Published by Oxford University Press on behalf of the International Epidemiological Association
International Journal of Epidemiology 2005;34:1319–1328
doi:10.1093/ije/dyi169

The choice of a genetic model in the meta-analysis of molecular association studies

Cosetta Minelli,[1†*] John R Thompson,[1†] Keith R Abrams,[1] Ammarin Thakkinstian[2,3] and John Attia[3]

Accepted 25 July 2005

Background To evaluate gene–disease associations, genetic epidemiologists collect information on the disease risk in subjects with different genotypes (for a bi-allelic polymorphism: gg, Gg, GG). Meta-analyses of such studies usually reduce the problem to a single comparison, either by performing two separate pairwise comparisons or by assuming a specific underlying genetic model (recessive, co-dominant, dominant). A biological justification for the choice of the genetic model is seldom available.

Methods We present a genetic model-free approach, which does not assume that the underlying genetic model is known in advance but still makes use of the information available on all genotypes. The approach uses OR_{GG}, the odds ratio between the homozygous genotypes, to capture the magnitude of the genetic effect, and λ, the heterozygote log odds ratio as a proportion of the homozygote log odds ratio, to capture the genetic mode of inheritance. The analysis assumes that the same unknown genetic model, i.e. the same λ, applies in all studies, and this is investigated graphically. The approach is illustrated using five examples of published meta-analyses.

Results Analyses based on specific genetic models can produce misleading estimates of the odds ratios when an inappropriate model is assumed. The genetic model-free approach gives appropriately wider confidence intervals than genetic model-based analyses because it allows for uncertainty about the genetic model. In terms of assessment of model fit, it performs at least as well as a bivariate pairwise analysis in our examples.

Conclusions The genetic model-free approach offers a unified approach that efficiently estimates the genetic effect and the underlying genetic model. A bivariate pairwise analysis should be used if the assumption of a common genetic model across studies is in doubt.

Keywords Meta-analysis, population genetics, polymorphism, genetic models, association studies

Population-based genetic epidemiology, which evaluates the risk of a disease associated with a specific genetic polymorphism, often seeks to identify relatively small effects against a noisy background of biological and social complexity. Because of this, most genetic association studies tend to be statistically underpowered.[1,2] While the need for large-scale population-based association studies has recently been recognized,[3,4] data from such studies will not be available in the near future. In the meantime, evidence synthesis from multiple small studies has the potential to play an important role in advancing biomedical knowledge by increasing the statistical power.[5] However, the appropriate use of meta-analysis within genetic epidemiology has been researched less than might be anticipated, and the general methodological quality of published meta-analyses of genetic association studies is poor.[6]

A recent review by Attia *et al.*[6] showed how meta-analyses of genetic association studies often fail to address general meta-analytical concerns and ignore important issues specific

[1] Department of Health Sciences, Centre for Biostatistics and Genetic Epidemiology, University of Leicester, Leicester, UK.

[2] Clinical Epidemiology Unit, Faculty of Medicine, Mahidol University, Bangkok, Thailand.

[3] Centre for Clinical Epidemiology and Biostatistics, Faculty of Health, University of Newcastle, Newcastle, Australia.

[†] The first two authors contributed equally to this work.

[*] Corresponding author. Department of Health Sciences, Centre for Biostatistics and Genetic Epidemiology, University of Leicester, 22–28 Princess Road West, Leicester LE1 6TP, UK. E-mail: cm109@le.ac.uk

to gene–disease associations. General concerns include the lack of explicit reporting of inclusion and exclusion criteria, a failure to explore possible sources of heterogeneity, and the absence of an investigation of publication bias. An important aspect of the inclusion criteria for a meta-analysis is outcome definition, since differences in the way outcome is defined and measured may well explain heterogeneity of study results.[3,7] Another important source of heterogeneity is diversity in the populations studied, in particular ethnic diversity.[3] Publication bias arises because studies showing either statistically significant results or large effect sizes are often more likely to be published than negative studies,[8,9] and thus the result of a meta-analysis based on published studies may be positively biased. Publication bias is particularly important in genetic epidemiology because it is possible to study many polymorphisms on the same subjects and then to select those that are submitted for publication.[3,10–13] Although simple graphical methods such as funnel plots can be used to detect publication bias,[8,9] in the review by Attia *et al.*[6] only 20% of the meta-analyses (7 out of 37) addressed this issue.

Methodological issues that are specific to genetic epidemiology include the checking of Hardy–Weinberg equilibrium and the choice of a genetic model.[6,7] In the meta-analysis of genetic association studies there are always at least three possible genotypes to compare. This contrasts with the two treatment groups characteristic of most biomedical meta-analyses. In practice, the number of possible comparisons between genotypes is often reduced by assuming a specific genetic model, such as dominant or recessive, but the conclusions might be sensitive to this assumption.[6]

In the simplest case of a polymorphism with two alleles (G and g), one of which is thought to be associated with a disease (G), association studies will usually collect information on the numbers of diseased and disease-free subjects with each of the three genotypes (gg, Gg, and GG). To date almost all meta-analyses of genetic association studies have reduced the three groups to two by (i) ignoring the heterozygotes and comparing gg with GG, (ii) performing separate pairwise comparisons, (iii) assuming a recessive model to justify combining the gg and Gg genotypes and comparing gg + Gg with GG, (iv) assuming a dominant model and comparing gg with Gg + GG, and (v) assuming a per-allele effect that places Gg mid-way between gg and GG, also called the co-dominant model. When unsure about the genetic model, some investigators fit multiple models and/or perform pairwise comparisons. However, adjustment for multiple testing is seldom made, and the pairwise estimates of the odds ratio of GG vs gg (subsequently referred to as OR_{GG}) and the odds ratio of Gg vs gg (subsequently referred to as OR_{Gg}) are usually obtained by carrying out two separate meta-analyses, thus ignoring the correlation between the two odds ratios induced by the common baseline group.

