X_1 and the OR for X_2 would be given by

$$\begin{aligned} OR_{YX_1} &= \frac{P\,(Y=1|X_1=1)\,P\,(Y=0|X_1=0)}{P\,(Y=0|X_1=1)\,P\,(Y=1|X_1=0)} \\ OR_{YX_2} &= \frac{P\,(Y=1|X_2=1)\,P\,(Y=0|X_2=0)}{P\,(Y=0|X_2=1)\,P\,(Y=1|X_2=0)} \end{aligned} \tag{1}$$

Usually, we work with the logarithms of these ORs, here denoted β_1 and β_2. Then, the sample estimates of these log-odds ratios, which under the Poisson sampling scheme discussed above can be regarded as contrasts between the log counts [8, 38], are given by

$$\begin{aligned} \hat{\beta}_1 &= \log \widehat{OR}_{YX_1} = \log\left(\frac{n_{11+}n_{00+}}{n_{01+}n_{10+}}\right) \\ \hat{\beta}_2 &= \log \widehat{OR}_{YX_2} = \log\left(\frac{n_{1+1}n_{0+0}}{n_{0+1}n_{1+0}}\right) \end{aligned} \tag{2}$$

These log-odds ratios are asymptotically normally distributed [40] and asymptotic estimates of their variances are easily computed by treating the sample counts n_{ijk} as independent Poisson variables, with $E[n_{ijk}] = \text{var}[n_{ijk}] = m_{ijk}$ [8, 38] (see Appendix A):

$$\begin{aligned} \widehat{\text{var}}\left(\hat{\beta}_1\right) &= \frac{1}{n_{11+}} + \frac{1}{n_{00+}} + \frac{1}{n_{01+}} + \frac{1}{n_{10+}} \\ \widehat{\text{var}}\left(\hat{\beta}_2\right) &= \frac{1}{n_{1+1}} + \frac{1}{n_{0+0}} + \frac{1}{n_{0+1}} + \frac{1}{n_{1+0}} \end{aligned} \tag{3}$$

The variance estimate given in Equation (3) is also known as the Woolf's estimator and it was originally derived using the large sample normal approximation to the binomial distribution [40]. Even though it is based on a large sample approximation, it performs quite well even for small sample sizes [41], justifying its popularity. In the general case described in Tables I and II, $\hat{\beta}_1$ and $\hat{\beta}_2$ are usually stochastically dependent because they are computed using overlapping counts (i.e., $n_{11+} = n_{111} + n_{110}$, $n_{1+1} = n_{111} + n_{101}$ and so on). This is easily understood if we consider a Poisson sampling scheme where S, U, W are independent Poisson random variables. Then, the linear combinations $Z = S + U$ and $V = S + W$ are also Poisson random variables following a bivariate Poisson distribution [42] with covariance equal to var$[S]$ (Appendix A).

The main purpose of this work is to derive an estimate for the covariance of $\hat{\beta}_1$ and $\hat{\beta}_2$ and express it using solely the observed counts of the contingency tables (Tables I and II). In Appendix A we show that the estimate of the covariance is given by

$$\widehat{\text{cov}}\left(\hat{\beta}_1, \hat{\beta}_2\right) = \sum_i \sum_j \sum_k (-1)^{j-k} \left(\frac{n_{ijk}}{n_{ij+}n_{i+k}}\right) \tag{4}$$

It is interesting to note at this point that the covariance depends only on the observed counts n_{ijk}, n_{ij+} and n_{i+k}. If n_{ij+}or n_{i+k} becomes zero, a simple correction can be employed adding $c = 1/2$ similar to what should have been made in Equation (3) in a similar situation. It is clear that calculating the covariance of Equation (4) requires knowledge of the full distribution of the counts in Table I (i.e., n_{ijk}). However, if we recall that n_{ij+}, n_{i+k} and n_{+jk} are the minimal sufficient statistics for obtaining the maximum likelihood (ML) estimates of the $2 \times 2 \times 2$ contingency table in the case of no three-factor interaction [38], we realize that the covariance can theoretically be calculated in certain cases even when the n_{ijk} are not directly observed, provided that we assume no three-way interaction.

Finally, we have to note that from Equation (4) we can deduce that $\widehat{\text{cov}}\left(\hat{\beta}_1, \hat{\beta}_2\right) = 0$ if and only if

$$\sum_i \sum_j \sum_{k=j} \left(\frac{n_{ijk}}{n_{ij+}n_{i+k}}\right) = \sum_i \sum_j \sum_{k\neq j} \left(\frac{n_{ijk}}{n_{ij+}n_{i+k}}\right) \tag{5}$$

3. Combining matched and unmatched case-control studies

We now return to the motivating example discussed in the introduction, that is, the problem of combining matched and unmatched case-control studies that use the same group of cases [20]. There is a considerable body of literature discussing the advantages and disadvantages of matched and unmatched designs [43–45]. As we discussed, the investigators choose to perform a matched and an unmatched case-control study within the same population using the same (or largely overlapping) set of cases, in an effort to control for different sources of confounding and simultaneously increase the power [21].

Recently, the particular problem has been addressed by le Cessie *et al.*, who developed a general method that uses the bootstrap or the sandwich estimator of variance [20]. This method is very general and can be applied to numerous study designs, but it requires access to individual data that may not be available to a meta-analyst. Thus, there is clearly a need for deriving analytical expressions for the covariance of the estimates using data available in the published reports. In matched case-control studies [46,47], the odds ratio is estimated by the ratio of discordant pairs of Table III. Then, the sample estimate for the logarithm of the odds-ratio is denoted

$$\hat{\beta}_2 = \log\left(\frac{b}{c}\right) \tag{6}$$

Consequently, an estimate for the asymptotic variance is given by

$$\widehat{\text{var}}\left(\hat{\beta}_2\right) = \frac{1}{b} + \frac{1}{c} \tag{7}$$

Then, the covariance between the two estimated log-odds ratios will be given by (we omit the subscript $_+$ for simplicity)

$$\begin{aligned}
\text{cov}\left(\hat{\beta}_1, \hat{\beta}_2\right) &= \text{cov}\left[\log\left(\frac{n_{11}n_{00}}{n_{01}n_{10}}\right), \log\left(\frac{b}{c}\right)\right] \\
&= \text{cov}\left[\log\left(n_{11}n_{00}\right) - \log\left(n_{01}n_{10}\right), \log\left(b\right) - \log\left(c\right)\right] \\
&= \text{cov}\left[\log\left(n_{11}\right), \log\left(b\right)\right] + \text{cov}\left[\log\left(n_{10}\right), \log\left(c\right)\right] \\
&= \frac{1}{m_{11}} + \frac{1}{m_{10}}
\end{aligned} \tag{8}$$

which, after replacing the expectations by the observed values, yields the estimate

$$\widehat{\text{cov}}\left(\hat{\beta}_1, \hat{\beta}_2\right) = \frac{1}{n_{11}} + \frac{1}{n_{10}} \tag{9}$$

