Statistical Applications in Genetics and Molecular Biology

Volume 7, *Issue* 1 2008 *Article* 31

A Unification of Multivariate Methods for Meta-Analysis of Genetic Association Studies

Pantelis G. Bagos*

*University of Central Greece, pbagos@ucg.gr

A Unification of Multivariate Methods for Meta-Analysis of Genetic Association Studies*

Pantelis G. Bagos

Abstract

Methods for multivariate meta-analysis of genetic association studies are reviewed, summarized and presented in a unified framework. Modifications of standard models are described in detail in order to be applied in genetic association studies. The model based on summary data is uniformly defined for both discrete and continuous outcomes and analytical expressions for the covariance of the two jointly modeled outcomes are derived for both cases. The models based on the binary nature of the data are fitted using both prospective and retrospective likelihood. Furthermore, formal tests for assessing the genetic model of inheritance are developed based on standard normal theory. The general model is compared to the recently proposed genetic model-free bivariate approach (either using summary or binary data), and it is clearly shown that the estimates provided by this approach are nearly identical to the estimates derived by the general bivariate model using the aforementioned tests for the genetic model. The methods developed here as well as the tests, are easily implemented in all major statistical packages, escaping the need of self written software. The methods are applied in several already published meta-analyses of genetic association studies (with both discrete and continuous outcomes) and the results are compared against the widely used univariate approach as well as against the genetic model free approaches. Illustrative examples of code in Stata are given in the appendix. It is anticipated that the methods developed in this work will be widely applied in the meta-analysis of genetic association studies.

KEYWORDS: meta-analysis, genetic epidemiology, random effects, multivariate methods

---

*The author would like to thank the editor and the two anonymous reviewers for the valuable comments and the constructive criticism that helped in improving the quality of the manuscript.

1. Introduction

The continuously increasing number of published genetic epidemiology studies (Becker et al, 2004; Hirschhorn et al, 2002), warrants the need for collecting and synthesizing the available information for a particular gene-disease association providing a quantitative overall estimate, a procedure known as meta-analysis (Normand, 1999; Petiti, 1994; van Houwelingen et al, 2002). Besides the problems encountered in meta-analysis of observational studies, special care is needed when dealing with meta-analyses of genetic factors speculated to be involved in a particular disease etiology (Attia et al, 2003; Ioannidis, 2005a; Ioannidis, 2005b; Ioannidis & Trikalinos, 2005; Salanti et al, 2005). Methods for combining evidence from family-based studies have been proposed (Gu et al, 2001), as well as methods for combining in a meta analysis the results from population-based and family-based studies (Evangelou et al, 2006; Kazeem & Farrall, 2005).

For population-based genetic association studies, which are of main interest and often fall under the well known in Epidemiology case-control design, several approaches have been proposed. Some of them are based on traditional approaches for meta-analysis of epidemiological studies with minor modifications to handle the genotype effects (Thakkinstian et al, 2005), whereas other more sophisticated methods have been proposed, using multivariate methods of meta-analysis (Minelli et al, 2005a; Minelli et al, 2005b). The major disadvantage of the multivariate methods proposed so far, is that they require specialized user-written software that is not currently available. Thus, in the vast majority of the published meta-analyses of genetic association studies (Attia et al, 2003; Ioannidis & Trikalinos, 2005; Ioannidis et al, 2003), the data are routinely analyzed in a univariate fashion ignoring the within studies pairwise correlation of genotype contrasts.

In this work, the various methods that were proposed in the past for performing meta-analysis of genetic association studies are summarized and a general multivariate framework for meta-analysis of genetic association studies is presented. The methodology is based on the general approach of bivariate meta-analysis (van Houwelingen et al, 2002), adapted for handling the correlation between the outcomes encountered in genetic association studies. The methods are capable of handling both discrete and continuous outcomes and can be easily applied in any major statistical package capable of fitting multivariate generalized linear models. Methods for performing inferences concerning the genetic model of inheritance are also presented and thus, a connection of the general multivariate framework with the genetic-model free approach is presented. The paper is organized as follows: In section 2, the problem is described in order to establish notation. The widely used univariate methods are presented and some multivariate

methods proposed earlier are discussed. In section 3, the bivariate approach using aggregate (summary) data is presented for both discrete and continuous outcomes and connections to the model-free approach (Minelli et al, 2005b) are discussed. In section 4 models using directly the binary nature of the data (for discrete outcomes) are presented, using either the retrospective or the prospective likelihood. In section 5, a general approach for inferring the genetic model of inheritance from the models presented here is given and the equivalence to the genetic model-free approach is shown. In section 6, application of the proposed methods in several published meta-analyses of genetic association studies concerning both discrete and continuous outcomes are presented and discussed. In section 7, the overall conclusions of this work are summarized and some other more special-purpose methods of multivariate meta-analysis that were developed for genetic association studies are discussed. Finally, in the Appendix, simple Stata programs are presented for performing the analyses reported in this paper.

2. Meta-analysis of genetic association studies

Consider a locus having two alleles (A, B) where the second (B) is assumed to be the risk factor for a particular disease. The possible genotypes of a person could be AA, AB and BB. Table 1 presents the data used in a meta-analysis of k studies, in which retrospectively sampled cases and controls were classified according to their genotypes. The majority of published genetic association studies involve dichotomous outcomes that can be represented in such a table. Another, even though not so commonly encountered situation (Table 2), is when the outcome of interest is continuous, such as Systolic Blood Pressure (SBP), cholesterol levels and so on. In this case, the studies are usually classified as having a cross-sectional design, where a representative sample from the general (or from a high-risk) population is collected, the individuals are genotyped and the mean values of the continuous measure that is under investigation are compared across genotypes.

Table 1. A typical layout of the data used in a meta-analysis of case-control genetic association studies involving a single bi-allelic locus with a dichotomous outcome. The distribution of the various genotypes (AA, AB and BB) is listed for cases and controls, whereas the different studies (i=1, 2 …k) included in the meta-analysis are listed in different rows.

Study	**Cases**			**Controls**		
1	AA_{11}	AB_{11}	BB_{11}	AA_{01}	AB_{01}	BB_{01}
2	AA_{12}	AB_{12}	BB_{12}	AA_{02}	AB_{02}	BB_{02}
…	…	…	…	…	…	…
k	AA_{1k}	AB_{1k}	BB_{1k}	AA_{0k}	AB_{0k}	BB_{0k}

Currently, the majority of meta-analyses of genetic association studies involving one locus with two variants (A vs. B), are performed by collapsing the genotypes in two categories assuming implicitly a particular genetic model and then performing comparisons by means of the Odds Ratio (OR, for dichotomous outcomes) or using the average difference (also known as weighted mean difference-WMD, for continuous outcomes). For instance, considering the contrast of BB+AB vs. AA genotypes, we implicitly assume a dominant model of inheritance, BB vs. AB+AA corresponds to a recessive model and so on. Another approach, which in the majority of the situations is performed in addition to the previous approach, is to compare allele frequencies between cases and controls (in the case of dichotomous outcomes) implying by this way an underlying co-dominant model of inheritance. Finally, another approach is to perform multiple comparisons (i.e. BB vs. AB, AB vs. AA etc) with the risk of an inflated Type I error rate.

Table 2. A typical layout of the data used in a meta-analysis of genetic association studies involving a single bi-allelic locus with a continuous outcome. The distribution of the various genotypes (AA, AB and BB) is listed for all participants (N), accompanied by the mean and the standard deviation of the measured continuous outcome in each group. The different studies (i=1, 2 …k) included in the meta-analysis are listed in different rows.

	genotypes					
Study	***AA***		***AB***		***BB***	
	mean (sd)	***N***	***mean (sd)***	***N***	***mean (sd)***	***N***
1	y_{AA1} (s_{AA1})	AA_1	y_{AB1} (s_{AB1})	AB_1	y_{BBA1} (s_{BB1})	BB_1
2	y_{AA2} (s_{AA2})	AA_2	y_{AB2} (s_{AB2})	AB_2	y_{BBA2} (s_{BB2})	BB_2
…	…	…	…	…	…	…
k	y_{AAk} (s_{AAk})	AA_k	y_{ABk} (s_{ABk})	AB_k	y_{BBAk} (s_{BBk})	BB_k

No matter which of the above-mentioned approaches is used, traditional methods of meta-analysis of summary measures are applicable either relying on random-effects or on fixed-effects models (Normand, 1999; Petiti, 1994). In a traditional fixed-effects meta-analysis using summary measures, the assumption usually made is that the individual estimates y_i of the logOR (or the mean difference) of each study, are distributed normally with mean equal to the true effect θ and variance σ^2, which is the estimated variance of the logOR (or the mean difference) of the particular contrast of each study (Normand, 1999; Petiti, 1994). In the presence of heterogeneity, an alternative and perhaps preferable method, is the method of random-effects, which assumes that the true effects vary randomly between studies and a random component of the between studies variance (τ^2) is introduced. The most common approach for estimation of τ^2 is the non-iterative method of moments (MM) proposed by DerSimonian and Laird

(DerSimonian & Laird, 1986). Other methods summarized by Thompson and Sharp (Thompson & Sharp, 1999), include iterative techniques, such as maximum likelihood (ML), restricted maximum likelihood (REML) and empirical Bayes estimation (EB). When heterogeneity is absent, τ^2 is essentially equal to zero and the fixed and random-effects methods coincide. Similar models have also been described in a Bayesian framework (Smith et al, 1995; Sutton & Abrams, 2001; Warn et al, 2002). In all of the above-mentioned methods, one could incorporate study-level covariates as linear predictors that potentially reduce the observed heterogeneity, resulting in a random-effects meta-regression (Thompson & Higgins, 2002; Thompson & Sharp, 1999).

Thakkinstian and coworkers (Thakkinstian et al, 2005), proposed a general framework for performing meta-analysis of genetic association studies that includes predefined steps involving tests for Hardy-Weinberg equilibrium (HWE), tests for heterogeneity and finally deciphering the most plausible genetic model. However, the authors still relied on fixed and random-effects methods based on summary measures. Most importantly however, they treated the inherently multivariate genetic data as if they were univariate, performing simultaneous inferences. The data on Table 1, can naturally been modeled by treating the two logORs derived from the mutant allele (AB vs. AA and BB vs. AA) as a bivariate response whose estimates are correlated. Taking this correlation into account is necessary when one attempts to draw simultaneously inferences concerning the statistical significance of the two logORs, as well as, when trying to compare them in order to decipher the genetic model of inheritance.