The review by Attia *et al.*[6] showed that 24 of 37 meta-analyses based their analysis on the assumption of an underlying genetic model, with half of these testing multiple modes of inheritance or multiple pairwise comparisons. A biological justification for the choice of the genetic model was provided in only eight meta-analyses. In nine of the meta-analyses the genetic effect was tested by comparing the allele frequency in cases and controls.

All of the methods of analysis in common use, with the exception of the pairwise comparisons, make the implicit assumptions that a particular genetic model applies in all studies, and, more importantly, that the model is known in advance; for instance, the gene might be assumed to be recessive in all populations. Here we suggest a genetic model-free approach to the meta-analysis of genetic association studies that also assumes a common genetic model across studies but which does not specify the mode of inheritance in advance. The underlying genetic model is instead estimated from the data. Although no specific genetic model is assumed, the analyses are, of course, still based on an assumed statistical model. The model is based on a simple reparameterization and uses the odds ratio between the homozygous genotypes (OR_{GG}) to captures the magnitude of the genetic effect, and λ, the ratio of log OR_{Gg} and log OR_{GG}, to capture the genetic mode of inheritance. λ is assumed to be common across studies, but if this assumption is in doubt then pairwise comparisons obtained using bivariate random-effect meta-analysis methods, which take into account the correlation between OR_{GG} and OR_{Gg}, should be used.[14,15] We describe graphical and statistical ways of investigating whether the assumption of a common λ is reasonable.

Allowing λ to take any value (unbounded analysis), is equivalent to allowing the possibility of heterosis, i.e. the risk of the Gg group can be higher or lower than either of the homozygous groups. Although rare, heterosis has been described.[16,17] If this possibility can be excluded on biological grounds then it is better to constrain λ between 0 and 1 (bounded analysis); this restricts the mode of effect to the spectrum between dominant, through co-dominant, to recessive.

Methods

Genetic model-free approach: a common but unrestricted genetic model

Consider the meta-analysis of a bi-allelic polymorphism, in which G is the risk allele, and a dichotomous disease outcome is ascertained for each genotype. We define two parameters: the odds ratio between the two homozygous genotypes, OR_{GG}; and λ, the ratio of log OR_{Gg} and log OR_{GG}. The value of λ is not restricted, but values equal to 0, 0.5, and 1 correspond to the recessive, co-dominant, and dominant genetic model, respectively, and values >1 or <0 would suggest positive or negative heterosis.

Log OR_{GG} could be modelled as a fixed-effect or as a random effect that allows for heterogeneity across studies.[8] In the analyses presented, the log OR_{GG} has been modelled as a random effect except in those situations where the heterogeneity of log OR_{GG} was very close to 0. λ is modelled as a fixed-effect, that is, the genetic model is assumed to be the same in all studies. It is usually not possible to model both log OR_{GG} and λ as random effects because, without extra information, it is very difficult to disentangle the heterogeneity of λ from that of log OR_{GG}.

The two log odds ratios from each study are modelled as being bivariate normally distributed. The within study variances and covariances are obtained from the reports of the individual studies and are treated as known. Any heterogeneity is assumed to be normally distributed. Full details of the model are reported in the Appendix. In the examples presented the parameters were estimated by maximum likelihood using the ml command in Stata.[18] Interval estimates can be obtained either from the approximate standard errors obtained as part of the

maximization, or from the appropriate profile likelihood. The profile likelihoods were used for the bounded analysis and were obtained by considering selected values of one of the parameters and maximizing the likelihood over the others. The corresponding intervals are the range of estimates that had a profile likelihood within $1.92 = 1/2[\chi_1^2(95\%)]$ of the maximum. In the bounded analysis λ was restricted to the range 0–1, that is heterosis was excluded. To obtain intervals under these conditions the maximization required for the profile likelihoods was performed over the restricted range. Values of Akaike's Information Criterion (AIC) are reported for model comparison,[19] with the best models showing the smallest AIC.

Prior to model fitting, it may be useful to plot, for each study, the log OR_{Gg} vs log OR_{GG}, as shown in Figure 1, in which the slope of the association between logOR_{Gg} and log OR_{GG} represents λ. Such a plot may help check the consistency of λ