The last equality is derived by applying Equation (4) (i.e., using Equation (A8) from Appendix A) with $w = 0$ and noticing also that $n_{11} = a + b$ and $n_{10} = c + d$ (because the same cases are used). The estimate of Equation (9) is simple, intuitive and more easily calculated compared with the ones given by le Cessie *et al.* [20] and additionally (and perhaps more importantly), it does not require

Table III. Presentation of data in an individually matched case-control study. The OR is given by the ratio of the discordant pairs (b/c). Similar rearrangements are performed in the TDT studies where the transmitted alleles are contrasted against the non-transmitted ones. When a population based case-control study is performed using the same cases (Y denotes the case control status and X_1 the exposure), we will have to consider this table along with the left-hand side of Table II and omit the subscript $_+$. Then we will have $n_{11} = a + b$ and $n_{10} = c + d$.

		Controls	
		Exposed	Not-exposed
Cases	Exposed	a	b
	Not-exposed	c	d

access to individual data. Moreover, Equation (9) applies also to situations where some of the cases (denoted by n_{11-} and n_{10-}) did not participate for some reason in the matched analysis. In such case, $n_{11} = n_{11*} + n_{11-} = a + b + n_{11-}$ and $n_{10} = n_{10*} + n_{10-} = c + d + n_{10}$ and thus, Equation (10) remains unchanged. Similar arguments hold for the reverse case (i.e., when some of the cases participate only in matched analysis and not in the unmatched one).

As we already discussed, a completely analogous situation is encountered when combining the results of family-based and population-based case-control association studies. It is obvious, that Equations (6)–(9) are directly applicable in this setting. The comparison of family-based and population-based case-control designs has attracted interest in the past [48] and pooling the estimates has been illustrated using summary data [49] and using individual data methods [31–35, 37]. The method of le Cessie *et al*. [20], being quite general is also applicable in this situation [50]. However, all the available methods, with the exception of the simple method proposed by Kazeem and Farall [49], which is applicable only when the cases are not overlapping (i.e., when the estimates are independent), require access to individual data. Thus, it is clear that the simple estimate given in Equation (9) is also applicable in combining odds ratios derived from TDT and traditional case-control association studies when the cases are shared between the two study designs.

4. Some other special cases known from the literature

In this section we show that some other well-known examples from the literature can be derived easily as special cases of Equation (4). For instance, in cases where we have two (or more) categories of exposure that are compared against the same baseline group (no exposure), we will have $n_{1+0} = n_{10+}$ and $n_{00+} = n_{0+0}$, and thus (Appendix A):

$$\begin{aligned}\operatorname{cov}\left(\hat{\beta}_1, \hat{\beta}_2\right) &= \operatorname{cov}\left(\log n_{1+0}, \log n_{10+}\right) + \operatorname{cov}\left(\log n_{00+}, \log n_{0+0}\right)\\ &= \operatorname{var}\left(\log n_{10+}\right) + \operatorname{var}\left(\log n_{00+}\right)\\ &= \frac{1}{m_{10+}} + \frac{1}{m_{00+}}\end{aligned}$$

which after plugging in the sample counts instead of the expectations, yields the final estimate

$$\widehat{\operatorname{cov}}\left(\hat{\beta}_1, \hat{\beta}_2\right) = \frac{1}{n_{10+}} + \frac{1}{n_{00+}} \tag{10}$$

Equation (10) has been used in the past in several situations such as in meta-analysis of dose–response epidemiological models, where we compare the various levels of exposure against the nonexposed category [8–10] and in multivariate meta-analysis of genetic association studies [11], where we compare the persons carrying the susceptibility allele *B* (i.e., individuals carrying the *AB* or *BB* genotype) against the persons homozygous for the common allele A (i.e., individuals carrying the *AA* genotype). A straightforward extension to include multiple categories has also been proposed recently, in the context of meta-analysis of haplotype association studies [51].

Another well-known easily derived example is encountered in observational studies that share the same group of controls [52] and in clinical trials with multiple treatments that share a common placebo group [6]. The latter has become very important recently, because it finds applications in the so-called multiple treatment comparison or network meta-analysis [53, 54]. In such a case we will have $n_{01+} = n_{0+1}$ and $n_{00+} = n_{0+0}$ and thus (Appendix A)

$$\begin{aligned}\operatorname{cov}\left(\hat{\beta}_1, \hat{\beta}_2\right) &= \operatorname{cov}\left(\log n_{01+}, \log n_{0+1}\right) + \operatorname{cov}\left(\log n_{00+}, \log n_{0+0}\right)\\ &= \operatorname{var}\left(\log n_{01+}\right) + \operatorname{var}\left(\log n_{00+}\right) = \frac{1}{m_{01+}} + \frac{1}{m_{00+}}\end{aligned}$$

which after plugging in the sample counts instead of the expectations, yields the estimate

$$\widehat{\operatorname{cov}}\left(\hat{\beta}_1, \hat{\beta}_2\right) = \frac{1}{n_{01+}} + \frac{1}{n_{00+}} \tag{11}$$

A slightly more general case could arise in case-control studies, in which the controls are only partially overlapping. The particular situation has been encountered recently within the context of genome-wide

association studies (GWAS), where cases of various common diseases were compared against a common set of controls [12, 13]. Expressions that are extended versions of Equation (11) have been proposed in these studies under different settings (i.e., within the logistic regression framework or concerning the allelic difference), using a different notation and method of derivation.