Even though general models for multivariate meta-analysis are available for years (Berkey et al, 1998; van Houwelingen et al, 1993), it is noteworthy that no attempt has been performed for adapting them in genetic-association studies. More importantly, the majority of published meta-analyses of genetic association studies treat the multiple outcomes as independent ones or by performing multiple comparisons (Attia et al, 2003; Ioannidis & Trikalinos, 2005; Ioannidis et al, 2003). On the other hand, during the last years some very interesting multivariate models have been proposed. Minelli and coworkers (Minelli et al, 2005b), proposed a very interesting (genetic) model-free approach for meta-analysis of genetic association studies, which does not specify in advance the genetic model but instead infers it from the data. In particular, they introduced the joint modeling of the logarithm of OR_{BB}, which is the OR of BB genotype vs. AA, and λ which is the ratio of the $logOR_{BB}$ and $logOR_{AB}$ (i.e the OR of AB genotype vs. AA), an approach that recognizes the fact that the two ORs are correlated. Lately, Minelli and coworkers extended their method in a Bayesian framework (Minelli et al, 2005a). Salanti and coworkers introduced another Bayesian method that incorporates directly in the meta-analysis the deviations from HWE using fixation coefficients (Salanti et al, 2006). A somewhat different approach was followed in

the so-called "Mendelian Randomization" method, which has been proposed in order to account for the pairwise correlations between phenotype-genotype and genotype-disease (Minelli et al, 2004; Thompson et al, 2005). Under this approach, we could finally decipher the association between phenotype and disease taking into account multiple sources of evidence from the literature. Lately, Bagos and Nikolopoulos (Bagos & Nikolopoulos, 2007) proposed an intuitive approach based on random coefficient logistic regression models easily implemented in Stata. However, most of the above mentioned methods are not widely used in practice for a series of reasons. For instance, the genetic model-free and the Mendelian randomization approaches cannot be easily implemented since the authors performed the analyses using self-written programs in Stata that were not published along with the respective papers. Bayesian methods on the other hand, even though appealing, they are difficult to be widely used from people performing meta-analysis, since an investigator should have knowledge of Bayesian statistics and programming skills in WinBUGS (Spiegelhalter et al, 2004). Furthermore, a Bayesian analysis requires a significant amount of time in order to monitor the convergence of MCMC and perform the necessary diagnostics. However, a major advantage of these methods is the fact that WinBUGS is freely available software. In any case though, it would be advantageous to have available Maximum Likelihood methods capable of performing the same analyses in a frequentist framework.

In the following section, the general model of bivariate random-effects meta-analysis is going to be presented and adapted for genetic association studies. The methodology is going to be presented separately for discrete and continuous outcome data. For comparison, the genetic model-free approach is going to be presented and the differences of the two approaches will be highlighted.

3. Multivariate meta-analysis using aggregate data

3.1 The general multivariate model of meta-analysis for discrete outcomes

In the general approach for bivariate meta-analysis (van Houwelingen et al, 2002), the two logORs derived from the mutant allele (AB vs. AA and BB vs. AA) could be modeled simultaneously as a bivariate response. The logarithm of OR_{AB} (the OR of heterozygous versus homozygous for the wild type) is given by:

$$y_{1i} = \log\left(\frac{AB_{1i}AA_{0i}}{AA_{1i}AB_{0i}}\right) \tag{3.1}$$

with an approximate variance calculated by:

$$s_{1i}^2 = 1/AA_{1i} + 1/AA_{0i} + 1/AB_{1i} + 1/AB_{0i} \quad (3.2)$$

Under the same rationale, the logarithm of OR_{BB} (the OR of homozygous for the mutant allele versus the homozygous for the wild type) is given by:

$$y_{2i} = \log\left(\frac{BB_{1i}AA_{0i}}{AA_{1i}BB_{0i}}\right) \quad (3.3)$$

with variance:

$$s_{2i}^2 = 1/AA_{1i} + 1/AA_{0i} + 1/BB_{1i} + 1/BB_{0i} \quad (3.4)$$

In a random-effects setting, we assume that the two logORs are distributed following a bivariate normal distribution:

$$\begin{bmatrix} y_{1i} \\ y_{2i} \end{bmatrix} \sim MVN\left\{\begin{bmatrix} \beta_{1i} \\ \beta_{2i} \end{bmatrix}, \begin{bmatrix} s_{1i}^2 & \rho_{W12}s_{1i}s_{2i} \\ \rho_{W12}s_{1i}s_{2i} & s_{1i}^2 \end{bmatrix}\right\} \quad (3.5)$$

with the means (β_{1i}, β_{2i}) which are considered random terms, distributed similarly as:

$$\begin{bmatrix} \beta_{1i} \\ \beta_{2i} \end{bmatrix} \sim MVN\left\{\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}, \begin{bmatrix} \tau_1^2 & \rho_{B12}\tau_1\tau_2 \\ \rho_{B12}\tau_1\tau_2 & \tau_2^2 \end{bmatrix}\right\} \quad (3.6)$$

Thus, the final marginal model on which we base the inference is:

$$\begin{bmatrix} y_{1i} \\ y_{2i} \end{bmatrix} \sim MVN\left\{\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}, \begin{bmatrix} s_{1i}^2 + \tau_1^2 & \rho_{W12}s_{1i}s_{2i} + \rho_{B12}\tau_1\tau_2 \\ \rho_{W12}s_{1i}s_{2i} + \rho_{B12}\tau_1\tau_2 & s_{2i}^2 + \tau_2^2 \end{bmatrix}\right\} \quad (3.7)$$

In the above notation, $\text{cov}(y_{1i}, y_{2i}) = \rho_{W12}s_{1i}s_{2i}$ is the within studies covariance (with ρ_{W12} being the within studies correlation that has to be known beforehand) whereas $\rho_{B12}\tau_1\tau_2$ is the between studies covariance of the random terms that is to be estimated from the data (ρ_{B12} is thus the between-studies correlation). In the general models of bivariate meta-analysis, the within-studies correlation is usually assumed zero or needs to be provided by the individual studies (Berkey et al, 1998; van Houwelingen et al, 1993). However, in genetic association studies, the two logORs are correlated since they both are estimating a comparison against the same baseline category (the individuals carrying AA genotype) and the within-studies correlation can be computed analytically using the genotype counts from the individual studies. The covariance (and hence the correlation) between the two logORs, can be derived by treating the observed counts in each 2×3 table representing a single study, as independent Poisson variables with $E[Y_i] = \text{var}[Y_i] = Y_i$ and the logORs as contrasts among the log counts. Then, using the delta-method and simple calculations we can compute the variance (for details, see Appendix I):

$$\text{cov}\left(y_{1i}, y_{2i}\right) = 1/AA_{0i} + 1/AA_{1i} \quad (3.8)$$

and the correlation:

$$\rho_{W12} = \left(1/AA_{0i} + 1/AA_{1i}\right)/s_{1i}s_{2i} \quad (3.9)$$

The result is quite intuitive since the covariance is equal to the amount of variance attributed to the common (shared) baseline group. A similar formula has been used in meta-analysis of dose-response epidemiological data, without however giving the details of the derivation (Berrington & Cox, 2003; Greenland & Longnecker, 1992). In Appendix III, a Stata program for fitting the model of Equation 3.7 is presented using the `ml` command. Similar results can be obtained using the `mvmeta` command (White, 2008).

3.2 The general multivariate model of meta-analysis for continuous outcomes

In case we have a continuous outcome such as the data presented in Table 2, from each study we will be provided with the mean and standard deviation for the outcome per genotype. In general, if we denote by y_{ij} the outcome of a person j in study i (which is assumed to be distributed normally) and use dummy variables such as z_{1ij} and z_{2ij} for individuals carrying the AB and BB genotype respectively, we can formulate the linear model:

$$y_{ij} = \alpha_0 + \alpha_i + \beta_1 z_{1ij} + \beta_2 z_{2ij} \quad (3.10)$$

If the data are in the form of Table 2, the observations are the studies and need to be weighted by the inverse of their variance. If however, we have access to individual patients' data we can use them directly in the linear model. In this model, α_0 is the overall mean associated with genotype AA which is considered the reference group and α_i is the study-specific fixed effects needed to preserve stratification by study. The coefficients β_1 and β_2 are of main interest here, representing the average differences of individuals carrying the AB and BB genotypes from the AA genotype respectively. This model was considered also by Thakkinstian and coworkers (Thakkinstian et al, 2005), without however extending it to include random effects. A random effects extension of the model can be formulated introducing random coefficients β_{1i} and β_{2i} for the genotypes:

$$y_{ij} = \beta_0 + \beta_i + \left(\beta_{1i} + \beta_1\right) z_{1ij} + \left(\beta_{2i} + \beta_2\right) z_{2ij} \quad (3.11)$$

The random coefficients are considered to be distributed normally with:

$$\begin{bmatrix} \beta_{1i} \\ \beta_{2i} \end{bmatrix} \sim MVN\left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_1^2 & \rho\tau_1\tau_2 \\ \rho\tau_2\tau_1 & \tau_2^2 \end{bmatrix} \right\}$$

Adding a random intercept for studies would result in linear mixed model, however such an option is not widely accepted for meta-analysis (see section 4). Models similar to that of Equation (3.11) have been proposed for meta-analysis of continuous outcome data from clinical trials (where covariates z, would be the treatment arm) but according to the author's knowledge have never been applied

in genetic association studies (Higgins et al, 2001). The particular model can be fitted in any statistical package capable of fitting random-coefficient (mixed) weighted regression models, such as SAS (using `PROC MIXED`), R (using `lme`) or Stata (using `gllamm`). In Appendix III illustrative code in Stata is presented for fitting the particular model.

A slightly different approach can be followed however, modeling directly the pairwise differences of the outcomes arising from individuals carrying the two mutant genotypes (AB, BB) from the reference genotype (AA):

$$y_{1i} = y_{ABi} - y_{AAi} \text{ and } y_{2i} = y_{BBi} - y_{AAi}$$

It is straightforward to show that the two differences are distributed normally with mean equal to the difference of the means and variance equal to the sum of the variances. Hence, it is obvious that the bivariate model presented in Equation (3.7) can be directly applied in order to perform bivariate modeling of the two differences. Similarly, the two quantities are correlated since they are both estimating a difference from a baseline (reference) category. By simple probability calculations and noting that the random variables $y_{AA\text{i}}$, $y_{AB\text{i}}$ and $y_{AB\text{i}}$ are mutually independent, we get the covariance as follows:

$$\begin{aligned}\operatorname{cov}\left(y_{1i}, y_{2i}\right) &= \operatorname{cov}\left(y_{ABi} - y_{AAi}, y_{BBi} - y_{AAi}\right) \\ &= \operatorname{cov}\left(y_{ABi}, y_{BBi}\right) - \operatorname{cov}\left(y_{ABi}, y_{AAi}\right) \\ &\quad - \operatorname{cov}\left(y_{AAi}, y_{BBi}\right) + \operatorname{cov}\left(y_{AAi}, y_{AAi}\right) \\ &= \operatorname{var}\left(y_{AAi}\right) = \frac{sd_{AAi}^2}{AA_i} \end{aligned} \quad (3.12)$$

Similarly to the discrete outcome measures in Equation (3.9) the covariance is equal to the amount of variance attributed to the baseline group.

The fixed effect model in equation (3.10) is completely equivalent to the fixed effects analogue of the model of equation (3.7) for continuous outcomes. However, the random-effects counterparts are likely to yield different results in situations where there is a large between studies heterogeneity, since model 3.10 calculates a common intercept (representing the baseline genotype AA) whereas model 3.7 calculates directly the pairwise differences (AB-AA, BB-AA).

In conclusion, we have seen that either we have a discrete outcome arising from genetic association case-control studies or a continuous outcome derived from cross-sectional genetic association studies, the general model described in Equation (3.7) can be applied in order to draw simultaneously conclusions about the correlated outcomes. This model can be fitted in any statistical package capable of fitting random-effects weighted regression models with an arbitrary covariance matrix, such as SAS (using `PROC MIXED` or `PROC NLMIXED`), R (using `lme`) or Stata (using `mvmeta`). Stata code for fitting the model is presented in Appendix III. `mvmeta` performs inferences based on either

Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML), by direct maximization of the approximate likelihood using a Newton-Raphson algorithm (White, 2008). Additionally in the Appendix, a program for fitting the same model using the `ml` command is presented mainly for pedagogical reasons as well as for making clear the connections to the genetic-model free approach which is presented below.