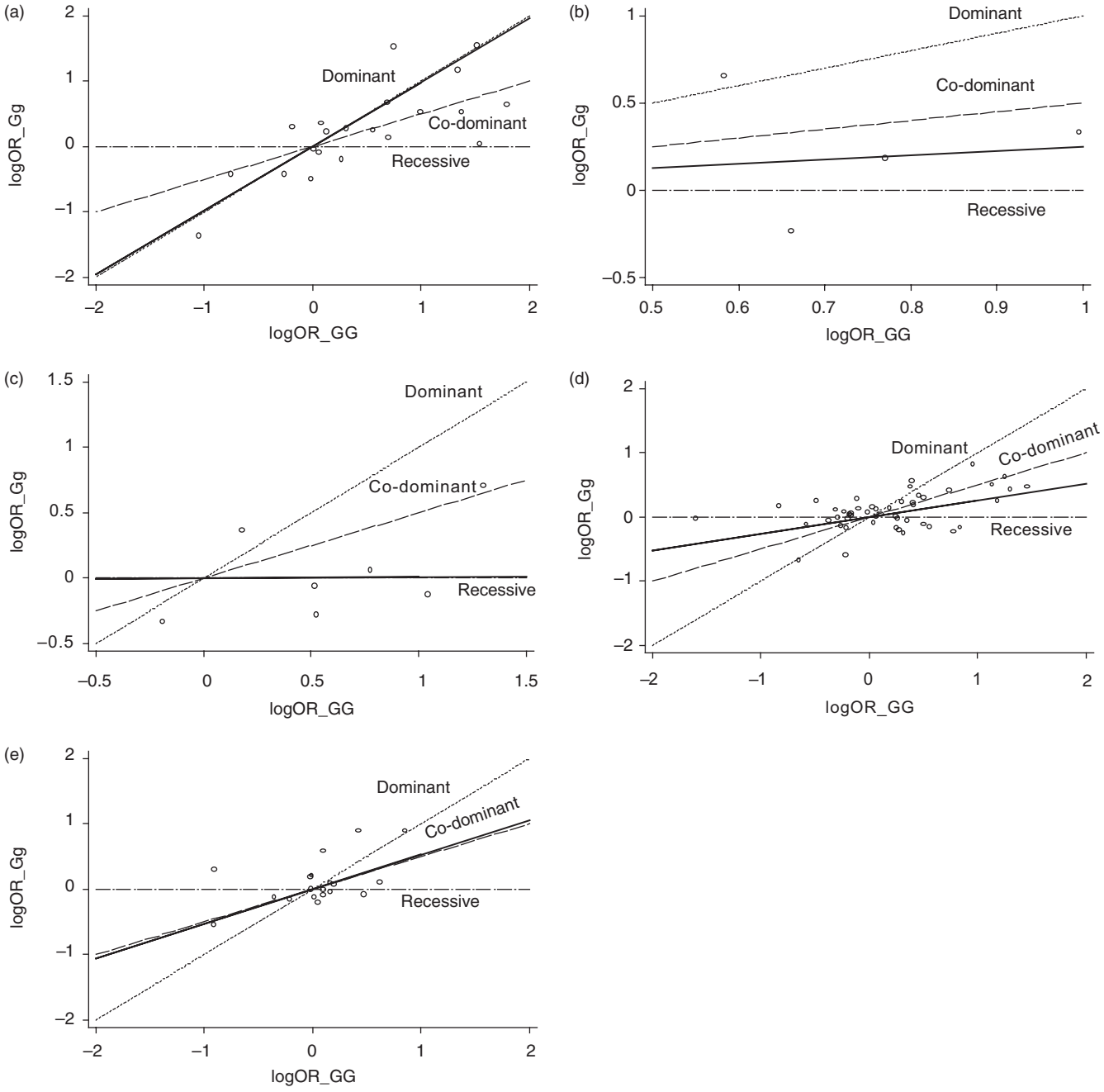

Figure 1 Plot of the log OR_{Gg} against the log OR_{GG} for: (a) *ACE* gene and diabetic nephropathy, (b) *KIR6.2* gene and Type II diabetes, (c) *AGT* gene and essential hypertension, (d) *MTHFR* gene and coronary heart disease, and (e) *PON1* Q192R polymorphism and myocardial infarction. The solid line represents the slope λ estimated by the genetic model-free approach; the three dotted lines correspond to the dominant, co-dominant, and recessive genetic models, respectively

across studies and identify outlying studies. Study-specific estimates of λ and bootstrapped 95% confidence intervals (CIs), as shown in Figure 2, help assess whether the variation in λ across studies might be explained by sampling error. Figure 2 is based on 1000 bootstrap samples from each study. If the genetic model does not seem to be consistent across studies then it may be better to perform joint pairwise comparisons using a general bivariate meta-analysis model,[14] which does not assume that λ is common but still takes into account the correlation between OR_{GG} and OR_{Gg}. Details of this model are also given in the Appendix.

Examples

The genetic model-free approach is illustrated using five published examples of the meta-analysis of genetic association studies. For each meta-analysis, the number of studies included, frequency of the risk allele, methods used by their authors, and main reported results, are given in Table 1.

ACE gene and diabetic nephropathy

This meta-analysis was carried out to evaluate the controversial association of the I/D polymorphism of the *ACE* gene with