Moreover, Equation (4) can be also used to derive the covariance in the case of clinical trials with mutually exclusive outcomes [7]. In this particular situation (i.e., when we have death from cancer, death from other causes, and no death at all), the odds-ratios are calculated against all other alternatives and not only against the 'alive' category. Thus, using the notation of Table II, we will have Y denoting the treatment and X_1 and X_2 denoting the mutually exclusive outcomes, and it is easily understood that the two log-odds ratios will be negatively correlated. To reconstruct this scenario using the notation followed here, we have to resort to Table I and observe that $n_{111} = n_{011} = 0$ by design (a person cannot die from both causes) and that the remaining counts are disjoint. Then, using Equation (4) it can be shown that

$$\begin{aligned}\widehat{\mathrm{cov}}\left(\hat{\beta}_1, \hat{\beta}_2\right) &= -\frac{(n_{001}+n_{010}+n_{000})}{(n_{001}+n_{000})(n_{010}+n_{000})} - \frac{(n_{101}+n_{110}+n_{100})}{(n_{101}+n_{100})(n_{110}+n_{100})} \\ &= -\frac{(n_{00+}+n_{010})}{n_{00+}n_{0+0}} - \frac{(n_{10+}+n_{110})}{n_{10+}n_{1+0}}\end{aligned} \tag{12}$$

This estimate is identical to the one obtained recently by Trikalinos and Olkin [7], even though they used a different notation and method of derivation (i.e. they assumed independent binomial sampling and subsequently they used the normal approximation to the binomial).

Finally, in a recent work Bagos and Liakopoulos [14] used Equation (4) to derive the covariance to model the joint effects of two SNPs that are in linkage disequilibrium (LD) [55]. They used only the marginal genotype counts and an external estimate for LD derived from Hapmap. The additional complication of this method was the fact that X's were actually the genotypes (AA, AB, and BB for both polymorphisms) and given that the LD holds for alleles (A, B), some strong assumptions concerning Hardy–Weinberg equilibrium, and the equality of LD patterns between cases and controls (i.e., no three-way interaction) had to be made to reconstruct the combined genotype counts [14].

5. Application of the method

In this section, we apply the methods developed in this work in the three motivating examples presented in the Introduction. The first dataset comes from the Multiple Environmental and Genetic Assessment of Risk Factors for Venous Thrombosis Study [25] where the primary interest was the effect of smoking. The dataset was analysed by le Cessie *et al*. [20] using the sandwich variance estimator and an estimator derived from bootstrap analysis. From Table III of the original publication it is clear that 2544 patients were current or former smokers whereas 1391 patients were never smokers. Thus, applying Equation (10) results in a covariance equal to $1/1391 + 1/2544 = 0.00111198$. This corresponds to a correlation 0.309 if we use the estimated variances from the model without putative confounders, and a correlation equal to 0.282 if we use the estimated variances from models adjusting for age, sex, BMI (Body Mass Index), and pregnancy. In the first case, the method of le Cessie *et al.* [20] produced correlations equal to 0.31 and 0.28 by the sandwich variance estimator and bootstrap estimator, respectively. In the second case, the respective correlations were 0.30 and 0.28. Thus, it is clear that the estimator we proposed produces nearly identical correlations with the analyses using the individual data; needless to say that in both cases the final pooled estimates are identical to the ones obtained by le Cessie *et al.* up to the second decimal place.

In the study concerning the association of CTLA4 polymorphisms with T1DM [27], we present the results of the haplotype analysis considering GCGG as the susceptibility haplotype and comparing it against the remaining ones. From the available published report data, we conclude that the covariance is equal to $1/153 + 1/197 = 0.011612$. In Table IV we present the results obtained from the case-control analysis, the TDT analysis and the pooled estimate. The two estimates (TDT and case-control) are close, but because of smaller sample size the TDT analysis failed to show evidence for association. It is clear that by combining the results we get a pooled estimate pointing towards a significant association of the haplotype with T1DM. The correlation is rather large (0.665) and the pooled estimate has smaller variance than both individual estimates producing thus narrower confidence intervals. The test of equality of the two logORs does not provide evidence against the null hypothesis ($z = -0.6708$), justifying this way the choice of pooling the estimates in a combined analysis.

Table IV. Results for the haplotype analysis concerning the association of GCGG haplotype of CTLA4 with T1DM [27]. The OR is derived by contrasting GCGG haplotype against the remaining ones.

	Odds ratio	95% CI		log(OR)	s.e.(logOR)	ρ (correlation)
TDT	1.247	0.955	1.629	0.2211	0.1363	—
CC	1.341	1.044	1.724	0.2938	0.1280	—
Pooled	1.302	1.029	1.648	0.2642	0.1202	0.6654

Finally, we reanalysed the results of a GWAS conducted for the identification of susceptibility loci for multiple sclerosis [30]. In the original analysis the authors chose not to combine the two estimates but instead to retain for the replication phase SNPs that showed significant association with either of the tests (TDT or case-control). These 174 SNPs were subsequently genotyped in an additional set of 609 family trios, 2322 case subjects, and 2987 controls. Here, we analysed only the results of the screening phase to provide a combined estimate. Because the original data were not available, we used the estimates for the 16 SNPs reported in the respective publication (i.e., see Table II in [30]) and we reconstructed the summary counts as advised previously [56]. The results of this analysis are presented in Table V, where we list the odds-ratios along with their 95% confidence interval (CI) and the respective p-values for the 16 SNPs. It is clear that the combined analysis produces p-values with nominal statistical significance ($p < 0.05$) for all the analysed SNPs, whereas in the separate analyses two SNPs (rs1321172, rs10984447) produced nonsignificant estimates in one of the tests. Note that there is a rather large correlation (ranging from 0.464 to 0.682) between the two estimates for all the SNPs analysed, and thus a pooling procedure that ignores it would have produced biased results.

6. Discussion

We presented here a simple method for calculating the covariance of two correlated logORs. The covariance may be needed in several situations such as when pooling the estimates to obtain a combined estimate, when we are performing a joint modeling in a form of bivariate meta-analysis or when we simply want to perform a comparison of the two estimates and construct a confidence interval for the difference. The general method presented here is very simple and straightforward, it can be applied using only data that are available in the published reports and more importantly, it is very general, because it is shown to encompass several previously proposed estimates as special cases. The particular work apart from theoretical and pedagogical has also practical importance. By encompassing the previous approaches in a unifying framework that is derived from a simple procedure, the method allows easily deriving estimates for the covariance concerning problems that are not easy to obtain otherwise.

We have shown that a straightforward application of the method could be used to derive the covariance of log-odds ratios arising from matched and unmatched case control studies that use the same group of cases, a situation that has been addressed in the past only using individual data. Application of the method in three real-life applications revealed that the method performs quite satisfactorily, because although it uses only published data, it produces nearly identical estimates compared with methods that require individual data. Applying the same procedure to the analogous situation of combining the results of TDT and case-control genetic association studies that share the same cases also yields very encouraging results. Family-based and population-based studies have been shown to be comparable in empirical evaluations [57] and thus, integrating the results in meta-analyses using summary data derived from the literature would increase the power of detecting weak genotype effects. The method has also shown to be very useful in the case of GWAS. In particular, the derivation of closed form expressions for the covariance makes the method very attractive in terms of computational time. This is reinforced by recent evidence suggesting that using time-consuming individual data methods for meta-analysis of GWAS does not necessarily result in increased efficiency [58].