3.3 The genetic model-free approach

The genetic model-free approach of Minelli and coworkers (Minelli et al, 2005b) is an extension of the general bivariate model using summary data. It consists of joint modeling of β_2 and λ which is the ratio of β_1 and β_2. The marginal model is:

$$\begin{bmatrix} y_{1i} \\ y_{2i} \end{bmatrix} \sim MVN\left\{ \begin{bmatrix} \lambda\beta_2 \\ \beta_2 \end{bmatrix}, \begin{bmatrix} s_{1i}^2 + \lambda^2\tau^2 & \rho_{W12}s_{1i}s_{2i} + \lambda\tau^2 \\ \rho_{W12}s_{1i}s_{2i} + \lambda\tau^2 & s_{2i}^2 + \tau^2 \end{bmatrix} \right\} \quad (3.13)$$

It is obvious that it is a special case of model of Equation (3.7) with re-parameterization of β_1, β_2, using $\lambda = \beta_1/\beta_2$. However, the model makes some additional assumptions since it imposes a single between-studies variance τ^2, implying this way that β_s share a common random component of variance, whereas λ is treated as a fixed-effects parameter. Minelli and coworkers in the respective publication (in the Appendix), considered also the general case of the bivariate model without explicitly defining the within-studies correlations. Furthermore, in their implementation of the general bivariate model, by acknowledging the fact that in case of small number of studies the between-studies correlation is poorly estimated, they used a fixed value of ρ_{B12}=0.9. The particular model is more parsimonious compared to the general model of Equation (3.7) since it invokes only three freely estimated parameters (i.e. λ, β_2, τ^2) compared to five (i.e. $\beta_1, \beta_2, \tau_1^2, \tau_2^2, \rho_{B12}$).

Although the authors did not consider this possibility, the model that was initially destined to be used on summary data, can also be applied in a straightforward manner to continuous outcome data, using the approach described in the previous section. Besides the theoretical discussion on the merits of the model (based on the assumptions that it makes), a major drawback of this approach is that in order to be applied, specialized code has to be written. Minelli and coworkers programmed the model in Stata using the `ml` command but the software is not widely available. In Appendix III, a Stata program that uses the `ml` command is presented for fitting the particular model. Although the implementation is (probably) different from the one used by Minelli and coworkers the results are in agreement up to the third decimal place. Minelli and coworkers, report also results from a variant of their program (i.e. bounding λ

between 0 and 1), however this version was not implemented here. The methods presented in this section use summary data and a normal approximation. In the next section, models using the binary nature of the data are going to be presented.

4. Multivariate meta-analysis using binary data

Another approach for performing multivariate meta-analysis equivalent to the one proposed in section (3.1) is to use directly the binary nature of the data. Instead of calculating the logORs and assume that they are normally distributed, we can use directly the genotype counts and perform the analysis using logistic regression. Similar models have been proposed for years for performing univariate random effects meta-analysis in a multilevel framework using logistic regression (Thompson & Sharp, 1999; Turner et al, 2000). The models that are going to be presented below can be considered as extensions of the above models in case of a multivariate response. Under the same rationale, they can be also viewed as extensions of the general multivariate approach of (van Houwelingen et al, 1993), with the difference being the fact that the two logORs are calculated from a comparison against the same baseline category.

In the binary case, the data of Table 1 are re-arranged and the 3 genotypes are encoded as j (j= 0, 1, 2) whereas we use δ_{ij} to denote the case or control status (case=0, control=1) of subjects having the j^{th} genotype in study i (i=1,2,…k). Then, we have two alternatives for modeling, either using the prospective or the retrospective likelihood. The former case, assumes a binomial sampling scheme where fixed numbers of cases and controls are sampled independently, whereas in the latter, which assumes a multinomial sampling scheme subjects are selected dependent on their disease status and then their exposure status is ascertained.

Using the prospective likelihood (the likelihood based on the probability of the disease given the exposure), the case/control status is the dependent variable and the genotypes are treated as covariates. Then, we denote $\pi_{ij} = P(\delta_{ij}=1)$ the underlying risk (i.e. the probability of being a case) of a person having the j^{th} genotype in the i^{th} study respectively. Since allele B is considered as the risk factor, a reasonable choice would be to consider the AA genotype as the reference category and create dummy variables such as z_{1i}=1 if the genotype is AB and z_{2i} =1 if the genotype is BB. This is the model described by Bagos and Nikolopoulos (Bagos & Nikolopoulos, 2007), which is formulated as:

$$\begin{aligned} \text{logit}\left(\pi_{ij}\right) &= \text{logit}\left[P\left(\delta_{ij}=1 \mid j\right)\right] \\ &= \alpha_0 + \alpha_i + \left(\beta_1 + \beta_{1i}\right) z_{1i} + \left(\beta_2 + \beta_{2i}\right) z_{2i} \end{aligned} \tag{4.1}$$

In this model, the random terms are distributed as:

$$\begin{pmatrix} \beta_{1i} \\ \beta_{2i} \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_1^2 & \rho\tau_1\tau_2 \\ \rho\tau_2\tau_1 & \tau_2^2 \end{pmatrix} \right) \tag{4.2}$$

The dummy variables α_i are indicators of the study-specific fixed-effects, whereas β_1 and β_2 obtained by fitting the model are the estimates of the logORs (AB vs. AA and BB vs. AA, respectively). Another similar approach is that of assuming beforehand an additive model across the genotypes (a co-dominant analysis). Following this rationale, the model would include as a predictor a variable z taking values 0, 1 and 2 (for AA, AB and BB genotypes respectively):

$$\text{logit}\left(\pi_{ij}\right) = \text{logit}\left[P\left(\delta_{ij} = 1 \mid j\right)\right] = \alpha_0 + \alpha_i + \left(\beta + \beta_i\right) z_i \tag{4.3}$$

This random-coefficient logistic regression model is the analogue of a univariate meta-analysis, and is thus more parsimonious. However, it will provide reliable results only in case of co-dominant inheritance. In the next section we will see that the general model of Equation (4.1) can be readily used to provide inferences for the genetic model without assuming it beforehand.

Alternatively, the model may be parameterized assuming a multinomial sampling scheme utilizing the retrospective likelihood (the likelihood based on the probability of genotypes given disease status). In this case, the genotypes are treated as dependent variables and the case/control status as the predictor in a multinomial (polytomous) logistic regression (McCullagh & Nelder, 1989):

$$p_{ij} = P\left(j \mid \delta_{ij} \right) = \frac{\exp\left(\alpha_0 + \alpha_i + \beta_j \delta_{ij}\right)}{\sum_{r=0}^{j} \exp\left(\alpha_0 + \alpha_i + \beta_r \delta_{ir}\right)} \tag{4.4}$$

By observing that the linear predictor in the above model becomes:

$$U_{ij} = \log\left(\frac{p_{ij}}{p_{i1}}\right) = \alpha_0 + \alpha_i + \beta_j \delta_{ij},\ j = 1, 2 \tag{4.5}$$

it is easy to understand that β_1 and β_2 obtained by fitting the model are estimates of the logORs (i.e. AB vs. AA and BB vs. AA, respectively) in equivalence to the respective coefficients of the model in Equation (4.1). Obviously, β_0=0 for identifiability since genotype j=0 (i.e. AA) is used as the reference category. Similar to the model based on prospective likelihood, the variables α_i are indicators of the study-specific fixed-effects. If in the above model, we introduce a genotype-specific random coefficient (for genotypes j=1, 2) the linear predictor becomes (Skrondal & Rabe-Hesketh, 2003):

$$U_{ij} = \log\left(\frac{p_{ij}}{p_{i1}}\right) = \alpha_0 + \alpha_i + \beta_j \delta_{ij} + \beta_{ij} \delta_{ij},\ j = 1, 2 \tag{4.6}$$

and the model is completely specified as a random effects bivariate meta-analysis, with random terms distributed similarly as:

$$\begin{pmatrix}\beta_{1i}\\ \beta_{2i}\end{pmatrix} \sim MVN\left(\begin{pmatrix}0\\ 0\end{pmatrix},\begin{pmatrix}\tau_1^2 & \rho\tau_1\tau_2\\ \rho\tau_2\tau_1 & \tau_2^2\end{pmatrix}\right) \tag{4.7}$$

The retrospective likelihood analogue of the model of Equation (4.3) would be an ordinal logistic regression model. Even though such models have been considered for meta-analysis of ordinal outcomes (Whitehead et al, 2001), according to the author's knowledge they have never been applied in meta-analysis of genetic association studies. It has been shown (in a fixed-effects framework) that maximum likelihood estimates obtained (with the exception of the intercept, which nevertheless is being treated as a nuisance parameter) from the retrospective likelihood are the same as those obtained from the prospective likelihood (Chen, 2003; Prentice & Pyke, 1979); this observation gave rise to the widely used approach of fitting logistic regression models for adjusting for confounders in observational studies.

Minelli and coworkers compared the retrospective with the prospective likelihood approaches in a Bayesian framework, concluding also, that the results in most cases are comparable (Minelli et al, 2005a). However, the particular approach is different from the one presented here in several aspects. First of all, it is a Bayesian method implemented using the freely available package WinBUGS (Spiegelhalter et al, 2004). Secondly, the model of Minelli and coworkers is re-parameterized involving λ, whereas the method presented here is developed following the general approach. In the next section, it will be made clear that equivalent estimates for the inheritance model (λ) could also be derived from the general model presented here. Another minor difference is the fact that in the models presented here, the study-specific effects are considered fixed parameters. Random study effects could also be added but in such case the complexity of the calculations is increased, without significant change in the accuracy of the estimates. It should be noted at this point, that usually the term "random effects" in meta-analyses refers to random treatment effects (here, treatment is the genotype) and even though methods for both random treatment and random study effects have been proposed, their use is limited and questionable; see also (Higgins et al, 2001; Turner et al, 2000) as well as the discussion in (Thompson et al, 1997; van Houwelingen & Senn, 1999).

The models presented here can be easily fitted in Stata using `gllamm`, in SAS using `PROC NLMIXED`, in R using `glmmPQL`, or utilizing specialized software for multi-level modeling such as `Mlwin` (Rasbash et al, 1998). These models are expected to perform better compared to the models presented in the previous section in case the normality assumption for logORs does not hold. Furthermore, a major advantage of these models is that can be directly used for pooled meta-analyses performed under large collaborative efforts. This is the reason why these models are usually termed Individual Patients Data (IPD)

methods (Turner et al, 2000). The disadvantage is that are significantly slower than the methods based on summary data (however, they are faster compared to their Bayesian counterparts). In Appendix III, Stata programs for fitting the two models developed in this section are presented. The models were fitted using the `gllamm` module for Stata (Rabe-Hesketh et al, 2002; Rabe-Hesketh et al, 2005). `gllamm` uses numerical integration by adaptive quadrature in order to integrate out the latent variables and obtain the marginal log-likelihood. Afterwards, the log-likelihood is maximized by Newton-Raphson using numerical first and second derivatives.

5. Inferences concerning the genetic model of inheritance

Let's now return to the bivariate model of meta-analysis described in section 3 or to the binary data methods of section 4. It would be advantageous to have general tests to assess the genetic model of inheritance without having to resort to the re-parameterization of Minelli and coworkers. Having such tests available, we will be able to make inferences concerning the genetic model using general purpose software. We will see here, that estimates of the genetic model of inheritance can be derived along with their confidence intervals and that these estimates in most cases coincide with the estimates provided by the genetic model free approach.