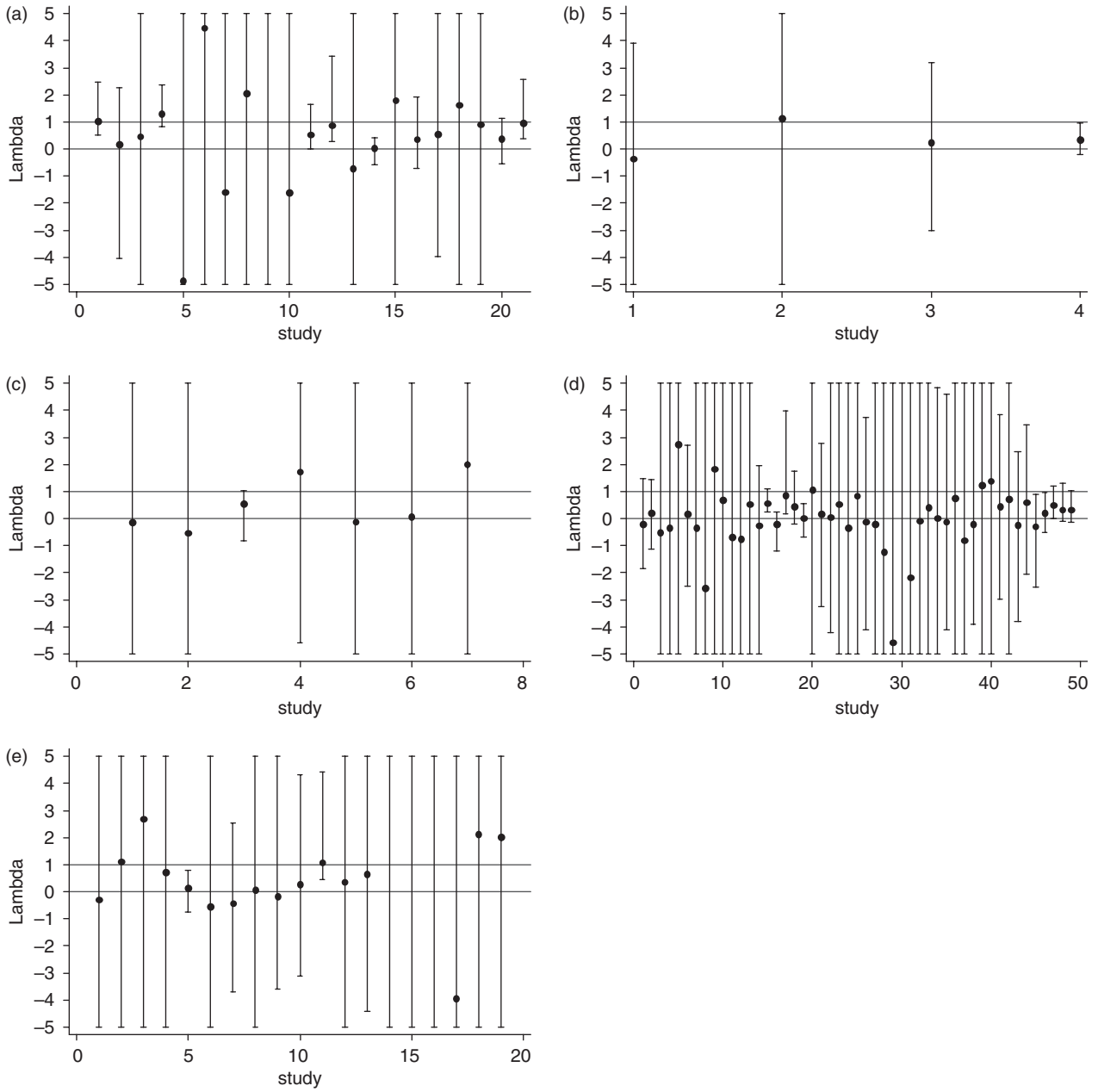

Figure 2 Plot of the study-specific estimates of λ (with 95% CI) for: (a) *ACE* gene and diabetic nephropathy, (b) *KIR6.2* gene and Type II diabetes, (c) *AGT* gene and essential hypertension, (d) *MTHFR* gene and coronary heart disease, and (e) *PON1* Q192R polymorphism and myocardial infarction. To better investigate the region in the middle, where the two lines correspond to the recessive and dominant models, the 95% CIs have been truncated at ±5

Table 1 Five published meta-analyses used for illustration, with methods and results reported in the original articles

Author, year	Association evaluated	Number of studies	Risk allele frequency	Reported analysis	
				Method	Results
Fujisawa, 1998[20]	*ACE* gene and diabetic nephropathy	21	0.46	Assumed dominant genetic model	1.32 (1.15–1.51)
Hani, 1998[21]	*KIR6.2* gene and Type II diabetes	4	0.34	Only *P*-value, under dominant and recessive genetic models	Dominant: $P < 0.05$ Recessive: $P < 0.01$
Kato, 1999[22]	*AGT* gene and essential hypertension	7	0.75	Allele frequencies cases vs controls	1.22 (1.05–1.42)
Wald, 2002[23]	*MTHFR* gene and coronary heart disease	49	0.32	Heterozygotes ignored, pairwise comparison for OR_{GG}	1.21 (1.06–1.39)
Wheeler, 2004[25]	*PON1* Q192R polymorphism and myocardial infarction	19	0.33	Per-allele relative risk	1.12 (1.15–1.51)

diabetic microangiopathy (nephropathy and retinopathy).[20] Here we consider only the meta-analysis assessing the effect on nephropathy. A dominant model was assumed and 21 studies were pooled to give an odds ratio of 1.32 (95% CI 1.15–1.51). The average allele frequency for the genetic variant was 0.46.

KIR6.2 gene and Type II diabetes

The K+ inwardly rectifier (KIR) channel is a protein that plays a major role in glucose-stimulated insulin secretion. Its encoding gene, *KIR6.2*, has been suggested as a candidate for inherited defects in Type II diabetes. This meta-analysis was carried out assuming dominant, recessive, and co-dominant models with *P*-values corrected for multiple testing.[21] The result of the meta-analysis, based on four studies, was a significant association between *KIR6.2* and Type II diabetes. The average frequency for the risk allele was 0.34.

AGT gene and essential hypertension

The genetic variant Thr235 of the angiotensinogen (*AGT*) gene has been found to be associated with hypertension in some linkage and association studies. This meta-analysis of seven Japanese case–control studies reported an odds ratio for the Thr235 allele of 1.22 (95% CI 1.05–1.42), with an average allele frequency of 0.75.[22]

MTHFR gene and coronary heart disease

The 677C→T is a polymorphism of the *MethyleneTetraHydroFolate Reductase* (*MTHFR*) gene involved in folate metabolism, which causes elevated homocysteine levels and has been associated with an increased risk of coronary heart disease. This meta-analysis of 49 studies reported an odds ratio of 1.21 (95% CI 1.06–1.39) for the TT vs CC comparison,[23] in close agreement with another meta-analysis published around the same time.[24] The average frequency for the T allele was 0.32.

PON1 Q192R polymorphism and myocardial infarction

PON1 is one of the genes encoding for paraoxonase, a serum enzyme that has been implicated in the prevention of atherogenesis and coronary heart disease through its association with high-density-lipoprotein particles. This recent meta-analysis of 19 studies investigated the effect of the Q192R polymorphism in the *PON1* gene on the risk of myocardial infarction.[25] The reported per-allele relative risk was 1.08 (95% CI 1.02–1.14), and the average allele frequency was 0.33.

Results

Figure 1 shows, for each meta-analysis, a plot of log OR_{Gg} against log OR_{GG}. All meta-analyses show variation in the genetic effect as represented by the two log odds ratios. This might be explained by a number of factors, including sampling error, differences in the study methods and differences in the true genetic risk across study populations. In the absence of heterogeneity in the genetic model and sampling error, all studies would be expected to lie along a straight line with slope λ. The solid line in Figure 1 represents the slope, λ, estimated by the genetic model-free approach, while the three dotted lines corresponding to the dominant, co-dominant, and recessive genetic models are plotted for comparison. The figure allows visual identification of any outliers or influential studies. Figure 2 plots the study-specific estimates of λ and their 95% bootstrap CIs, and is used to investigate whether any departures from linearity in Figure 1 are consistent with sampling error. Within individual studies λ is often poorly estimated, but there is little indication in any of the meta-analyses that the genetic models are not common across studies.