The statistical theory behind the proposed methodology is very simple and is based on standard large sample approximations [38, 59] that are in every-day use by researchers performing meta-analyses of published data. Theoretically, the method is expected to fail when the sample size is very small or when there are several studies, with small or even zero, cell counts. However, in the majority of the situations that we presented, the particular counts needed to compute the covariance should be, one way or another, sufficiently large for the analysis to be carried out using the odds ratio in the first place. In several situations however, such as in systematic reviews assessing drug safety or in epidemiological studies of

Table V. The results obtained from reanalysis of the data of a GWAS conducted for identification of susceptibility loci for multiple sclerosis [30]. We list the odds-ratios derived from the TDT and the case-control analysis along with their 95% CI and the respective p-values for the 16 SNPs. We also present the correlation between the two estimates (ρ) and the estimates derived from the combined analysis. The combined analysis produces p-values with nominal statistical significance ($p < 0.05$) for all the analysed SNPs, whereas in the original analysis two SNPs (rs1321172, rs10984447 denoted by an asterisk) produced nonsignificant estimates in one of the samples.

		931 Family trios (TDT)		Cases from 931 family trios vs 2431 unrelated controls (case-control)		Combined analysis of 931 cases, their parents and 2431 unrelated controls		
Gene	SNP	OR (95% C.I.)	P_{TDT}	OR (95% CI)	P_{CC}	OR (95% CI)	Correlation (ρ)	P_{comb}
IL2RA	rs12722489	1.35 (1.13, 1.62)	1.28×10^{-3}	1.3 (1.11, 1.52)	9.61×10^{-4}	1.31 (1.13, 1.53)	0.663	3.85×10^{-4}
IL2RA	rs2104286	1.26 (1.08, 1.47)	3.25×10^{-3}	1.26 (1.11, 1.43)	2.85×10^{-4}	1.26 (1.11, 1.42)	0.613	2.16×10^{-4}
IL7R	rs6897932	1.24 (1.07, 1.44)	5.78×10^{-3}	1.17 (1.03, 1.32)	1.65×10^{-2}	1.19 (1.05, 1.34)	0.617	4.51×10^{-3}
KIAA0350	rs6498169	1.16 (1.02, 1.33)	2.90×10^{-2}	1.17 (1.04, 1.31)	6.51×10^{-3}	1.17 (1.05, 1.30)	0.618	5.96×10^{-3}
RPL5	rs6604026	1.29 (1.11, 1.5)	4.45×10^{-4}	1.25 (1.11, 1.4)	2.34×10^{-4}	1.26 (1.13, 1.41)	0.543	5.58×10^{-5}
*DBC1**	rs10984447*	1.36 (1.16, 1.59)	1.18×10^{-4}	1.09 (0.95, 1.24)	2.13×10^{-1}	1.16 (1.02, 1.32)	0.619	2.20×10^{-2}
CD58	rs12044852	1.48 (1.17, 1.87)	9.71×10^{-4}	1.54 (1.26, 1.89)	3.01×10^{-5}	1.52 (1.25, 1.85)	0.682	2.79×10^{-5}
ALK	rs7577363	2.14 (1.43, 3.2)	1.14×10^{-4}	1.44 (1.08, 1.92)	1.21×10^{-2}	1.57 (1.19, 2.06)	0.464	1.52×10^{-3}
FAM69A	rs7536563	1.34 (1.16, 1.55)	2.53×10^{-5}	1.14 (1.02, 1.27)	2.48×10^{-2}	1.18 (1.06, 1.31)	0.542	2.21×10^{-3}
FAM69A	rs11164838	1.32 (1.15, 1.52)	6.01×10^{-5}	1.18 (1.06, 1.32)	3.28×10^{-3}	1.21 (1.09, 1.35)	0.574	4.00×10^{-4}
ANKRD15	rs10975200	1.26 (1.06, 1.5)	8.05×10^{-3}	1.37 (1.19, 1.58)	9.95×10^{-6}	1.34 (1.17, 1.53)	0.554	2.44×10^{-5}
EVI5	rs10735781	1.29 (1.12, 1.5)	2.21×10^{-4}	1.17 (1.05, 1.3)	6.05×10^{-3}	1.19 (1.07, 1.32)	0.529	9.00×10^{-4}
EVI5	rs6680578	1.29 (1.11, 1.49)	3.46×10^{-4}	1.17 (1.05, 1.31)	4.88×10^{-3}	1.19 (1.07, 1.33)	0.537	1.15×10^{-3}
KLRB1	rs4763655	1.15 (1.00, 1.32)	4.55×10^{-2}	1.19 (1.07, 1.33)	2.16×10^{-3}	1.18 (1.06, 1.31)	0.561	2.00×10^{-3}
CBLB	rs12487066	1.22 (1.05, 1.41)	7.65×10^{-3}	1.29 (1.14, 1.46)	4.09×10^{-5}	1.27 (1.13, 1.43)	0.627	8.49×10^{-5}
*PDE4B**	rs1321172*	1.12 (0.98, 1.27)	8.77×10^{-2}	1.15 (1.04, 1.28)	9.57×10^{-3}	1.14 (1.03, 1.26)	0.614	9.59×10^{-3}

rare diseases, zero cell counts are expected to occur and the investigator should be aware that the method will fail completely [60–62]. In such situations, relying on the usual continuity correction that consists of adding 1/2 to the cell counts may introduce bias [62]. Finally, when it comes to prospective studies, for which the use of odds-ratio is not mandatory, the development of methods based on the arcsine transform [61] could be pursued in future studies, pretty much in the spirit of Trikalinos and Olkin [7].

Even though some of the other methods, especially those in the context of genetic epidemiology, are more sophisticated (i.e., allow for more complex study designs or genetic models of inheritance), we should emphasize once again that the proposed method is the only currently available alternative for several of the examples presented, which allows pooling the data using only data derived from the published reports and thus we expect to be useful in case the individual data are not available (as usually is the case).