Once the model is fitted, comparison of the two estimates $\left(\hat{\beta}_1, \hat{\beta}_2\right)$ could provide evidence for the underlying genetic model. If both estimates are significantly different than zero and equal one to each other, a dominant model would be suggested, whereas if $\hat{\beta}_2 > \hat{\beta}_1$ a a co-dominant model will be more plausible. Of course, if $\hat{\beta}_1 = 0$ and $\hat{\beta}_2 > 0$ the recessive model would be the only choice (Sasieni, 1997). Thakkinstian and coworkers, considered these possibilities, but they did not provide formal tests in their univariate framework (Thakkinstian et al, 2005). As a matter of fact, formal statistical tests can only be provided in a multivariate framework. The most plausible test would be that of the equality of the two coefficients (i.e. $\hat{\beta}_1 = \hat{\beta}_2$). This could lead to the formulation of the following null hypothesis:

$$H_0 : d = \beta_1 - \beta_2 = 0, H_a : d \neq 0 \qquad (5.1)$$

The sample difference is normally distributed with mean $\hat{d} = \hat{\beta}_1 - \hat{\beta}_2$, whereas its variance could be calculated from:

$$\mathrm{var}\left(\hat{\beta}_1 - \hat{\beta}_2\right) = \mathrm{var}\left(\hat{\beta}_1\right) + \mathrm{var}\left(\hat{\beta}_2\right) - 2\,\mathrm{cov}\left(\hat{\beta}_1, \hat{\beta}_2\right) \qquad (5.2)$$

Thus, under H_0 the following statistic will be normally distributed:

$$\frac{\hat{d}}{\sqrt{\text{var}\left(\hat{d}\right)}} = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\sqrt{\text{var}\left(\hat{\beta}_1 - \hat{\beta}_2\right)}} \sim N\left(0,1\right) \tag{5.3}$$

and a 95% approximate confidence interval for the difference would be computed according to:

$$\hat{d} - 1.96\sqrt{\text{var}\left(\hat{d}\right)}, \hat{d} + 1.96\sqrt{\text{var}\left(\hat{d}\right)} \tag{5.4}$$

Instead of testing the equality of the two coefficients, another approach would be to test their ratio, a null hypothesis which is typically formulated as:

$$H_0 : \lambda = \frac{\beta_1}{\beta_2} = 0, H_a : \lambda \neq 0 \tag{5.5}$$

This is exactly the test statistic reported in the genetic model-free approach of Minelli, and it is perhaps more easily understood since it models simultaneously the magnitude and the significance of the two coefficients as well as their relative size. Technically, we are interested in making inferences concerning the ratio of two correlated normally distributed variables. This is a problem widely studied since the sixties (Hinkley, 1969; Marsaglia, 1965), but it is under research up to now (Marsaglia, 2006; Pham-Gia et al, 2006). It is interesting to note that in the initial publication Marsaglia (Marsaglia, 1965), was motivated by the need to calculate the distribution of the ratio of two regression coefficients. The ratio of two uncorrelated standard normal variables is distributed according to the Cauchy distribution; however the situation is complicated in case of non-zero means or in case of correlation. A simpler approach for calculating the variance of $\hat{\lambda}$ under the null hypothesis could be used if we define a function $f(\beta_1, \beta_2) = \beta_1/\beta_2$ and expand it using a bivariate 1st order Taylor approximation around the means:

$$\hat{f}\left(\beta_1, \beta_2\right) = f\left(\hat{\beta}_1, \hat{\beta}_2\right) + \frac{\partial f\left(\hat{\beta}_1, \hat{\beta}_2\right)}{\partial \beta_1}\left(\beta_1 - \hat{\beta}_1\right) + \frac{\partial f\left(\hat{\beta}_1, \hat{\beta}_2\right)}{\partial \beta_2}\left(\beta_2 - \hat{\beta}_2\right)$$

Then, by using the delta method and after replacing the population values with the sample ones, the variance will be (the details are presented in Appendix II):

$$\text{var}\left(\hat{\lambda}\right) = \frac{\text{var}\left(\hat{\beta}_1\right)}{\hat{\beta}_2^2} + \frac{\text{var}\left(\hat{\beta}_2\right)\hat{\beta}_1^2}{\hat{\beta}_2^4} - 2\,\text{cov}\left(\hat{\beta}_1, \hat{\beta}_2\right)\frac{\hat{\beta}_1}{\hat{\beta}_2^3} \tag{5.6}$$

Finally, under H_0 we will have λ=0 and thus the following statistic will be normally distributed:

$$z = \frac{\hat{\lambda}}{\sqrt{\text{var}\left(\hat{\lambda}\right)}} \sim N\left(0,1\right) \tag{5.7}$$

Consequently, a 95% confidence interval can be calculated, using:

$$\hat{\lambda} - 1.96\sqrt{\text{var}\left(\hat{\lambda}\right)}, \hat{\lambda} + 1.96\sqrt{\text{var}\left(\hat{\lambda}\right)} \quad (5.8)$$

It is interesting to notice at this point, that such problems (i.e. testing the difference of two coefficients) are usually encountered in econometrics, where a large literature is developed during the last decades. It can easily be shown that the previously developed test for *d* is a special case of the so called Wald test for "linear hypotheses" or linear restrictions (Judge et al, 1985). Similarly, the test for λ is an extension suitable for the so-called test for "non-linear hypotheses" (or non-linear restrictions). In any case, if we define a function **R** returning a $q \times 1$ vector **r** given by **R**(**b**) = **r**, then, the variance of **R**(**b**) - **r** will be equal to $\mathbf{GVG}'$

$$\text{var}\left[\mathbf{R}\left(\mathbf{b}\right) - \mathbf{r}\right] = \mathbf{GVG}' \quad (5.9)$$

where **V** is the estimated variance-covariance matrix and **G** is the derivative matrix of **R**(**b**) with respect to the vector of coefficients **b** (Green, 2008). In case of *d*, if we define: $\mathbf{R}\left(\mathbf{b}\right) = \hat{\beta}_1 - \hat{\beta}_2$, with **r**=0, we will have:

$$\mathbf{G} = \frac{\partial \mathbf{R}\left(\mathbf{b}\right)}{\partial \mathbf{b}} = \begin{bmatrix} 1 & -1 \end{bmatrix}, \text{ with: } \mathbf{G}' = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

Thus:

$$\mathbf{GVG}' = \text{var}\left(\hat{\beta}_1\right) + \text{var}\left(\hat{\beta}_2\right) - 2\,\text{cov}\left(\hat{\beta}_1, \hat{\beta}_2\right) \quad (5.10)$$

In the second case (i.e. λ), if we define a function **R** such as:

$$\mathbf{R}\left(\mathbf{b}\right) = \frac{\hat{\beta}_1}{\hat{\beta}_2}, \text{ with } \mathbf{r}=0,$$

then, the derivative matrix will be:

$$\mathbf{G} = \frac{\partial \mathbf{R}\left(\mathbf{b}\right)}{\partial \mathbf{b}} = \begin{bmatrix} 1/\hat{\beta}_2 & -\hat{\beta}_1/\hat{\beta}_2^2 \end{bmatrix}, \text{ with: } \mathbf{G}' = \begin{bmatrix} 1/\hat{\beta}_2 \\ -\hat{\beta}_1/\hat{\beta}_2^2 \end{bmatrix}$$

and finally, we will have:

$$\mathbf{GVG}' = \frac{\text{var}\left(\hat{\beta}_1\right)}{\hat{\beta}_2^2} + \frac{\text{var}\left(\hat{\beta}_2\right)\hat{\beta}_1^2}{\hat{\beta}_2^4} - 2\,\text{cov}\left(\hat{\beta}_1, \hat{\beta}_2\right)\frac{\hat{\beta}_1}{\hat{\beta}_2^3} \quad (5.11)$$

Details of the particular derivations are presented in Appendix II. The particular tests are implemented in the `testnl` and `nlcom` (`test` and `lincom` for the linear hypotheses, respectively) commands in Stata and in `delta.method` and `deltamethod` functions in R using the general formula for the delta method (5.9); however, using Equations 5.2 and 5.6 they can easily be calculated in every statistical package. As we will see in the next section, applying the last tests, results in nearly identical estimates and 95% confidence intervals for $\hat{\lambda}$ when compared to the estimates provided by the genetic model-free approach.

6. Application in published meta-analyses

In this section we present results obtained by using the methods proposed in previous sections in several already published meta-analyses of genetic association studies. We chose to include in the analysis, three published meta-analyses of case-control studies (binary outcome) that were also used as working examples by Minelli and coworkers (Minelli et al, 2005a; Minelli et al, 2005b), in order to be able to directly compare the results. These were:

- The meta-analysis of Hani and coworkers (Hani et al, 1998) concerning the association of Inwardly rectifying K+ channel (KIR 6.2) E23K polymorphism with Type II Diabetes Mellitus, which included 4 studies with 521 cases and 367 controls.
- The meta-analysis of Kato and coworkers (Kato et al, 1999) concerning the association of Angiotensinogen (AGT) M235T polymorphism with Essential Hypertension (EH), which included 7 studies with 1336 cases and 1225 controls.
- The meta-analysis of Wheeler and coworkers (Wheeler et al, 2004) concerning the association of Paraoxonase (PON1) Q192R polymorphism with Myocardial Infarction (MI), which included 19 studies with 5725 cases and 8072 controls.

Two other meta-analyses of studies involving a continuous outcome were also re-analyzed:

- The meta-analysis of Sayed-Tabatabaei and coworkers (Sayed-Tabatabaei et al, 2003) concerning the association of Angiotensin-converting enzyme gene I/D polymorphism with carotid artery wall thickness, which included 23 studies with 7795 participants.
- The individual patients' data meta-analysis of Boekholdt and coworkers (Boekholdt et al, 2005) concerning the relation of Cholesteryl Ester Transfer Protein TaqIB polymorphism with serum HDL-C levels, which included 10 studies with 13667 participants.

The results are summarized in Table 3, where the methods proposed in this work were used in comparison to the genetic model free methods of Minelli and coworkers as well as the commonly used univariate summary data methods. For reasons of brevity, fixed effects multivariate methods are not presented.

In the meta-analysis of Hani and coworkers (Hani et al, 1998), which is the one with the smaller number of included studies and individuals, Bayesian methods produce wider confidence intervals including unity for both ORs, reflecting both the uncertainty implied by the prior distribution on the parameters and the imprecision in estimating the residual heterogeneity. All other methods considered, yielded nearly identical point estimates and 95% CIs. The estimate of the genetic model (λ) is nearly identical in all cases ranging from 0.24 to 0.27,

suggesting a closer to recessive model of inheritance. In the meta-analysis concerning the relationship between the M235T AGT polymorphism and Essential hypertension (Kato et al, 1999) all methods agree that there is no significant risk associated with the heterozygous genotype, pointing indirectly to a recessive model of inheritance. The methods using the retrospective likelihood produce slightly increases estimates for OR_{BB} compared to those using the prospective likelihood, as well as, to methods using summary data. In all cases multivariate methods produce wider confidence intervals compared to univariate ones. The estimates of λ range from 0.11 to 0.25 in all case, suggesting that a recessive model is more likely. Finally, in the meta-analysis concerning the association of Paraoxonase (PON1) Q192R polymorphism with Myocardial Infarction (Wheeler et al, 2004), which included 19 studies, all methods yielded nearly identical estimates. The estimates of λ range from 0.53 to 0.69 suggesting that a co-dominant model is more likely.