Table 2 summarizes the results for the different meta-analytical methods in common use; namely, separate pairwise comparisons, where log OR_{Gg} is pooled independently of log OR_{GG}, and methods based on assumed genetic models. In these analyses the log OR_{GG} has been modelled as a random effect except in two cases, marked in Table 2, where the heterogeneity of log OR_{GG} was very close to zero. The result for the ACE example when assuming a dominant model (Table 2) differs from the published result, which also assumed a dominant model (Table 1), because the main result in the original paper was based on a fixed-effect meta-analysis rather than our random effect meta-analysis.[8] The choice of the genetic model in model-based methods can have a marked impact on the estimates of OR_{GG} and OR_{Gg}. For instance, in the *KIR6.2* example, the estimates of OR_{GG} vary between 1.38 (95% CI 1.04–1.82) and 1.94 (95% CI 1.30–2.90). Separate pairwise comparisons give a consistent estimate of OR_{GG} of 2.21, but with an unnecessarily wide CI (95% CI 1.43–3.40) because they do not incorporate any of the information on OR_{Gg} when estimating OR_{GG}. Values of the AIC can be used to identify genetic models that are not consistent with the data. For instance, in the ACE example the possibility of a recessive model can be eliminated.

Table 3 presents the results of the genetic model-free approach, with λ unbounded and bounded between 0 and 1, and of the joint pairwise comparisons. The pooled estimates of λ obtained from the genetic model-free approach tend not to be very precise,

Table 2 Results of currently used meta-analytical methods for the five meta-analyses

Meta-analysis	Method	OR_{GG} (95% CI)	OR_{Gg} (95% CI)	Implicit λ	AIC
ACE gene and diabetic nephropathy	Separate pairwise comparisons	1.44 (1.07–1.93)	1.23 (0.94–1.60)	–	94.4
	Recessive model	1.16 (1.01–1.32)	–	0	89.8
	Co-dominant model	1.42 (1.09–1.87)	1.19 (1.04–1.37)	0.5	76.2
	Dominant model	1.29 (1.00–1.66)	–	1	68.2
KIR6.2 gene and Type II diabetes[a]	Separate pairwise comparisons	2.21 (1.43–3.40)	1.22 (0.91–1.64)	–	8.4
	Recessive model	1.93 (1.29–2.88)	–	0	8.0
	Co-dominant model	1.94 (1.30–2.90)	1.39 (1.14–1.70)	0.5	7.9
	Dominant model	1.38 (1.04–1.82)	–	1	13.3
AGT gene and essential hypertension	Separate pairwise comparisons	1.58 (1.06–2.35)	1.16 (0.77–1.76)	–	26.0
	Recessive model	1.64 (1.17–2.29)	–	0	20.7
	Co-dominant model	2.15 (1.26–3.65)	1.47 (1.12–1.91)	0.5	24.2
	Dominant model	1.41 (0.95–2.09)	–	1	40.6
MTHFR gene and coronary heart disease	Separate pairwise comparisons	1.19 (1.04–1.36)	1.05 (0.99–1.12)	–	88.6
	Recessive model	1.16 (1.02–1.31)	–	0	76.2
	Co-dominant model	1.18 (1.05–1.32)	1.08 (1.02–1.15)	0.5	75.8
	Dominant model	1.08 (1.01–1.16)	–	1	88.3
PON1 Q192R polymorphism and myocardial infarction[a]	Separate pairwise comparisons	1.16 (1.02–1.32)	1.08 (1.00–1.17)	–	22.4
	Recessive model	1.13 (1.00–1.27)	–	0	20.0
	Co-dominant model	1.17 (1.05–1.31)	1.08 (1.03–1.14)	0.5	15.8
	Dominant model	1.10 (1.02–1.18)	–	1	17.4

[a] Fixed-effect model.

Table 3 Results of the proposed genetic model-free approach, for both unbounded and bounded λ, and the joint pairwise comparisons obtained using bivariate meta-analysis

Meta-analysis	Method	OR_{GG} (95% CI)	OR_{Gg} (95% CI)	λ (95% CI)	AIC
ACE gene and diabetic nephropathy	Genetic model-free approach				
	Unbounded λ	1.30 (0.98–1.72)	1.29 (1.01–1.66)	0.98 (0.61–1.34)	70.2
	Bounded λ	1.30 (1.00–1.77)	1.29 (1.00–1.69)	0.98 (0.61–1.00)	70.2
	Joint pairwise comparisons	1.39 (1.07–1.81)	1.23 (0.96–1.58)	–	71.4
KIR6.2 gene and Type II diabetes[a]	Genetic model-free approach				
	Unbounded λ	2.14 (1.39–3.29)	1.21 (0.90–1.63)	0.25 (−0.11 to 0.61)	8.4
	Bounded λ	2.14 (1.43–3.29)	1.21 (1.08–1.63)	0.25 (0.00–0.69)	8.4
	Joint pairwise comparisons	2.14 (1.39–3.29)	1.21 (0.90–1.63)	–	8.4
AGT gene and essential hypertension	Genetic model-free approach				
	Unbounded λ	1.64 (0.99–2.72)	1.00 (0.66–1.53)	0.01 (−0.83 to 0.85)	22.7
	Bounded λ	1.64 (1.15–3.05)	1.00 (1.00–1.62)	0.01 (0.00–0.52)	22.7
	Joint pairwise comparisons	1.86 (1.14–3.05)	1.16 (0.77–1.76)	–	24.2
MTHFR gene and coronary heart disease	Genetic model-free approach				
	Unbounded λ	1.20 (1.05–1.37)	1.05 (0.99–1.11)	0.26 (0.04–0.47)	73.6
	Bounded λ	1.20 (1.05–1.38)	1.05 (1.01–1.12)	0.26 (0.04–0.49)	73.6
	Joint pairwise comparisons	1.20 (1.05–1.37)	1.06 (0.99–1.13)	–	75.3
PON1 Q192R polymorphism and myocardial infarction[a]	Genetic model-free approach				
	Unbounded λ	1.17 (1.04–1.33)	1.08 (1.00–1.17)	0.53 (−0.03 to 1.13)	17.8
	Bounded λ	1.17 (1.04–1.33)	1.08 (1.01–1.17)	0.53 (0.09–1.00)	17.8
	Joint pairwise comparisons	1.17 (1.04–1.33)	1.08 (1.00–1.17)	–	17.8

[a] Fixed-effect model.

but like the AIC, they can usually rule out some of the commonly assumed genetic models. For example, the *KIR6.2* gene and the *ACE* gene examples rule out the dominant and recessive models, respectively, while the *MTHFR* gene example suggests that λ is different from any of the values corresponding to the standard genetic models. In the example of the *ACE* gene, the estimate of λ is very close to 1, that is, close to dominant. Compared with an assumed dominant model, the model-free approach gives very similar estimates of OR_{GG}, but the CI is wider reflecting uncertainty about the true mode of inheritance.