Although we have focused on the odds ratio, the method is easily applied to other effect sizes as well (i.e., risk ratio and so on). Moreover, it can also be applied to other ratio-type measures for achieving various goals. For instance, a recently proposed [63] Bayesian method for meta-analysis of genetic association studies that adjusts for deviations from Hardy–Weinberg equilibrium can be easily performed in frequentist framework by using a ratio measure for the departures such as the α coefficient [64, 65]. This measure is calculated from the genotype frequencies and as expected, is stochastically dependent with the odds ratio derived from the same genotypes. The covariance however, is easily calculated by the method proposed here allowing thus, to fit the model very easily in a frequentist framework using standard software.

Moreover, we have to mention that the method proposed here can also be used for extending the recently proposed method of 'synthesis analysis' [66]. Whereas traditional meta-analysis search for a single summary measure for the relation of two variables (Y, X, i.e., exposure and disease) across k studies, synthesis analysis seeks to integrate estimates for two or more predictors (X_1, X_2), in a multivariate model for predicting Y, using only information from the pairwise comparisons, for instance using estimates from the relations (Y, X_1), (Y, X_2), and (X_1, X_2) [66, 67]. However, the methodology was originally proposed only for continuous variables and an extension to the binary case requires the calculation of the covariances between correlated logORs that can be easily performed with the method proposed here. Such applications should be pursued in the near future where we expect the method to be widely used.

Finally, in the higher level of generality, one may encounter situations where the joint distribution needed to form the *partial contingency table* (Table I) may not be available. Even in such a case, there is a lot to be done if we have some strong prior beliefs concerning the association of Y, X_1, and X_2. For instance, assuming that X_1 and X_2 are marginally dependent (i.e., ignoring Y), but their correlation does not vary with the levels of Y (i.e., there is no three-factor interaction in the contingency table), we can use log-linear modelling [38] and obtain maximum likelihood estimates of the cell frequencies of Table II by the iterative proportional fitting algorithm [68, 69]. Thus, the method can be used in the future in a number of applications that are not easily addressed otherwise.

Appendix A

The sample estimates of the variances that are given by the well-know formulae of Equation (2), also known as the Woolf's estimator, were originally derived using the large sample normal approximation to the binomial distribution [40]. The same variance however, is easily computed by treating the sample counts (n_{ijk}) as independent Poisson variables with $E[n_{ijk}] = \mathrm{var}[n_{ijk}] = m_{ijk}$ and the log-odds ratios as contrasts among the log counts [8]. Then, using the delta-method [38] we can compute the variance of log-counts

$$\mathrm{var}\left[\log\left(n_{ijk}\right)\right] \approx \mathrm{var}\left[n_{ijk}\right]\left(\frac{\partial \log\left(E\left[n_{ijk}\right]\right)}{\partial\left(E\left[n_{ijk}\right]\right)}\right)^2 = m_{ijk}\left(\frac{1}{m_{ijk}}\frac{\partial N_{ijk}}{\partial N_{ijk}}\right)^2 = \frac{1}{m_{ijk}} \tag{A1}$$

Afterwards, by replacing the expectations (m_{ijk}) by the observed values (n_{ijk}), we can derive the estimate for the variance of log-counts

$$\widehat{\mathrm{var}}\left[\log\left(n_{ijk}\right)\right] = \frac{1}{n_{ijk}} \tag{A2}$$

Subsequently, by making use of standard properties of the covariance function, such as

$$\begin{aligned}\operatorname{cov}(aX+bY, cW+dV) = ac\,\operatorname{cov}(X,W) + ad\,\operatorname{cov}(X,V)\\ + bc\,\operatorname{cov}(Y,W) + bd\,\operatorname{cov}(Y,V)\end{aligned}$$

we can derive the estimate for the variance of $\hat{\beta}_1$:

$$\widehat{\operatorname{var}}\left(\hat{\beta}_1\right) = \frac{1}{n_{11+}} + \frac{1}{n_{00+}} + \frac{1}{n_{10+}} + \frac{1}{n_{01+}} \tag{A3}$$

Similar calculations yield the estimated variance of $\hat{\beta}_2$. This variance, even though based on a large sample approximation, performs quite well even for small sample sizes [41] justifying this way its popularity. Accordingly, the covariance of $\hat{\beta}_1$ and $\hat{\beta}_2$ will be given by

$$\begin{aligned}\operatorname{cov}\left(\hat{\beta}_1, \hat{\beta}_2\right) &= \operatorname{cov}\left[\log\left(\frac{n_{11+}n_{00+}}{n_{01+}n_{10+}}\right), \log\left(\frac{n_{1+1}n_{0+0}}{n_{0+1}n_{1+0}}\right)\right]\\ &= \operatorname{cov}\left[\log\left(n_{11+}n_{00+}\right) - \log\left(n_{01+}n_{10+}\right), \log\left(n_{1+1}n_{0+0}\right) - \log\left(n_{0+1}n_{1+0}\right)\right]\end{aligned}$$

which after calculations using the properties of the covariance function reduces to

$$\begin{aligned}\operatorname{cov}\left(\hat{\beta}_1, \hat{\beta}_2\right) = &\operatorname{cov}\left(\log n_{11+}, \log n_{1+1}\right) + \operatorname{cov}\left(\log n_{11+}, \log n_{0+0}\right)\\ &+ \operatorname{cov}\left(\log n_{00+}, \log n_{1+1}\right) + \operatorname{cov}\left(\log n_{00+}, \log n_{0+0}\right)\\ &- \operatorname{cov}\left(\log n_{11+}, \log n_{0+1}\right) - \operatorname{cov}\left(\log n_{11+}, \log n_{1+0}\right)\\ &- \operatorname{cov}\left(\log n_{00+}, \log n_{0+1}\right) - \operatorname{cov}\left(\log n_{00+}, \log n_{1+0}\right)\\ &- \operatorname{cov}\left(\log n_{01+}, \log n_{1+1}\right) - \operatorname{cov}\left(\log n_{01+}, \log n_{0+0}\right)\\ &- \operatorname{cov}\left(\log n_{10+}, \log n_{1+1}\right) - \operatorname{cov}\left(\log n_{10+}, \log n_{0+0}\right)\\ &+ \operatorname{cov}\left(\log n_{01+}, \log n_{0+1}\right) + \operatorname{cov}\left(\log n_{01+}, \log n_{1+0}\right)\\ &+ \operatorname{cov}\left(\log n_{10+}, \log n_{0+1}\right) + \operatorname{cov}\left(\log n_{10+}, \log n_{1+0}\right)\end{aligned} \tag{A4}$$