The results obtained by re-analyzing the data of two meta-analyses of continuous outcome data are presented in Table 4. In the individual patients' data meta-analysis of Boekholdt and coworkers (Boekholdt et al, 2005) concerning the relation of Cholesteryl Ester Transfer Protein TaqIB polymorphism with serum HDL-Cholesterol levels, all methods produce nearly identical results indicative of absence of heterogeneity. Once again, the confidence intervals produced by the multivariate methods are slightly wider compared to the ones produced by univariate methods. The estimates of λ (0.35-0.36) suggest that a co-dominant model is more likely. Similar results were obtained using the mixed model of Equation (3.11). In the meta-analysis of Sayed-Tabatabaei and coworkers (Sayed-Tabatabaei et al, 2003) concerning the association of Angiotensin-converting enzyme gene I/D polymorphism with carotid artery wall thickness, the significant heterogeneity resulted in different estimates arising from fixed effects methods compared to random effects. Multivariate methods, once again produce wider confidence intervals for the mean difference. λ is estimated equal to 0.60 suggesting that a co-dominant model is more likely. In this meta-analysis the mixed model presented in Equation (3.11) produces different results compared to the bivariate model of Equation (3.7). The fixed effects analogues of both models yielded however identical results (data not shown). The reason for this discrepancy should be attributed to the large between studies variability, since this analysis includes studies performed on low-risk (15 studies) as well as high-risk populations (8 studies). Whereas the mean AWT of persons carrying the QQ (i.e. AA) genotype in the low risk populations is 7.2×10^{-1} mm, the respective value in high risk populations is 9.2×10^{-1} mm suggesting a significant heterogeneity. Furthermore, it seems that there is a kind of interaction between the type of population (high- or low-risk) with the genotype effects. Including in the mixed model the type of the population (low/high) as a predictor, the results were

unaffected but the estimate for the risk variable was highly significant. By analyzing separately the two group of studies, the bivariate ML model yielded insignificant results for both genotypes (for AB: 0.0494 with 95% CI: -0.04801, 0.1471 and for BB: 0.0954 with 95% CI: -0.0244, 0.2152) concerning the 15 studies of low-risk populations, whereas the same model yielded highly significant results for the 8 studies of high-risk populations (for AB: 0.4258, with 95% CI: 0.1122, 0.7393 and for BB: 0.6872 with 95% CI: 0.2841, 1.090). The mixed model for the same analyses produces somewhat different results since in the low-risk populations yields significant findings (for AB: 0.1952, with 95% CI: 0.0763, 0.3141 and for BB: 0.2485 with 95% CI: 0.1373, 0.3597) whereas for the high-risk populations the significance as well as the magnitude of the estimates are larger (for AB: 0.5059, with 95% CI: 0.2619, 0.7501 and for BB: 1.020 with 95% CI: 0.5265, 1.5136). In this meta-analysis, the underlying risk of the population seems to be an important source of heterogeneity that should be taken into account. Furthermore, it seems that this variability is extended in the genetic-model since the estimates of λ, appear to move towards the co-dominant inheritance (i.e. λ=0.5) when considering the low risk populations.

7. Discussion

Methods for multivariate meta-analysis have been proposed for years either based on summary (van Houwelingen et al, 2002) or on individual patients' data (van Houwelingen et al, 1993). It has been for long argued that multiple outcomes from the same study should be analyzed this way in order to account for the within studies correlations. Even though it is widely recognized that data arising from genetic association studies in the form of ORs derived from genotype contrasts are inherently correlated, no detailed adaptation of the multivariate methodology has been presented in the past. Most importantly, the vast majority of meta-analysis of genetic association studies is performed using univariate techniques (Attia et al, 2003; Ioannidis & Trikalinos, 2005; Ioannidis et al, 2003). In this work, the general bivariate model for random-effects meta-analysis was modified in order to be applied to genetic association studies.

Methods based on aggregate (summary) data as well as methods based on the binary nature of the data were considered. Concerning aggregate data, the model was defined with the same formulation for both discrete and continuous outcomes and analytical expressions for the covariance of the two jointly modeled outcomes were derived. The methods presented here based on summary data, resemble closely the general multivariate methods proposed by Berkey and coworkers (Berkey et al, 1998), and van Houwelingen and coworkers (van Houwelingen et al, 2002). However, the major difference is that in these methods, the within studies correlation should have been made available from the original

articles, whereas in genetic association studies it can be derived from the published genotype counts as we described. Trikalinos and Olkin (Trikalinos & Olkin, 2008), recently proposed a similar framework for multivariate meta-analysis of data arising from clinical trials where the outcomes are mutually exclusive and gave formulae for calculating the within studies correlation. Although very similar, the particular method was not extended to genetic association studies. The issue of within studies correlation in bivariate meta-analysis has been under extensive examination recently (Riley et al, 2007a; Riley et al, 2007b) and lately Riley and coworkers developed a very interesting method which does not distinguish between and within studies correlation (Riley et al, 2008). This model could be used in general applications when the within studies correlation is unknown, but as discussed already, this is not the case concerning genetic association studies where the correlation can be computed analytically.

Using the binary nature of the data, both prospective and retrospective likelihood based methods were proposed. These models are extensions of the general method of multivariate meta-analysis proposed by van Houwelingen and coworkers (van Houwelingen et al, 1993). However, with the presented approach, the two elements of the bivariate response are calculated compared to the same reference group. Similarly, the methods could be viewed as extensions of the univariate multilevel methods for random effects meta-analysis (Higgins et al, 2001; Thompson & Sharp, 1999; Turner et al, 2000). As we already mentioned, these models are expected to perform better compared to the models based on summary data when the normality assumption for logORs is invalid. Furthermore, a major advantage of these models is that can be directly used for pooled meta-analyses performed under large collaborative efforts. This is the reason why these models are usually termed Individual Patients Data (IPD) methods (Turner et al, 2000). A disadvantage is that these models are significantly slower than the models based on summary data.

In all of the presented models, formal tests for assessing the genetic model of inheritance were developed based on standard normal theory. In this respect, the general model was compared to a recently proposed, genetic model-free bivariate approach, and it was clearly shown that estimates provided by these two approaches are similar. It is argued that the slight differences between the estimates produced by the general framework described here and the genetic model-free approaches are not attributable solely to the different parameterization of the genetic model-free approach but also to the extra assumptions made. As already mentioned, the genetic model-free approach makes some additional assumptions, since it imposes a single between studies variance τ^2, implying this way that β_s are sharing a common random component of variance, whereas λ is treated as a fixed-effects parameter. Minelli and coworkers in the respective publication, considered also the general case of the bivariate model, but in their

implementation, they acknowledged the fact that in case of small number of studies, the between studies correlation is poorly estimated and thus they used a fixed value of 0.9. These assumptions made by Minelli and coworkers, could be responsible for the slight difference in the estimates produced in some cases by both model.

Another advantage of the proposed methods is the fact that they can be readily extended to studies of loci having more than two alleles such as the apolipoprotein E alleles (Song et al, 2004). Assuming n alleles, the possible genotypes (combinations of 2 with replacement) would be equal to $n(n+1)/2$. In such a situation, the multiple genotypes can simply be incorporated into the multivariate models presented in sections 3 and 4. The only limitation is that methods using the binary nature of the data (section 4) are very time-consuming in such an application. However, the summary-data methods of section 3 converge in less than a minute in a standard PC providing an attractive approach for analysing such data. Analysis of the data concerning the association of Apolipoprotein E (ApoE) alleles with Coronary Heart Disease (Wilson et al, 1996), where the three alleles of ApoE, resulting in 6 possible genotypes modeled simultaneously was performed, providing very encouraging results (data not shown).

Other multivariate methods have also been proposed for meta-analysis of genetic association studies, however in a different context from the one discussed here. In the Mendelian Randomization approach, bivariate modeling has been proposed in order to account for the multiple pairwise correlations between phenotype-genotype and genotype-disease associations, in order to finally decipher the association between phenotype and disease (Minelli et al, 2004; Thompson et al, 2005). Under this model, one element of the bivariate response is the logOR for the association of a contrast of genotypes (i.e. AB+BB vs. AA) with the disease, whereas the other is an estimate of the genotype-phenotype association for the same contrast of genotypes (i.e. the mean difference between plasma levels of a metabolite in persons carrying AB+BB and AA genotypes). Under this perspective, the methodology is applicable only in few situations (when the intermediate phenotype is known and data are available) and provides no information for the genetic model of inheritance.

Salanti and coworkers described a multivariate Bayesian method for meta-analysis adjusting for deviations from HWE (Salanti et al, 2006). They used the retrospective likelihood and parameterized the model assuming a priori a genetic model of inheritance (dominant, recessive etc). Similar to the other Bayesian methods, this approach cannot be widely used by primary researchers, whereas the a priori assumption concerning the inheritance model may be problematic and the authors suggested that their model could be modified according to the method of Minelli and coworkers. Very recently, Thompson and

coworkers extended the Bayesian genetic model free method to incorporate deviations from HWE (Thompson et al, 2008). Salanti and Higgins developed another special-purpose multivariate method with which meta-analysis can be performed when data from some of the included studies are reported as merged genotypes (Salanti & Higgins, 2008).

8. Conclusions

In this work, general multivariate methods for meta-analysis of genetic association studies based on aggregate (summary) data as well as methods based on the binary nature of the data were considered. Concerning aggregate data, the model was defined with the same formulation for both discrete and continuous outcomes and analytical expressions for the covariance of the two jointly modeled outcomes were derived for both cases. Similar models were presented in a linear mixed model formulation that allows also the use of individual patients' data in case of continuous outcome. Using the binary nature of the data, both prospective and retrospective likelihood based methods were considered. The general methods presented here were compared against some advanced methods for multivariate meta-analysis (i.e. the genetic model free approach) and the theoretical differences and similarities were highlighted. In all models considered here under the general framework, estimates for the genetic model of inheritance can be derived using standard normal theory. The methods developed here as well as the tests, are easily implemented in all major statistical packages, escaping the need of self written software. The methods were applied in several already published meta-analyses of genetic association studies (with both discrete and continuous outcomes) and the results are compared against the widely used univariate approaches as well as against the genetic model free approaches. Illustrative examples of code in Stata are given in the Appendix. It is anticipated that the methods developed in this work will be widely applied in the meta-analysis of genetic association studies.

Table 3. Results obtained from re-analyzing the data of the three published meta-analyses concerning dichotomous outcomes. The results obtained by using traditional summary data methods as well the various multivariate methods described in the text, are presented. Multivariate fixed effects methods are not presented for reasons of brevity. In all cases, the two Odds Rations (ORs) for homozygous and heterozygous for the mutant allele individuals, the estimate of the genetic model of inheritance (λ) and the respective 95% confidence intervals are listed. FE: fixed effects; RE: random effects, ML: Maximum Likelihood; REML: Restricted Maximum Likelihood; MM-DL: Method of moments of Dersimonian and Laird.