In all of the examples, the AIC shows that the genetic model-free approach fits at least as well as the joint pairwise comparisons. Since the two approaches only differ for the assumption of common λ, these findings support those in Figure 2, and suggest that there is no evidence against the assumption of common λ in any of the five examples.

Under a fixed-effect assumption there is no between-study heterogeneity and so the model-free approach is exactly equivalent to the joint pairwise comparison as both models adjust for within-study correlation. For a random-effects model they give different answers because the model-free approach implies a structured covariance pattern as well as assuming a common mode of inheritance (see Appendix). The bounded analysis, in which λ must lie between 0 and 1, did not alter the point estimates of any of the parameters in our examples, because the maximum likelihood estimates of λ were all within the required range. The intervals for λ in the bounded analysis are truncated at 0 and 1 and are based on profile likelihoods rather than approximate standard errors, which accounts for some small differences from the unbounded analysis. The bounded analysis can have an effect on the interval estimates. For instance, in the *AGT* example, where the fitted model is very close to recessive, the restriction on λ implies that OR_{Gg} cannot fall <1.00 as this would either require a negative λ or a protective effect of the GG genotype; the bound rules out the former and the data contradict the latter.

The *AGT* example appears to be close to recessive, $\lambda = 0.01$, but with the largest study pointing to a co-dominant effect, as shown in Figure 2c. If the constancy of λ is doubted then joint pairwise comparisons could be used; such an analysis does not down weight the OR_{GG} and OR_{Gg} estimates from the largest study to the same extent and so produces larger pooled estimates. The AIC prefers the genetic model-free approach because it requires three parameters instead of four.

Discussion

When synthesizing the evidence on the association between a genetic polymorphism and a disease the main issue is the size of any association, but an important additional question is the mode of action of the gene. In practice, the estimate of the size of the association is influenced by our assumptions about the underlying genetic model. A review of the literature on meta-analysis of genetic association studies reveals how currently used approaches fail to address this issue.[6] Investigators often base their meta-analyses on the assumption of a specific genetic model and ignore their uncertainty about the mode of inheritance. Moreover, since it may be that no a priori biological evidence is available to justify the choice, different common genetic models are sometimes tested and the different results reported. Apart from the problem of multiple testing, this leaves the reader with a set of estimates and significance tests to interpret, all based on different assumptions. A number of investigators compare allele frequencies between cases and controls; however, this method yields a per-allele effect that is equivalent to assuming a co-dominant model with Hardy–Weinberg equilibrium. Additionally, the issue of whether the genetic model is actually common across populations does not seem to have been addressed.

The results for the five meta-analysis examples show that adopting the wrong genetic model can lead to erroneous pooled estimates with deceptively high precision. The only meta-analytical approach currently in use that does not assume a common known underlying genetic model is analysis by separate pairwise comparisons, i.e. independent meta-analyses comparing genotype groups two at a time. This method ignores the correlation between the two estimated odds ratios induced by the common baseline group and thus is inefficient, as the estimates cannot 'borrow strength' from one another as they would in a multivariate meta-analysis.[14,15] The genetic model-free approach is likely to be particularly beneficial compared with pairwise comparisons when either of the alleles is rare. Moreover, separate pairwise comparisons run into the problem of multiple testing, which becomes especially important when a polymorphism with more than two alleles is considered.

As Table 1 illustrates, published meta-analysis of genetic association studies have used a variety of methods for presenting their results. The genetic model-free approach offers a single method that could have been used in all of these examples giving a consistent presentation and avoiding the pitfall of overly strong assumptions about the genetic model or of inefficient estimates.

The genetic model-free approach provides an integrated way of synthesizing the evidence on genetic associations, which yields not only the magnitude of the genetic effect (OR), but also an indication of the operating genetic model based on the available data. The underlying genetic model is not constrained to correspond to one of the classical modes of inheritance (recessive, co-dominant, dominant), in recognition of the fact that the gene's mode of action in complex diseases might differ from that found in Mendelian traits, where the association between genotype and disease tend to be of a deterministic nature and, hence, the mode of inheritance is relatively clearly apparent. For example, a value of 0.26 for λ, as in the *MTHFR* meta-analysis, might be interpreted in two ways:

(i) The polymorphism is recessive in some studies and co-dominant in others, so that the average result is between the two.
(ii) In complex diseases, the genotype is only one of many factors acting in a complex causal cascade leading to the disease. Although, at the molecular level, the polymorphism of interest might act in a clearly Mendelian manner on some intermediate phenotype, that Mendelian 'signal' may be 'diluted' or 'distorted' when measured at the level of the final step in the cascade. Hence, λ may be a more flexible and appropriate way to discuss genetic models in complex disease.

In the meta-analysis of genetic association studies there are two important types of heterogeneity that need to be addressed: heterogeneity in the genetic effect and heterogeneity in the genetic model. There are a number of reasons why we might see heterogeneity in the genetic effect, including differences in study

methods and differences in the underlying genetic risk associated with gene–gene or gene–environment interactions. Heterogeneity of the genetic effect might also arise if the polymorphism under study does not act directly on the disease risk, i.e. it is not a 'functional' or 'causal' polymorphism but is simply a marker, which tends to be inherited together with the causal polymorphism (linkage disequilibrium). Populations may have different patterns of linkage disequilibrium, which lead to differences in the marker association with disease. It is important to note that causes of heterogeneity in the genetic effect will not necessarily cause heterogeneity in the genetic model. In fact, in order to act on the genetic model, interactions need to influence the disease risk in heterozygotes to a different extent to the risk in homozygotes.