By observing that $\operatorname{cov}\left(\log n_{i'jk}, \log n_{ijk}\right) = 0, \forall i \neq i'$, because cases and controls (or treated and nontreated individuals) are always nonoverlapping, we can derive

$$\begin{aligned}\operatorname{cov}\left(\hat{\beta}_1, \hat{\beta}_2\right) = &\operatorname{cov}\left(\log n_{11+}, \log n_{1+1}\right) + \operatorname{cov}\left(\log n_{00+}, \log n_{0+0}\right)\\ &- \operatorname{cov}\left(\log n_{11+}, \log n_{1+0}\right) - \operatorname{cov}\left(\log n_{00+}, \log n_{0+1}\right)\\ &- \operatorname{cov}\left(\log n_{01+}, \log n_{0+0}\right) - \operatorname{cov}\left(\log n_{10+}, \log n_{1+1}\right)\\ &+ \operatorname{cov}\left(\log n_{01+}, \log n_{0+1}\right) + \operatorname{cov}\left(\log n_{10+}, \log n_{1+0}\right)\\ = &\sum_i \sum_j \sum_{k=j} \operatorname{cov}\left(\log n_{ij+}, \log n_{i+k}\right) - \sum_i \sum_j \sum_{k\neq j} \operatorname{cov}\left(\log n_{ij+}, \log n_{i+k}\right)\\ = &\sum_i \sum_j \sum_k (-1)^{j-k} \operatorname{cov}\left(\log n_{ij+}, \log n_{i+k}\right)\end{aligned} \tag{A5}$$

In the general case (i.e., see Table I) we assume that $n_{ij+} = n_{ij1} + n_{ij0}$ and $n_{i+k} = n_{i0k} + n_{i1k}$ and thus, we can rewrite Equation (A5) in the form

$$\begin{aligned}&\operatorname{cov}\left(\hat{\beta}_1, \hat{\beta}_2\right)\\ &= \operatorname{cov}\left[\log\left(n_{111}+n_{110}\right), \log\left(n_{111}+n_{101}\right)\right] + \operatorname{cov}\left[\log\left(n_{001}+n_{000}\right), \log\left(n_{010}+n_{000}\right)\right]\\ &\quad - \operatorname{cov}\left[\log\left(n_{111}+n_{110}\right), \log\left(n_{110}+n_{100}\right)\right] - \operatorname{cov}\left[\log\left(n_{001}+n_{000}\right), \log\left(n_{011}+n_{001}\right)\right]\\ &\quad - \operatorname{cov}\left[\log\left(n_{011}+n_{010}\right), \log\left(n_{010}+n_{000}\right)\right] - \operatorname{cov}\left[\log\left(n_{101}+n_{100}\right), \log\left(n_{111}+n_{101}\right)\right]\\ &\quad + \operatorname{cov}\left[\log\left(n_{011}+n_{010}\right), \log\left(n_{011}+n_{001}\right)\right] + \operatorname{cov}\left[\log\left(n_{101}+n_{100}\right), \log\left(n_{110}+n_{100}\right)\right]\\ &= \sum_i \sum_j \sum_{k=j} \operatorname{cov}\left(\log \sum\nolimits_c n_{ijc}, \log \sum\nolimits_c n_{ick}\right) - \sum_i \sum_j \sum_{k\neq j} \operatorname{cov}\left(\log \sum\nolimits_c n_{ijc}, \log \sum\nolimits_c n_{ick}\right)\\ &= \sum_i \sum_j \sum_k (-1)^{j-k} \operatorname{cov}\left(\log \sum\nolimits_c n_{ijc}, \log \sum\nolimits_c n_{ick}\right)\end{aligned} \tag{A6}$$

We now observe that the right-hand side of Equation (A6) includes terms of the form cov[log($S+U$), log($S+W$)], which need to be computed. If we assume that S, U, W are independent Poisson random variables with $E[S]=\text{var}[S]=s$, $E[U]=\text{var}[U]=u$ and $E[W]=\text{var}[W]=w$, then the linear combinations $Z=S+U$ and $V=S+W$ are also Poisson random variables with $E[Z]=\text{var}[Z]=z=s+u$ and $E[V]=\text{var}[V]=v=s+w$, respectively. Thus, (V, Z) will follow a bivariate Poisson distribution [42] and the covariance between Z and V would then be

$$\text{cov}[S+U, S+W]=\text{cov}[S,S]=\text{var}[S]=s \tag{A7}$$

The covariance between functions of random variables Z and V, requires analytical integration involving the joint and the marginal cumulative distribution functions [70]. However, we can easily approximate it using the delta method [38] with a first-order Taylor expansion around the means

$$\begin{aligned}\text{cov}\left[f\left(Z\right),g\left(V\right)\right] &\approx \text{cov}\left[\frac{\partial f\left(E\left[Z\right]\right)}{\partial Z}\left(Z-E\left[Z\right]\right),\frac{\partial g\left(E\left[V\right]\right)}{\partial V}\left(V-E\left[V\right]\right)\right]\\ &=\frac{\partial f\left(E\left[Z\right]\right)}{\partial Z}\frac{\partial g\left(E\left[V\right]\right)}{\partial V}\text{cov}\left(Z,V\right)\end{aligned}$$

Thus, we will have

$$\begin{aligned}\text{cov}\left[\log\left(S+U\right),\log\left(S+W\right)\right] &\approx \text{cov}\left[\log\left(Z\right),\log\left(V\right)\right]\\ &=\text{cov}\left(Z,V\right)\frac{\partial \log Z}{\partial Z}\frac{\partial \log V}{\partial V}\\ &=\frac{s}{zv}=\frac{s}{\left(s+u\right)\left(s+w\right)}\end{aligned} \tag{A8}$$

Thus, we get the intuitive result that the covariance is equal to the common component of the counts divided by the product of the marginal counts. Finally, by applying recursively Equation (A8), we can further rewrite Equation (A6) and, after plugging in the observed counts, we can derive the final formula for the approximate estimate of the covariance

$$\begin{aligned}\widehat{\text{cov}}\left(\hat{\beta}_1,\hat{\beta}_2\right) &= \frac{n_{111}}{n_{11+}n_{1+1}}+\frac{n_{000}}{n_{00+}n_{0+0}}-\frac{n_{110}}{n_{11+}n_{1+0}}-\frac{n_{001}}{n_{00+}n_{0+1}}\\ &\quad -\frac{n_{010}}{n_{01+}n_{0+0}}-\frac{n_{101}}{n_{10+}n_{1+1}}+\frac{n_{011}}{n_{01+}n_{0+1}}+\frac{n_{100}}{n_{10+}n_{1+0}}\\ &=\sum_i\sum_j\sum_{k=j}\left(\frac{n_{ijk}}{n_{ij+}n_{i+k}}\right)-\sum_i\sum_j\sum_{k\neq j}\left(\frac{n_{ijk}}{n_{ij+}n_{i+k}}\right)\\ &=\sum_i\sum_j\sum_k(-1)^{j-k}\left(\frac{n_{ijk}}{n_{ij+}n_{i+k}}\right)\end{aligned} \tag{A9}$$

Acknowledgements

The author would like to thank the associate editor and two anonymous reviewers whose comments and constructive criticism helped in improving the quality of the manuscript.