	Meta-analysis								
	KIR 6.2 E23K and Type II Diabetes Mellitus (Hani et al, 1998)			AGT M235T and Essential Hypertension (Kato et al, 1999)			PON1 Q192R and Myocardial Infarction (Wheeler et al, 2004)		
Method	**OR_{BB} (95% C.I.)**	**OR_{AB} (95% C.I.)**	**λ (95% C.I.)**	**OR_{BB} (95% C.I.)**	**OR_{AB} (95% C.I.)**	**λ (95% C.I.)**	**OR_{BB} (95% C.I.)**	**OR_{AB} (95% C.I.)**	**λ (95% C.I.)**
Univariate methods									
Summary methods (FE)	2.21 (1.43, 3.40)	1.22 (0.90, 1.64)	-	1.58 (1.06, 2.35)	1.16 (0.77, 1.76)	-	1.15 (1.01, 1.31)	1.09 (1.01, 1.17)	-
Summary methods (RE, MM-DL)	2.21 (1.43, 3.40)	1.24 (0.85, 1.80)	-	1.58 (1.06, 2.35)	1.16 (0.77, 1.76)	-	1.14 (0.99, 1.32)	1.10 (0.99, 1.23)	-
Summary methods (RE, ML)	2.21 (1.43, 3.40)	1.23 (0.88, 1.72)	-	1.58 (1.06, 2.35)	1.16 (0.77, 1.76)	-	1.15 (1.00, 1.31)	1.10 (1.00, 1.21)	-
Summary methods (RE, REML)	2.21 (1.43, 3.40)	1.24 (0.84, 1.81)	-	1.58 (1.06, 2.35)	1.16 (0.77, 1.76)	-	1.15 (1.00, 1.31)	1.10 (0.99, 1.21)	-
multivariate methods									
Bivariate regression (ML)	2.16 (1.40, 3.33)	1.23 (0.87, 1.73)	0.27 (-0.15, 0.69)	1.77 (1.10, 2.86)	1.09 (0.69, 1.74)	0.16 (-0.56, 0.88)	1.15 (1.00, 1.31)	1.09 (1.00, 1,20)	0.65 (-0.17, 1.47)
Bivariate regression (REML)	2.17 (1.41, 3.35)	1.24 (0.84, 1.83)	0.27 (-0.21, 0.76)	1.78 (1.09, 2.91)	1.09 (0.69,1.75)	0.15 (-0.57, 0.88)	1.14 (1.00, 1.32)	1.09 (0.99, 1.20)	0.66 (-0.20, 1.53)
Logistic regression (ML-prospective)	2.15 (1.40, 3.30)	1.22 (0.91, 1.63)	0.25 (-0.10, 0.61)	1.72 (1.13, 2.62)	1.06 (0.67, 1.66)	0.11 (-0.67, 0.88)	1.14 (1.00, 1.29)	1.09 (1.01, 1.19)	0.69 (-0.25, 1.62)
Multinomial logistic regression (ML-retrospective)	2.15 (1.40, 3.30)	1.21 (0.91, 1.63)	0.25 (-0.10, 0.61)	1.87 (1.16, 3.02)	1.17 (0.76, 1.78)	0.25 (-0.32, 0.81)	1.15 (1.02, 1.30)	1.09 (1.01, 1.17)	0.59 (-0.03, 1.20)
Model-free approach (ML-unbounded λ)	2.14 (1.39, 3.29)	1.21 (0.90, 1.63)	0.25 (-0.11, 0.61)	1.64 (0.99, 2.72)	1.00 (0.66, 1.53)	0.01 (-0.83, 0.85)	1.17 (1.04, 1.33)	1.08 (1.00, 1.17)	0.53 (-0.03, 1.13)
Model-free approach (ML-bounded λ)	2.14 (1.43, 3.29)	1.21 (1.08, 1.63)	0.25 (0.00, 0.69)	1.64 (1.15, 3.05)	1.00 (1.00, 1.62)	0.01 (0.00, 0.52)	1.17 (1.04, 1.33)	1.08 (1.01, 1.17)	0.53 (0.09, 1.00)
Bayesian model-free approach (prospective)	2.01 (0.97, 4.09)	1.16 (0.99, 1.77)	0.24 (0.00, 0.81)	1.81 (1.05, 3.66)	1.08 (1.00, 1.74)	0.16 (0.00, 0.54)	1.15 (1.01, 1.33)	1.08 (1.00, 1.21)	0.63 (0.03, 0.99)
Bayesian model-free approach (retrospective)	2.03 (0.96, 3.96)	1.16 (0.99, 1.80)	0.24 (0.00, 0.79)	1.83 (1.06, 3.60)	1.09 (1.00, 1.71)	0.17 (0.00, 0.54)	1.15 (1.02, 1.34)	1.08 (1.00, 1.21)	0.62 (0.03; 0.99)

Table 4. Results obtained from re-analyzing the data of the two published meta-analyses concerning continuous outcomes. The results obtained by using traditional summary data methods as well the various multivariate methods described in the text, are presented. Multivariate fixed effects methods are not presented for reasons of brevity. The average differences (β) for homozygous and heterozygous for the mutant allele individuals, the estimate of the genetic model of inheritance (λ) and the respective 95% confidence intervals are listed. FE: fixed effects; RE: random effects, ML: Maximum Likelihood; REML: Restricted Maximum Likelihood; MM-DL: Method of moments of Dersimonian and Laird.

	Meta-analysis					
	CETP TaqIB polymorphism and serum HDL-C levels (mmol/L) (Boekholdt et al, 2005)			ACE I/D polymorphism and carotid artery wall thickness ($x10^{-1}$ mm) (Sayed-Tabatabaei et al, 2003)		
Method	β_{BB} **(95% C.I.)**	β_{AB} **(95% C.I.)**	λ **(95% C.I.)**	β_{BB} **(95% C.I.)**	β_{AB} **(95% C.I.)**	λ **(95% C.I.)**
Univariate methods						
Summary methods (FE)	0.1118 (0.0980, 0.1255)	0.0397 (0.0297, .0497)	-	0.1136 (0.0351,0.1915)	0.0672 (0.000, 0.133)	-
Summary methods (RE, MM-DL)	0.1118 (0.0980, 0.1255)	0.0397 (0.0297, .0497)	-	0.2080 (0.0738, 0.3423)	0.1015 (0.0085, 0.1945)	-
Summary methods (RE, ML)	0.1118 (0.0980, 0.1255)	0.0397 (0.0297, 0.0497)	-	0.1905 (0.0696, 0.3114)	0.0918 (0.0072, 0.1763)	-
Summary methods (RE, REML)	0.1118 (0.0980, 0.1255)	0.0399 (0.0297, 0.0499)	-	0.1968 (0.0714, 0.3223)	0.0947 (0.0077, 0.1816)	-
multivariate methods						
Bivariate regression (ML)	0.1107 (0.0964, 0.1250)	0.0397 (0.0288, 0.0506)	0.36 (0.26, 0.45)	0.2001 (0.0549, 0.3452)	0.1192 (0.0059, 0.2325)	0.60 (0.22, 0.97)
Bivariate regression (REML)	0.1112 (0.0960, 0.1265)	0.0394 (0.0281, 0.0508)	0.35 (0.25, 0.46)	0.2099 (0.0614, 0.3584)	0.1269 (0.0114, 0.2425)	0.60 (0.24, 0.97)
Random coefficient model (ML)	0.1151 (0.1002, 0.1299)	0.0378 (0.0236, 0.0521)	0.33 (0.20, 0.45)	0.4704 (0.2149, 0.7258)	0.2210 (0.0733, 0.3687)	0.47 (0.22, 0.72)
Model-free approach (ML-unbounded λ)	0.1103 (0.0967, 0.1241)	0.0397 (0.0298, 0.0497)	0.36 (0.28, 0.44)	0.1999 (0.0548, 0.3451)	0.1191 (0.0056, 0.2325)	0.60 (0.22, 0.97)

Appendix I

Below, is the proof for the covariance of Equation 3.8. The covariance (and hence the correlation) between the two logORs, can be derived by treating the observed counts in each 2×3 table representing a single study, as independent Poisson variables with $E[Y_i] = var[Y_i] = Y_i$ and the logORs as contrasts among the log counts. Then, using the delta-method we can compute the variance by:

$$\text{var}\left[f\left(Y\right)\right] \approx \text{var}\left(Y\right)\left(\frac{\partial f\left(E\left[Y\right]\right)}{\partial\left(E\left[Y\right]\right)} \right)^2$$

Thus:

$$\text{var}\left[\log\left(Y_i\right)\right] \approx \text{var}\left(Y_i\right)\left(\frac{\partial \log\left(E\left[Y_i\right]\right)}{\partial\left(E\left[Y_i\right]\right)} \right)^2 = Y_i\left(\frac{\partial \log\left(Y_i\right)}{\partial Y_i} \right)^2$$

$$= Y_i\left(\frac{1}{Y_i}\frac{\partial Y_i}{\partial Y_i} \right)^2 = Y_i\left(\frac{1}{Y_i} \right)^2 = \frac{1}{Y_i}$$

For calculating the covariance $\text{cov}\left(y_{1i}, y_{2i}\right)$, we also make use of the following identities:

1) $\text{cov}\left(aX + bY, cW + dV\right) = ac\,\text{cov}\left(X, W\right) + ad\,\text{cov}\left(X, V\right)$
$+bc\,\text{cov}\left(Y, W\right) + bd\,\text{cov}\left(Y, V\right)$
2) $\text{cov}\left(Y_i, Y_i\right) = \text{var}\left(Y_i\right)$, and
3) $\text{cov}\left(Y_i, Y_j\right) = 0, \forall i \neq j$

Finally, we have:

$$\begin{aligned}
\text{cov}\left(y_{1i}, y_{2i}\right) &= \text{cov}\left[\log\left(AB_{1i}AA_{0i}\right) - \log\left(AA_{1i}AB_{0i}\right), \log\left(BB_{1i}AA_{0i}\right) - \log\left(AA_{1i}BB_{0i}\right)\right] \\
&= \text{cov}\left[\log\left(AB_{1i}AA_{0i}\right), \log\left(BB_{1i}AA_{0i}\right)\right] - \text{cov}\left[\log\left(AB_{1i}AA_{0i}\right), \log\left(AA_{1i}BB_{0i}\right)\right] \\
&\quad - \text{cov}\left[\log\left(AA_{1i}AB_{0i}\right), \log\left(BB_{1i}AA_{0i}\right)\right] + \text{cov}\left[\log\left(AA_{1i}AB_{0i}\right), \log\left(AA_{1i}BB_{0i}\right)\right] \\
&= \text{cov}\left[\log\left(AB_{1i}AA_{0i}\right), \log\left(BB_{1i}AA_{0i}\right)\right] + \text{cov}\left[\log\left(AA_{1i}AB_{0i}\right), \log\left(AA_{1i}BB_{0i}\right)\right] \\
&= \text{cov}\left[\log\left(AB_{1i}\right) - \log\left(AA_{0i}\right), \log\left(BB_{1i}\right) - \log\left(AA_{0i}\right)\right] \\
&\quad + \text{cov}\left[\log\left(AA_{1i}\right) - \log\left(AB_{0i}\right), \log\left(AA_{1i}\right) - \log\left(BB_{0i}\right)\right]
\end{aligned}$$

$$
\begin{aligned}
&= \operatorname{cov}\left[\log\left(AB_{1i}\right), \log\left(BB_{1i}\right)\right] - \operatorname{cov}\left[\log\left(AB_{1i}\right), \log\left(AA_{0i}\right)\right] \\
&\quad - \operatorname{cov}\left[\log\left(AA_{0i}\right), \log\left(BB_{1i}\right)\right] + \operatorname{cov}\left[\log\left(AA_{0i}\right), \log\left(AA_{0i}\right)\right] \\
&\quad + \operatorname{cov}\left[\log\left(AA_{1i}\right), \log\left(AA_{1i}\right)\right] - \operatorname{cov}\left[\log\left(AA_{1i}\right), \log\left(BB_{0i}\right)\right] \\
&\quad - \operatorname{cov}\left[\log\left(AB_{0i}\right), \log\left(AA_{1i}\right)\right] + \operatorname{cov}\left[\log\left(AB_{0i}\right), \log\left(BB_{0i}\right)\right] \\
&= \operatorname{cov}\left[\log\left(AA_{0i}\right), \log\left(AA_{0i}\right)\right] + \operatorname{cov}\left[\log\left(AA_{1i}\right), \log\left(AA_{1i}\right)\right] \\
&= \operatorname{var}\left[\log\left(AA_{0i}\right)\right] + \operatorname{var}\left[\log\left(AA_{1i}\right)\right] \\
&= \frac{1}{AA_{0i}} + \frac{1}{AA_{1i}}
\end{aligned}
$$

Appendix II

To calculate the variance of $\hat{\lambda}$, we have to define it as a function $f(\beta_1, \beta_2) = \beta_1/\beta_2$ and expand it using a bivariate 1st order Taylor approximation around the means $\left(\hat{\beta}_1, \hat{\beta}_2\right)$:

$$
\hat{f}\left(\beta_1, \beta_2\right) = f\left(\hat{\beta}_1, \hat{\beta}_2\right) + \frac{\partial f\left(\hat{\beta}_1, \hat{\beta}_2\right)}{\partial \beta_1}\left(\beta_1 - \hat{\beta}_1\right) + \frac{\partial f\left(\hat{\beta}_1, \hat{\beta}_2\right)}{\partial \beta_2}\left(\beta_2 - \hat{\beta}_2\right)
$$