The absence of heterogeneity in the genetic model is an important assumption of the genetic model-free analysis and, although this assumption is likely to hold in most cases, it still needs to be assessed. For example, the effect of genotype on allergy to pollens appears to follow different modes of inheritance for different ethnic groups and different forms of allergy.[26,27] Although these studies are based on segregation analyses, and are relatively weak, they do raise the possibility that the mode of action may vary from study to study, perhaps owing to complex gene–environment interactions that have different impact on the disease risk in heterozygotes compared with homozygotes for the polymorphism. Thus, the assumption of a common genetic model should be checked before applying the genetic model-free approach, for instance by using the graphs presented in Figures 1 and 2. Should this assumption be in doubt, then the best approach would be to carry out joint pairwise comparisons using a multivariate meta-analysis, where the correlations between the odds ratios for the different genotype groups are taken into account. In addition to the graphical investigation, the difference in fit, as measured by AIC, between the model with common λ and the corresponding pairwise analysis offers a guide to the appropriateness of the assumption of a common genetic model. In general the random-effects model-free approach is easier to fit than the corresponding pairwise bivariate model because it contains two fewer parameters. Only in very large meta-analyses will it be possible to estimate the correlation in the heterogeneities required for the pairwise model. So, even when the assumptions of the model-free analysis are not met exactly, the model-free analysis may still be the best way of summarizing the data and obtaining CIs that are not falsely optimistic.

All of the models considered in this paper have been based on the normal approximation to the distribution of the log odds ratio. In examples where some of the studies have very few subjects within one of the genotypes, as might happen with a rarer allele, it would be better to use a multinomial likelihood. In the case of a random-effects model this adds to the complexity because of the need to numerically integrate over the random effect before maximization. Within this multinomial framework we can still use the λ parameterization basic to the genetic model-free approach and interpret the results in the same way as with the normal approximation.

The results presented in this paper have been obtained using maximum likelihood methods, but a Bayesian approach with non-informative prior distributions gave very similar results to those in Tables 2 and 3 (data not shown). The choice of a Bayesian approach to implement the method might be more desirable when there is external information regarding the magnitude of the genetic effect and/or mode of inheritance, which might come from studies not included in the meta-analysis or from expert opinion.[28] When Markov chain Monte Carlo methods are used, it also makes the generalization to multinomial likelihoods with random effects more straightforward.[29]

In conclusion, we propose a new meta-analytical method based on a re-parameterization of the classical representation of genetic association studies, where the new parameters are biologically meaningful and informative. The approach makes maximum use of the information available by quantifying the magnitude of the genetic effect and estimating the genetic mode of action at the same time. The genetic model is estimated on the basis of the data rather than assumed, and this is important in all cases where no a priori knowledge about the underlying genetic model is available.

Acknowledgements

We would like to acknowledge the helpful comments on earlier drafts of the paper from Martin Tobin. Cosetta Minelli would like to thank the Department of Health, UK, for supporting this research via a National Research Scientist in Evidence Synthesis Award. We would also like to thank the two anonymous reviewers for their thoughtful and useful comments.

KEY MESSAGES

- Meta-analysis of molecular association studies is often based on the assumption of a specific genetic model (recessive, co-dominant, or dominant).
- Biological justification for the choice of the genetic model is seldom available, and results can be misleading when an inappropriate model is assumed.
- Specification of the genetic model is sometimes avoided by comparing genotype groups two at a time, but this is inefficient.
- We propose a genetic model-free approach where the information available on all genotypes is used and the genetic model is estimated rather than assumed.
- The approach assumes that all studies share the same unknown genetic model, and we suggest ways of investigating whether this assumption might hold.

References

[1] Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. *Genet Med* 2002;**4:**45–61.

[2] Clayton D, McKeigue PM. Epidemiological methods for studying genes and environmental factors in complex diseases. *Lancet* 2001;**358:**1356–60.

[3] Colhoun HM, McKeigue PM, Davey Smith G. Problems of reporting genetic associations with complex outcomes. *Lancet* 2003;**361:**865–72.

[4] Cardon LR, Bell JI. Association study designs for complex diseases. *Nat Rev Genet* 2001;**2:**91–99.

[5] Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet* 2003;**33:**177–82.

[6] Attia J, Thakkinstian A, D'Este C. Meta-analyses of molecular association studies: Methodologic lessons for genetic epidemiology. *J Clin Epidemiol* 2003;**56:**297–303.

[7] Salanti G, Sanderson S, Higgins JP. Obstacles and opportunities in meta-analysis of genetic association studies. *Genet Med* 2005;**7:**13–20.

[8] Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. *Methods for Meta-Analysis in Medical Research*. Chichester: Wiley, 2000.

[9] Sterne JA, Egger M, Smith GD. Systematic reviews in health care: Investigating and dealing with publication and other biases in meta-analysis. *BMJ* 2001;**323:**101–5.

[10] The Lancet. In search of genetic precision. *Lancet* 2003;**361:**357.

[11] Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. Replication validity of genetic association studies. *Nat Genet* 2001;**29:**306–9.

[12] Morgan TM, Coffey CS, Krumholz HM. Overestimation of genetic risks owing to small sample sizes in cardiovascular studies. *Clin Genet* 2003;**64:**7–17.

[13] Agema WR, Jukema JW, Zwinderman AH, van der Wall EE. A meta-analysis of the angiotensin-converting enzyme gene polymorphism and restenosis after percutaneous transluminal coronary revascularization: evidence for publication bias. *Am Heart J* 2002;**144:**760–68.

[14] Nam IS, Mengersen K, Garthwaite P. Multivariate meta-analysis. *Stat Med* 2003;**22:**2309–33.

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

[16] Williams J, Spurlock G, Holmans P *et al.* A meta-analysis and transmission disequilibrium study of association between the dopamine D3 receptor gene and schizophrenia [*erratum in Mol Psychiatr 1998;3:458*] *Mol Psychiatr* 1998;**3:**141–49.

[17] Juengel JL, Quirke LD, Tisdall DJ, Smith P, Hudson NL, McNatty KP. Gene expression in abnormal ovarian structures of ewes homozygous for the inverdale prolificacy gene. *Biol Reprod* 2000;**62:**1467–78.