References

1. Altman DG, Bland JM. Interaction revisited: the difference between two estimates. *British Medical Journal* 2003; **326**(7382):219.
2. Normand SL. Meta-analysis: formulating, evaluating, combining, and reporting. *Statistics in Medicine* 1999; **18**(3): 321–359.
3. van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine* 2002; **21**(4):589–624.
4. Gleser LJ, Olkin I. Stochastically dependent effect sizes. In *The Handbook of Research Synthesis*, Cooper HM, Hedges LV (eds). Russell Sage Foundation: New York, 1994; 339–355.
5. Berkey CS, Hoaglin DC, Antczak-Bouckoms A, Mosteller F, Colditz GA. Meta-analysis of multiple outcomes by regression with random effects. *Statistics in Medicine* 1998; **17**(22):2537–2550.

6. Higgins JP, Whitehead A. Borrowing strength from external trials in a meta-analysis. *Statistics in Medicine* 1996; **15**(24):2733–2749.
7. Trikalinos TA, Olkin I. A method for the meta-analysis of mutually exclusive binary outcomes. *Statistics in Medicine* 2008; **27**(21):4279–4300.
8. Berrington A, Cox DR. Generalized least squares for the synthesis of correlated information. *Biostatistics* 2003; **4**(3):423–431.
9. Greenland S, Longnecker MP. Methods for trend estimation from summarized dose-response data, with applications to meta-analysis. *American Journal of Epidemiology* 1992; **135**(11):1301–1309.
10. Berlin JA, Longnecker MP, Greenland S. Meta-analysis of epidemiologic dose-response data. *Epidemiology* 1993; **4**(3):218–228.
11. Bagos PG. A unification of multivariate methods for meta-analysis of genetic association studies. *Statistical Applications in Genetics and Molecular Biology* 2008; **7**. Article31.
12. Lin DY, Sullivan PF. Meta-analysis of genome-wide association studies with overlapping subjects. *American Journal of Human Genetics* 2009; **85**(6):862–872.
13. Zaykin DV, Kozbur DO. P-value based analysis for shared controls design in genome-wide association studies. *Genetic Epidemiology* 2010; **34**(7):725–738.
14. Bagos PG, Liakopoulos TD. A multipoint method for meta-analysis of genetic association studies. *Genetic Epidemiology* 2010; **34**(7):702–715.
15. Daniels MJ, Hughes MD. Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine* 1997; **16**(17):1965–1982.
16. Thompson JR, Minelli C, Abrams KR, Tobin MD, Riley RD. Meta-analysis of genetic studies using Mendelian randomization–a multivariate approach. *Statistics in Medicine* 2005; **24**(14):2241–2254.
17. Riley RD, Abrams KR, Lambert PC, Sutton AJ, Thompson JR. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. *Statistics in Medicine* 2007; **26**(1):78–97.
18. Riley RD, Abrams KR, Sutton AJ, Lambert PC, Thompson JR. Bivariate random-effects meta-analysis and the estimation of between-study correlation. *BMC Medical Research Methodology* 2007; **7**:3.
19. Riley RD, Thompson JR, Abrams KR. An alternative model for bivariate random-effects meta-analysis when the within-study correlations are unknown. *Biostatistics* 2008; **9**(1):172–186.
20. le Cessie S, Nagelkerke N, Rosendaal FR, van Stralen KJ, Pomp ER, van Houwelingen HC. Combining matched and unmatched control groups in case-control studies. *American Journal of Epidemiology* 2008; **168**(10):1204–1210.
21. Pomp ER, Van Stralen KJ, Le Cessie S, Vandenbroucke JP, Rosendaal FR, Doggen CJ. Experience with multiple control groups in a large population-based case-control study on genetic and environmental risk factors. *European Journal of Epidemiology* 2010; **25**(7):459–466.
22. Duffy SW, Rohan TE, Altman DG. A method for combining matched and unmatched binary data. Application to randomized, controlled trials of photocoagulation in the treatment of diabetic retinopathy. *American Journal of Epidemiology* 1989; **130**(2):371–378.
23. Moreno V, Martin ML, Bosch FX, de Sanjose S, Torres F, Munoz N. Combined analysis of matched and unmatched case-control studies: comparison of risk estimates from different studies. *American Journal of Epidemiology* 1996; **143**(3):293–300.
24. Huberman M, Langholz B. Re: Combined analysis of matched and unmatched case-control studies: comparison of risk estimates from different studies. *American Journal of Epidemiology* 1999; **150**(2):219–220.
25. Pomp ER, Rosendaal FR, Doggen CJ. Smoking increases the risk of venous thrombosis and acts synergistically with oral contraceptive use. *American Journal of Epidemiology* 2008; **83**(2):97–102.
26. Lewis CM. Genetic association studies: design, analysis and interpretation. *Briefings in Bioinformatics* 2002; **3**(2):146–153.
27. Zhernakova A, Eerligh P, Barrera P, Wesoly JZ, Huizinga TW, Roep BO, Wijmenga C, Koeleman BP. CTLA4 is differentially associated with autoimmune diseases in the Dutch population. *Human Genetics* 2005; **118**(1):58–66.
28. Benyamin B, Visscher PM, McRae AF. Family-based genome-wide association studies. *Pharmacogenomics* 2009; **10**(2):181–190.
29. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nature Genetics* 2006; **38**(2):209–213.
30. Hafler DA, Compston A, Sawcer S, Lander ES, Daly MJ, De Jager PL, de Bakker PI, Gabriel SB, Mirel DB, Ivinson AJ, *et al.* Risk alleles for multiple sclerosis identified by a genomewide study. *New England Journal of Medicine* 2007; **357**(9):851–862.
31. Pfeiffer RM, Pee D, Landi MT. On combining family and case-control studies. *Genetic Epidemiology* 2008; **32**(7):638–646.
32. Gebregziabher M, Guimaraes P, Cozen W, Conti DV. A polytomous conditional likelihood approach for combining matched and unmatched case-control studies. *Statistics in Medicine* 2010; **29**(9):1004–1013.
33. Nagelkerke NJ, Hoebee B, Teunis P, Kimman TG. Combining the transmission disequilibrium test and case-control methodology using generalized logistic regression. *European Journal of Human Genetics* 2004; **12**(11):964–970.
34. Zou GY. Statistical Methods for the Analysis of Genetic association studies. *Annals of Human Genetics* 2005; **69**:1–15.
35. Hsu L, Starr JR, Zheng Y, Schwartz SM. On combining triads and unrelated subjects data in candidate gene studies: an application to data on testicular cancer. *Human Heredity* 2009; **67**(2):88–103.
36. Infante-Rivard C, Mirea L, Bull SB. Combining case-control and case-trio data from the same population in genetic association analyses: overview of approaches and illustration with a candidate gene study. *American Journal of Epidemiology* 2009; **170**(5):657–664.
37. Chen YH, Lin HW. Simple association analysis combining data from trios/sibships and unrelated controls. *Genetic Epidemiology* 2008; **32**(6):520–527.