Then, using the delta method we will have:

$$
\begin{aligned}
\operatorname{var}\left[f\left(\beta_1, \beta_2\right)\right] &\approx \operatorname{var}\left[\hat{f}\left(\beta_1, \beta_2\right)\right] \\
&= \operatorname{var}\left[\frac{\partial f\left(\hat{\beta}_1, \hat{\beta}_2\right)}{\partial \beta_1}\left(\beta_1 - \hat{\beta}_1\right) + \frac{\partial f\left(\hat{\beta}_1, \hat{\beta}_2\right)}{\partial \beta_2}\left(\beta_2 - \hat{\beta}_2\right)\right] \\
&= \operatorname{var}\left[\frac{\partial f\left(\hat{\beta}_1, \hat{\beta}_2\right)}{\partial \beta_1}\left(\beta_1 - \hat{\beta}_1\right)\right] + \operatorname{var}\left[\frac{\partial f\left(\hat{\beta}_1, \hat{\beta}_2\right)}{\partial \beta_2}\left(\beta_2 - \hat{\beta}_2\right)\right] \\
&\quad + 2\operatorname{cov}\left[\frac{\partial f\left(\hat{\beta}_1, \hat{\beta}_2\right)}{\partial \beta_1}\left(\beta_1 - \hat{\beta}_1\right), \frac{\partial f\left(\hat{\beta}_1, \hat{\beta}_2\right)}{\partial \beta_2}\left(\beta_2 - \hat{\beta}_2\right)\right]
\end{aligned}
$$

Thus:

$$\operatorname{var}\left[f\left(\beta_1,\beta_2\right)\right] \approx \left[\frac{\partial f\left(\hat{\beta}_1,\hat{\beta}_2\right)}{\partial \beta_1}\right]^2 \operatorname{var}\left(\beta_1\right) + \left[\frac{\partial f\left(\hat{\beta}_1,\hat{\beta}_2\right)}{\partial \beta_2}\right]^2 \operatorname{var}\left(\beta_2\right) + 2\operatorname{cov}\left(\beta_1,\beta_2\right)\frac{\partial f\left(\hat{\beta}_1,\hat{\beta}_2\right)}{\partial \beta_1}\frac{\partial f\left(\hat{\beta}_1,\hat{\beta}_2\right)}{\partial \beta_2}$$

and after replacing the population values with the sample ones, the variance will be:

$$\operatorname{var}\left(\hat{\lambda}\right) = \frac{\operatorname{var}\left(\hat{\beta}_1\right)}{\hat{\beta}_2^2} + \operatorname{var}\left(\hat{\beta}_2\right)\frac{\hat{\beta}_1^2}{\hat{\beta}_2^4} - \operatorname{cov}\left(\hat{\beta}_1,\hat{\beta}_2\right)\frac{\hat{\beta}_1}{\hat{\beta}_2^3} - \frac{\hat{\beta}_1}{\hat{\beta}_2^3}\operatorname{cov}\left(\hat{\beta}_1,\hat{\beta}_2\right)$$

$$= \frac{\operatorname{var}\left(\hat{\beta}_1\right)}{\hat{\beta}_2^2} + \frac{\operatorname{var}\left(\hat{\beta}_2\right)\hat{\beta}_1^2}{\hat{\beta}_2^4} - 2\operatorname{cov}\left(\hat{\beta}_1,\hat{\beta}_2\right)\frac{\hat{\beta}_1}{\hat{\beta}_2^3}$$

It can easily be shown that the test for *d* is a special case of the so called Wald test for "linear hypotheses" or linear restrictions (Judge et al, 1985). Similarly, the test for λ is an extension suitable for the so-called test for "non-linear hypotheses" (or non-linear restrictions). In any case, if we define a function **R** returning a $q \times 1$ vector **r** given by **R**(**b**) = **r**, then, the variance of **R**(**b**) - **r** will be equal to $\mathbf{GVG}'$

$$\operatorname{var}\left[\mathbf{R}\left(\mathbf{b}\right)-\mathbf{r}\right]=\mathbf{GVG}'$$

where **G**, is the derivative matrix of **R**(**b**) with respect to **b**. From the estimated model we will have:

$$\mathbf{b} = \begin{bmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{bmatrix} \text{ and } \mathbf{V} = \begin{bmatrix}\operatorname{var}\left(\hat{\beta}_1\right) & \operatorname{cov}\left(\hat{\beta}_1,\hat{\beta}_2\right)\\ \operatorname{cov}\left(\hat{\beta}_1,\hat{\beta}_2\right) & \operatorname{var}\left(\hat{\beta}_2\right)\end{bmatrix}$$

In case of *d*, if we define: $\mathbf{R}\left(\mathbf{b}\right) = \hat{\beta}_1 - \hat{\beta}_2$, with **r**=0, we will have:

$$\mathbf{G} = \frac{\partial \mathbf{R}\left(\mathbf{b}\right)}{\partial \mathbf{b}} = \begin{bmatrix}\frac{\partial \mathbf{R}\left(\mathbf{b}\right)}{\partial \hat{\beta}_1} & \frac{\partial \mathbf{R}\left(\mathbf{b}\right)}{\partial \hat{\beta}_2}\end{bmatrix} = \begin{bmatrix}\frac{\partial\left(\hat{\beta}_1\right)}{\partial \hat{\beta}_1} & -\frac{\partial\left(\hat{\beta}_2\right)}{\partial \hat{\beta}_2}\end{bmatrix} = \begin{bmatrix}1 & -1\end{bmatrix} \quad \text{with: } \mathbf{G}' = \begin{bmatrix}1\\ -1\end{bmatrix}$$

Thus:

$$\mathbf{GVG}' = \begin{bmatrix}1 & -1\end{bmatrix}\begin{bmatrix}\operatorname{var}\left(\hat{\beta}_1\right) & \operatorname{cov}\left(\hat{\beta}_1,\hat{\beta}_2\right)\\ \operatorname{cov}\left(\hat{\beta}_1,\hat{\beta}_2\right) & \operatorname{var}\left(\hat{\beta}_2\right)\end{bmatrix}\begin{bmatrix}1\\ -1\end{bmatrix}$$

$$= \left[\operatorname{var}\left(\hat{\beta}_1\right) - \operatorname{cov}\left(\hat{\beta}_1, \hat{\beta}_2\right) \quad \operatorname{cov}\left(\hat{\beta}_1, \hat{\beta}_2\right) - \operatorname{var}\left(\hat{\beta}_2\right) \right] \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

$$= \operatorname{var}\left(\hat{\beta}_1\right) + \operatorname{var}\left(\hat{\beta}_2\right) - 2\operatorname{cov}\left(\hat{\beta}_1, \hat{\beta}_2\right)$$

In the second case (λ), if we define a function **R** such as:

$$\mathbf{R}(\mathbf{b}) = \frac{\hat{\beta}_1}{\hat{\beta}_2}, \text{ with } \mathbf{r}=0,$$

then, the derivative matrix will be:

$$\mathbf{G} = \frac{\partial \mathbf{R}(\mathbf{b})}{\partial \mathbf{b}} = \begin{bmatrix} \frac{\partial \mathbf{R}(\mathbf{b})}{\partial \hat{\beta}_1} & \frac{\partial \mathbf{R}(\mathbf{b})}{\partial \hat{\beta}_2} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial\left(\hat{\beta}_1 / \hat{\beta}_2\right)}{\partial \hat{\beta}_1} & \frac{\partial\left(\hat{\beta}_1 / \hat{\beta}_2\right)}{\partial \hat{\beta}_2} \end{bmatrix} = \begin{bmatrix} \frac{1}{\hat{\beta}_2} & -\frac{\hat{\beta}_1}{\hat{\beta}_2^2} \end{bmatrix} \quad \text{with: } \mathbf{G}' = \begin{bmatrix} \frac{1}{\hat{\beta}_2} \\ -\frac{\hat{\beta}_1}{\hat{\beta}_2^2} \end{bmatrix}$$

and finally, we will have:

$$\mathbf{GVG}' = \begin{bmatrix} \frac{1}{\hat{\beta}_2} & -\frac{\hat{\beta}_1}{\hat{\beta}_2^2} \end{bmatrix} \begin{bmatrix} \operatorname{var}\left(\hat{\beta}_1\right) & \operatorname{cov}\left(\hat{\beta}_1, \hat{\beta}_2\right) \\ \operatorname{cov}\left(\hat{\beta}_1, \hat{\beta}_2\right) & \operatorname{var}\left(\hat{\beta}_2\right) \end{bmatrix} \begin{bmatrix} \frac{1}{\hat{\beta}_2} \\ -\frac{\hat{\beta}_1}{\hat{\beta}_2^2} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{\hat{\beta}_2} & -\frac{\hat{\beta}_1}{\hat{\beta}_2^2} \end{bmatrix} \begin{bmatrix} \frac{\operatorname{var}\left(\hat{\beta}_1\right)}{\hat{\beta}_2} - \operatorname{cov}\left(\hat{\beta}_1, \hat{\beta}_2\right)\frac{\hat{\beta}_1}{\hat{\beta}_2^2} \\ \frac{\operatorname{cov}\left(\hat{\beta}_1, \hat{\beta}_2\right)}{\hat{\beta}_2} - \operatorname{var}\left(\hat{\beta}_2\right)\frac{\hat{\beta}_1}{\hat{\beta}_2^2} \end{bmatrix}$$

$$= \frac{1}{\hat{\beta}_2}\left\{\frac{\operatorname{var}\left(\hat{\beta}_1\right)}{\hat{\beta}_2} - \operatorname{cov}\left(\hat{\beta}_1, \hat{\beta}_2\right)\frac{\hat{\beta}_1}{\hat{\beta}_2^2}\right\} - \frac{\hat{\beta}_1}{\hat{\beta}_2^2}\left\{\frac{\operatorname{cov}\left(\hat{\beta}_1, \hat{\beta}_2\right)}{\hat{\beta}_2} - \operatorname{var}\left(\hat{\beta}_2\right)\frac{\hat{\beta}_1}{\hat{\beta}_2^2}\right\}$$

$$= \frac{\operatorname{var}\left(\hat{\beta}_1\right)}{\hat{\beta}_2^2} + \operatorname{var}\left(\hat{\beta}_2\right)\frac{\hat{\beta}_1^2}{\hat{\beta}_2^4} - \operatorname{cov}\left(\hat{\beta}_1, \hat{\beta}_2\right)\frac{\hat{\beta}_1}{\hat{\beta}_2^3} - \frac{\hat{\beta}_1}{\hat{\beta}_2^3}\operatorname{cov}\left(\hat{\beta}_1, \hat{\beta}_2\right)$$

$$= \frac{\operatorname{var}\left(\hat{\beta}_1\right)}{\hat{\beta}_2^2} + \frac{\operatorname{var}\left(\hat{\beta}_2\right)\hat{\beta}_1^2}{\hat{\beta}_2^4} - 2\operatorname{cov}\left(\hat{\beta}_1, \hat{\beta}_2\right)\frac{\hat{\beta}_1}{\hat{\beta}_2^3}$$

References

Attia J, Thakkinstian A, D'Este C (2003) Meta-analyses of molecular association studies: methodologic lessons for genetic epidemiology. *J Clin Epidemiol* **56**(4)**:** 297-303