[18] Gould W, Sribney W. *Maximum Likelihood Estimation with Stata*. College Station, Texas: Stata Press, 1999.

[19] Akaike H. Fitting autoregressive models for prediction. *Ann Inst Stat Math* 1969;**21:**243–47.

[20] Fujisawa T, Ikegami H, Kawaguchi Y *et al*. Meta-analysis of association of insertion/deletion polymorphism of angiotensin I-converting enzyme gene with diabetic nephropathy and retinopathy. *Diabetologia* 1998;**1:**47–53.

[21] Hani EH, Boutin P, Durand E *et al*. 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* 1998;**41:**1511–15.

[22] Kato N, Sugiyama T, Morita H, Kurihara H, Yamori Y, Yazaki Y. Angiotensinogen gene and essential hypertension in the Japanese: extensive association study and meta-analysis on six reported studies. *J Hypertens* 1999;**17:**757–63.

[23] Wald DS, Law M, Morris JK. Homocysteine and cardiovascular disease: evidence on causality from a meta-analysis. *BMJ* 2002;**325:**1202.

[24] Klerk M, Verhoef P, Clarke R, Blom HJ, Kok FJ, Schouten EG. MTHFR 677C→T polymorphism and risk of coronary heart disease: a meta-analysis. *JAMA* 2002;**288:**2023–31.

[25] Wheeler JG, Keavney BD, Watkins H, Collins R, Danesh J. Four paraoxonase gene polymorphisms in 11212 cases of coronary heart disease and 12786 controls: meta-analysis of 43 studies. *Lancet* 2004;**363:**689–95.

[26] Marsh DG, Huang SK. Molecular genetics of human immune responsiveness to pollen allergens. *Clin Exp Allergy* 1991;**21(suppl.1):**168–72.

[27] Sasazuki T, Nishimura Y, Muto M, Ohta N. HLA-linked genes controlling immune response and disease susceptibility. *Immunol Rev* 1983;**70:**51–75.

[28] Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation.* Chichester: Wiley, 2004.

[29] Minelli C, Thompson JR, Abrams KR, Lambert PC. Bayesian implementation of a genetic model-free approach to the meta-analysis of genetic association studies. *Stat Med* (in press).

Appendix

Bivariate meta-analysis

Consider the meta-analysis of a set of case–control association studies of a bi-allelic polymorphism. Let z_{1i} represent the value of log OR_{Gg} estimated from the ith study and z_{2i} the log OR_{GG}. Assuming approximate bivariate normality

$$\begin{bmatrix} z_{1i} \\ z_{2i} \end{bmatrix} \sim N\left\{ \begin{bmatrix} \mu_{1i} \\ \mu_{2i} \end{bmatrix}, \begin{bmatrix} v_{1i} & v_{12i} \\ v_{12i} & v_{2i} \end{bmatrix} \right\},$$

where the μ_is are the true log odds ratios for that study. The values of the variances and covariances are treated as known and can be derived from the number of cases and controls in each genotype in that study. If we assume that the studies come from a population in which the log odds ratios are also normally distributed, then

$$\begin{bmatrix} \mu_{i1} \\ \mu_{i2} \end{bmatrix} \sim N\left\{ \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \tau_1 & \tau_{12} \\ \tau_{12} & \tau_2 \end{bmatrix} \right\},$$

where the τ's represent the heterogeneities between studies. The distribution of the observed data in the meta-analysis is thus

$$\begin{bmatrix} z_{i1} \\ z_{i2} \end{bmatrix} \sim N\left\{ \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} v_{1i}+\tau_1 & v_{12i}+\tau_{12} \\ v_{12i}+\tau_{12} & v_{2i}+\tau_2 \end{bmatrix} \right\}.$$

From which a likelihood can be formed and the parameters estimated. Unless the meta-analysis includes a large number of studies, the covariance between the heterogeneities is difficult to estimate, but the results for the other parameters are not very sensitive to τ_{12} so using an assumed value will not be misleading. In our analyses we used $\tau_{12} = 0.9\sqrt{\tau_1\tau_2}$ and checked the results in a sensitivity analysis. A fixed-effects model assumes that $\tau_1 = \tau_2 = \tau_{12} = 0$.

Genetic model-free analysis

The genetic model-free analysis is similar to the general bivariate meta-analysis. First we assume that

$$\begin{bmatrix} z_{1i} \\ z_{2i} \end{bmatrix} \sim N\left\{ \begin{bmatrix} \lambda\mu_{2i} \\ \mu_{2i} \end{bmatrix}, \begin{bmatrix} v_{1i} & v_{12i} \\ v_{12i} & v_{2i} \end{bmatrix} \right\},$$

where the parameter, λ, which describes the genetic model is common across studies. The heterogeneity between studies will be

$$\begin{bmatrix} \lambda\mu_{2i} \\ \mu_{2i} \end{bmatrix} \sim N\left\{ \begin{bmatrix} \lambda\mu_2 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \lambda^2\tau & \lambda\tau \\ \lambda\tau & \tau \end{bmatrix} \right\}.$$

The distribution of the observed data in the meta-analysis is thus

$$\begin{bmatrix} z_{1i} \\ z_{2i} \end{bmatrix} \sim N\left\{ \begin{bmatrix} \lambda\mu_2 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} v_{1i} + \lambda^2\tau & v_{12i} + \lambda\tau \\ v_{12i} + \lambda\tau & v_{2i} + \tau \end{bmatrix} \right\},$$

and once again the likelihood can be formed and maximized to estimate the parameters. In this model the covariance between the heterogeneities is controlled by λ and can thus be estimated. It is advisable to inspect the profile likelihood of each parameter as in small meta-analyses the log-likelihood can be far from quadratic. A fixed-effects model assumes that $\tau = 0$.

In some meta-analyses it may be appropriate to restrict λ to lie in the range (0,1), that is, to exclude heterosis. In this case, the overall maximization and the profile likelihood maximizations are over the restricted range.