38. Agresti A. *Categorical Data Analysis*, (2nd edn). John Wiley & Sons: New York, 2002.
39. Agresti A. A Survey of Exact Inference for Contingency Tables. *Statistical Science* 1992; **7**(1):131–153.
40. Woolf B. On estimating the relationship between blood group and disease. *Human Genetics* 1955; **19**:251–253.
41. Agresti A. On logit confidence intervals for the odds ratio with small samples. *Biometrics* 1999; **55**(2):597–602.
42. Marshall AW, Olkin I. A family of bivariate distributions generated by the bivariate bernoulli distribution. *Journal of the American Statistical Association* 1985; **80**(390):332–338.
43. Hamajima N, Hirose K, Inoue M, Takezaki T, Kuroishi T, Tajima K. Case-control studies: matched controls or all available controls? *Journal of Clinical Epidemiology* 1994; **47**(9):971–975.
44. Feinstein AR. Quantitative ambiguities in matched versus unmatched analyses of the 2x2 table for a case-control study. *International Journal of Epidemiology* 1987; **16**(1):128–134.
45. Thompson WD, Kelsey JL, Walter SD. Cost and efficiency in the choice of matched and unmatched case-control study designs. *American Journal of Epidemiology* 1982; **116**(5):840–851.
46. Hardy RJ, White C. Matching in retrospective studies. *American Journal of Epidemiology* 1971; **93**(2):75–76.
47. Miettinen OS. Matching and design efficiency in retrospective studies. *American Journal of Epidemiology* 1970; **91**(2):111–118.
48. Mitchell LE. Relationship between case-control studies and the transmission/disequilibrium test. *Genetic Epidemiology* 2000; **19**(3):193–201.
49. Kazeem GR. Farrall M. Integrating case-control and TDT studies. *Annals of Human Genetics* 2005; **69**(Pt 3):329–335.
50. Janssen R, Bont L, Siezen CL, Hodemaekers HM, Ermers MJ, Doornbos G, van't Slot R, Wijmenga C, Goeman JJ, Kimpen JL, *et al.* Genetic susceptibility to respiratory syncytial virus bronchiolitis is predominantly associated with innate immune genes. *Journal of Infectious Diseases* 2007; **196**(6):826–834.
51. Bagos PG. Meta-analysis of haplotype-association studies: comparison of methods and empirical evaluation of the literature. *BMC Genetics* 2011; **12**:8.
52. Bunin GR, Baumgarten M, Norman SA, Strom BL, Berlin JA. Practical aspects of sharing controls between case-control studies. *Pharmacoepidemiology and Drug Safety* 2005; **14**(8):523–530.
53. Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. *Statistics in Medicine* 2004; **23**(20):3105–3124.
54. Salanti G, Higgins JP, Ades AE, Ioannidis JP. Evaluation of networks of randomized trials. *Statistical Methods in Medical Research* 2008; **17**(3):279–301.
55. HapMap. The International HapMap Project. *Nature* 2003; **426**(6968):789–796.
56. Di Pietrantonj C. Four-fold table cell frequencies imputation in meta analysis. *Statistics in Medicine* 2006; **25**(13):2299–2322.
57. Evangelou E, Trikalinos TA, Salanti G, Ioannidis JP. Family-based versus unrelated case-control designs for genetic associations. *PLoS Genetics* 2006; **2**(8). e123.
58. Lin DY, Zeng D. Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. *Genetic Epidemiology* 2010; **34**(1):60–66.
59. Clayton D, Hills M. *Statistical Models in Epidemiology*. Oxford University Press: New York, 1993.
60. Cai T, Parast L, Ryan L. Meta-analysis for rare events. *Statistics in Medicine* 2010; **29**(20):2078–2089.
61. Rucker G, Schwarzer G, Carpenter J, Olkin I. Why add anything to nothing?The arcsine difference as a measure of treatment effect in meta-analysis with zero cells. *Statistics in Medicine* 2009; **28**(5):721–738.
62. Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing?Use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in Medicine* 2004; **23**(9):1351–1375.
63. Salanti G, Higgins JP, Trikalinos TA, Ioannidis JP. Bayesian meta-analysis and meta-regression for gene-disease associations and deviations from Hardy-Weinberg equilibrium. *Statistics in Medicine* 2007; **26**(3):553–567.
64. Lindley D. Statistical inference concerning Hardy-Weinberg equilibrium. *Bayesian Statistics* 1988; **3**:307–326.
65. Pereira C, Rogatko A. The Hardy-Weinberg equilibrium under a Bayesian perspective. *Revista Brasileira de Genetica* 1984; **4**:689–707.
66. Samsa G, Hu G, Root M. Combining information from multiple data sources to create multivariable risk models: illustration and preliminary assessment of a new method. *Journal of Biomedicine and Biotechnology* 2005; **2005**(2):113–123.
67. Zhou XH, Hu N, Hu G, Root M. Synthesis analysis of regression models with a continuous outcome. *Statistics in Medicine* 2009; **28**(11):1620–1635.
68. Deming WE, Stephan FF. On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known. *The Annals of Mathematical Statistics* 1940; **11**(4):427–444.
69. Fienberg SE. An Iterative Procedure for Estimation in Contingency Tables. *The Annals of Mathematical Statistics* 1970; **41**(3):907–917.
70. Cuadras CM. On the covariance between functions. *Journal of Multivariate Analysis* 2002; **81**:19–27.