Bagos PG, Nikolopoulos GK (2007) A method for meta-analysis of case-control genetic association studies using logistic regression. *Stat Appl Genet Mol Biol* **6:** Article17

Becker KG, Barnes KC, Bright TJ, Wang SA (2004) The genetic association database. *Nat Genet* **36**(5)**:** 431-432

Berkey CS, Hoaglin DC, Antczak-Bouckoms A, Mosteller F, Colditz GA (1998) Meta-analysis of multiple outcomes by regression with random effects. *Stat Med* **17**(22)**:** 2537-2550

Berrington A, Cox DR (2003) Generalized least squares for the synthesis of correlated information. *Biostatistics* **4**(3)**:** 423-431

Boekholdt SM, Sacks FM, Jukema JW, Shepherd J, Freeman DJ, McMahon AD, Cambien F, Nicaud V, de Grooth GJ, Talmud PJ, Humphries SE, Miller GJ, Eiriksdottir G, Gudnason V, Kauma H, Kakko S, Savolainen MJ, Arca M, Montali A, Liu S, Lanz HJ, Zwinderman AH, Kuivenhoven JA, Kastelein JJ (2005) Cholesteryl ester transfer protein TaqIB variant, high-density lipoprotein cholesterol levels, cardiovascular risk, and efficacy of pravastatin treatment: individual patient meta-analysis of 13,677 subjects. *Circulation* **111**(3)**:** 278-287

Chen HY (2003) A note on the prospective analysis of outcome-dependent samples. *J Roy Soc B* **65**(2)**:** 575-584

DerSimonian R, Laird N (1986) Meta-analysis in clinical trials. *Controlled Clinical Trials* **7:** 177-188

Evangelou E, Trikalinos TA, Salanti G, Ioannidis JP (2006) Family-based versus unrelated case-control designs for genetic associations. *PLoS Genet* **2**(8)**:** e123

Green W (2008) *Econometric Analysis*, 6th Edition edn.: Prentice Hall.

Greenland S, Longnecker MP (1992) Methods for trend estimation from summarized dose-response data, with applications to meta-analysis. *Am J Epidemiol* **135**(11)**:** 1301-1309

Gu C, Province MA, Rao DC (2001) Meta-analysis for model-free methods. *Adv Genet* **42:** 255-272

Hani EH, Boutin P, Durand E, Inoue H, Permutt MA, Velho G, Froguel P (1998) Missense mutations in the pancreatic islet beta cell inwardly rectifying K+ channel gene (KIR6.2/BIR): a meta-analysis suggests a role in the polygenic basis of Type II diabetes mellitus in Caucasians. *Diabetologia* **41**(12)**:** 1511-1515

Higgins JP, Whitehead A, Turner RM, Omar RZ, Thompson SG (2001) Meta-analysis of continuous outcome data from individual patients. *Stat Med* **20**(15)**:** 2219-2241

Hinkley DV (1969) On the ratio of two correlated normal random variables. *Biometrika* **56**(3)**:** 635-639

Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K (2002) A comprehensive review of genetic association studies. *Genet Med* **4**(2)**:** 45-61

Ioannidis JP (2005a) Molecular bias. *Eur J Epidemiol* **20**(9)**:** 739-745

Ioannidis JP (2005b) Why most published research findings are false. *PLoS Med* **2**(8)**:** e124

Ioannidis JP, Trikalinos TA (2005) Early extreme contradictory estimates may appear in published research: the Proteus phenomenon in molecular genetics research and randomized trials. *J Clin Epidemiol* **58**(6)**:** 543-549

Ioannidis JP, Trikalinos TA, Ntzani EE, Contopoulos-Ioannidis DG (2003) Genetic associations in large versus small studies: an empirical assessment. *Lancet* **361**(9357)**:** 567-571

Judge GG, Griffiths WE, Hill RC, Lutkepohl H, Lee T-C (1985) *The Theory and Practice of Econometrics*, 2nd edn.: John Wiley & Sons.

Kato N, Sugiyama T, Morita H, Kurihara H, Yamori Y, Yazaki Y (1999) Angiotensinogen gene and essential hypertension in the Japanese: extensive

association study and meta-analysis on six reported studies. *J Hypertens* **17**(6)**:** 757-763

Kazeem GR, Farrall M (2005) Integrating case-control and TDT studies. *Ann Hum Genet* **69**(Pt 3)**:** 329-335

Marsaglia G (1965) Ratios of Normal Variables and Ratios of Sums of Uniform Variables. *Journal of the American Statistical Association* **60**(309)**:** 193-204

Marsaglia G (2006) Ratios of Normal Variables. *Journal of Statistical Software* **16**(4)

McCullagh P, Nelder JA (1989) *Generalized Linear Models*, London: Chapman & Hall.

Minelli C, Thompson JR, Abrams KR, Lambert PC (2005a) Bayesian implementation of a genetic model-free approach to the meta-analysis of genetic association studies. *Stat Med* **24**(24)**:** 3845-3861

Minelli C, Thompson JR, Abrams KR, Thakkinstian A, Attia J (2005b) The choice of a genetic model in the meta-analysis of molecular association studies. *Int J Epidemiol* **34**(6)**:** 1319-1328

Minelli C, Thompson JR, Tobin MD, Abrams KR (2004) An integrated approach to the meta-analysis of genetic association studies using Mendelian randomization. *Am J Epidemiol* **160**(5)**:** 445-452

Normand SL (1999) Meta-analysis: formulating, evaluating, combining, and reporting. *Stat Med* **18**(3)**:** 321-359

Petiti DB (1994) *Meta-analysis Decision Analysis and Cost-Effectiveness Analysis*, Vol. 24: Oxford University Press.

Pham-Gia T, Turkkan N, Marchand E (2006) Density of the Ratio of Two Normal Random Variables and Applications. *Communications in Statistics -Theory and Methods* **35**(9)**:** 1569 — 1591

Prentice RL, Pyke R (1979) Logistic disease incidence models and case-control studies. *Biometrika* **66**(3)**:** 403-411

Rabe-Hesketh S, Skrondal A, Pickles A (2002) Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal* **2:** 1-21.

Rabe-Hesketh S, Skrondal A, Pickles A (2005) Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics* **128**(2)**:** 301-323

Rasbash J, Cameron B, Browne W, Healy M (1998) *The MlwiN software Package* Vol. version 1.0, London, UK: Institute of Education.

Riley RD, Abrams KR, Lambert PC, Sutton AJ, Thompson JR (2007a) An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. *Stat Med* **26**(1)**:** 78-97

Riley RD, Abrams KR, Sutton AJ, Lambert PC, Thompson JR (2007b) Bivariate random-effects meta-analysis and the estimation of between-study correlation. *BMC Med Res Methodol* **7:** 3

Riley RD, Thompson JR, Abrams KR (2008) An alternative model for bivariate random-effects meta-analysis when the within-study correlations are unknown. *Biostatistics* **9**(1)**:** 172-186

Salanti G, Higgins JP (2008) Meta-analysis of genetic association studies under different inheritance models using data reported as merged genotypes. *Stat Med* **27**(5)**:** 764-777

Salanti G, Higgins JP, Trikalinos TA, Ioannidis JP (2006) Bayesian meta-analysis and meta-regression for gene-disease associations and deviations from Hardy-Weinberg equilibrium. *Stat Med*

Salanti G, Sanderson S, Higgins JP (2005) Obstacles and opportunities in meta-analysis of genetic association studies. *Genet Med* **7**(1)**:** 13-20

Sasieni PD (1997) From genotypes to genes: doubling the sample size. *Biometrics* **53**(4)**:** 1253-1261

Sayed-Tabatabaei FA, Houwing-Duistermaat JJ, van Duijn CM, Witteman JC (2003) Angiotensin-converting enzyme gene polymorphism and carotid artery wall thickness: a meta-analysis. *Stroke* **34**(7)**:** 1634-1639

Skrondal A, Rabe-Hesketh S (2003) Multilevel logistic regression for polytomous data and rankings. *Psychometrika* **68**(2)**:** 267-287

Smith TC, Spiegelhalter DJ, Thomas A (1995) Bayesian approaches to random-effects meta-analysis: a comparative study. *Stat Med* **14**(24)**:** 2685-2699

Song Y, Stampfer MJ, Liu S (2004) Meta-analysis: apolipoprotein E genotypes and risk for coronary heart disease. *Ann Intern Med* **141**(2)**:** 137-147

Spiegelhalter DJ, Thomas A, Best NG, Lunn D (2004) *WinBUGS User Manual*, Vol. Version 1.4.1. , Cambridge, U.K.: MRC Biostatistics Unit.

Sutton AJ, Abrams KR (2001) Bayesian methods in meta-analysis and evidence synthesis. *Stat Methods Med Res* **10**(4)**:** 277-303

Thakkinstian A, McElduff P, D'Este C, Duffy D, Attia J (2005) A method for meta-analysis of molecular association studies. *Stat Med* **24**(9)**:** 1291-1306

Thompson JR, Minelli C, Abrams KR, Thakkinstian A, Attia J (2008) Combining information from related meta-analyses of genetic association studies. *Appl Statist* **57**(1)**:** 103-115

Thompson JR, Minelli C, Abrams KR, Tobin MD, Riley RD (2005) Meta-analysis of genetic studies using Mendelian randomization--a multivariate approach. *Stat Med* **24**(14)**:** 2241-2254

Thompson SG, Higgins JP (2002) How should meta-regression analyses be undertaken and interpreted? *Stat Med* **21**(11)**:** 1559-1573

Thompson SG, Sharp SJ (1999) Explaining heterogeneity in meta-analysis: a comparison of methods. *Stat Med* **18**(20)**:** 2693-2708

Thompson SG, Smith TC, Sharp SJ (1997) Investigating underlying risk as a source of heterogeneity in meta-analysis. *Stat Med* **16**(23)**:** 2741-2758

Trikalinos TA, Olkin I (2008) A method for the meta-analysis of mutually exclusive binary outcomes. *Stat Med*

Turner RM, Omar RZ, Yang M, Goldstein H, Thompson SG (2000) A multilevel model framework for meta-analysis of clinical trials with binary outcomes. *Stat Med* **19**(24)**:** 3417-3432

van Houwelingen H, Senn S (1999) Investigating underlying risk as a source of heterogeneity in meta-analysis. *Stat Med* **18**(1)**:** 110-115

van Houwelingen HC, Arends LR, Stijnen T (2002) Advanced methods in meta-analysis: multivariate approach and meta-regression. *Stat Med* **21**(4)**:** 589-624

van Houwelingen HC, Zwinderman KH, Stijnen T (1993) A bivariate approach to meta-analysis. *Stat Med* **12**(24)**:** 2273-2284

Warn DE, Thompson SG, Spiegelhalter DJ (2002) Bayesian random effects meta-analysis of trials with binary outcomes: methods for the absolute risk difference and relative risk scales. *Stat Med* **21**(11)**:** 1601-1623

Wheeler JG, Keavney BD, Watkins H, Collins R, Danesh J (2004) Four paraoxonase gene polymorphisms in 11212 cases of coronary heart disease and 12786 controls: meta-analysis of 43 studies. *Lancet* **363**(9410)**:** 689-695

White IR (2008) Multivariate random-effects meta-analysis. *Stata Journal* **in press**

Whitehead A, Omar RZ, Higgins JP, Savaluny E, Turner RM, Thompson SG (2001) Meta-analysis of ordinal outcomes using individual patient data. *Stat Med* **20**(15)**:** 2243-2260

Wilson PW, Schaefer EJ, Larson MG, Ordovas JM (1996) Apolipoprotein E alleles and risk of coronary disease. A meta-analysis. *Arterioscler Thromb Vasc Biol* **16**(10)**:** 1250-1255