STATISTICS IN MEDICINE
Statist. Med. 2004; **23**:1663–1682 (DOI: 10.1002/sim.1752)

Controlling the risk of spurious findings from meta-regression

Julian P. T. Higgins*,† and Simon G. Thompson

MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR, U.K.

SUMMARY

Meta-regression has become a commonly used tool for investigating whether study characteristics may explain heterogeneity of results among studies in a systematic review. However, such explorations of heterogeneity are prone to misleading false-positive results. It is unclear how many covariates can reliably be investigated, and how this might depend on the number of studies, the extent of the heterogeneity and the relative weights awarded to the different studies. Our objectives in this paper are two-fold. First, we use simulation to investigate the type I error rate of meta-regression in various situations. Second, we propose a permutation test approach for assessing the true statistical significance of an observed meta-regression finding. Standard meta-regression methods suffer from substantially inflated false-positive rates when heterogeneity is present, when there are few studies and when there are many covariates. These are typical of situations in which meta-regressions are routinely employed. We demonstrate in particular that fixed effect meta-regression is likely to produce seriously misleading results in the presence of heterogeneity. The permutation test appropriately tempers the statistical significance of meta-regression findings. We recommend its use before a statistically significant relationship is claimed from a standard meta-regression analysis. Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS: meta-analysis; meta-regression; weighted regression; false-positive results; permutation tests; randomization tests; simulation study

1. INTRODUCTION

Meta-regression aims to investigate whether heterogeneity among results of multiple studies is related to specific characteristics of the studies. For example, meta-regression has been used to relate vaccine efficacy to geographical latitude [1], coronary risk benefit to serum cholesterol reduction [2] and properties of diagnostic tests to methodological quality of diagnostic accuracy studies [3]. While meta-regression has become a commonly used tool for exploring heterogeneity among the results of studies in a systematic review, explorations of heterogeneity are widely noted to be potentially misleading [4–7]. Investigations of differences between studies and their results are observational associations and are subject to biases (such

*Correspondence to: Julian P. T. Higgins, MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge, CB2 2SR, U.K.

†E-mail: julian.higgins@mrc-bsu.cam.ac.uk

Contract/grant sponsor: MRC; contract/grant number: G9815466

Received July 2002
Accepted September 2003

as aggregation bias) and confounding (resulting from correlation between characteristics) [7]. Moreover, many systematic reviews include small numbers of studies, whereas innumerable characteristics of these studies may be identified as potential causes of heterogeneity. There is a clear danger of misleading conclusions if p-values from multiple meta-regression analyses are interpreted naïvely.

False-positive results are more likely in meta-regression than in conventional regression because of the potential presence of heterogeneity. For example, consider the case of just two studies producing estimates with non-overlapping confidence intervals. Any covariate whose value differs between these studies will be significantly related to the heterogeneity among the studies, and hence a potential explanation of it. It is clear, however, that the majority of such ‘explanations’ will be entirely spurious. In the case of three studies yielding non-overlapping confidence intervals, one may expect a third of unimportant covariates to appear to explain the heterogeneity. As the number of studies increases, the risk of identifying spurious associations decreases. It is unclear at what point the risk becomes acceptably small, and how this might depend on the number of studies, the number of covariates, the extent of the heterogeneity and the relative weights awarded to the different studies. The question is difficult to address analytically because of the complexities of heterogeneity and differential study weights, and results for conventional regression [8, 9] are not directly relevant.

Our objectives in this paper are two-fold. First, we use Monte Carlo simulation to investigate the type I error rate of frequentist meta-regression in various situations. Second, to address important limitations in currently available methodology, we propose a permutation test approach for assessing the true statistical significance of an observed meta-regression finding. In developing the permutation test we have in mind the availability of, typically, a small number of studies, with one or a moderate number of potential covariates. Although exploratory meta-regression analyses may sometimes throw up useful leads for further investigation, we aim here to maintain type I error so that spurious findings are strictly controlled.

We start with a brief review of meta-regression methodology in Section 2. In Section 3, we compare the type I error rates of different methods in simulations, and in Section 4, we describe the permutation test. We illustrate our methods using two published examples in Section 5, and discuss our findings and alternative approaches to the problem in Section 6.

2. A BRIEF REVIEW OF META-REGRESSION METHODOLOGY

2.1. Methods

A variety of frequentist methods for meta-regression has been described [10–12]. Here we provide a brief overview of approaches that we address in this paper. Suppose an effect size θ_i is estimated by y_i in study i ($i=1,\ldots,k$). These may for example be treatment effects from clinical trials, measures of association from observational studies or (log) diagnostic odds ratios from studies of diagnostic accuracy. We suppose the estimated variance of y_i is v_i, and we follow the convention of assuming this is known. Let $\mathbf{X}$ be a $k \times (m+1)$ matrix of m study-level covariates plus an intercept term, and $\boldsymbol{\beta}$ be an $(m+1)\times 1$ vector of coefficients. Early investigations of heterogeneity in meta-analyses made use of a fixed effect meta-regression model, in which the effects in different studies were assumed to be fully explained by regression on the study-level covariates [13]. Using $\mathbf{X}_i$ to denote the ith row of

X, the fixed effect meta-regression model is

$$y_i \sim \mathrm{N}(\theta_i, v_i)$$

$$\theta_i = \mathbf{X}_i\boldsymbol{\beta}$$

Such analyses are possible using standard weighted regression with weights $w_i = 1/v_i$. However, in weighted regression the *relative* weights are assumed known whereas in meta-regression the *actual* weights are assumed known. The correct standard errors for the regression slopes must therefore be obtained by dividing the usual standard errors by the mean square error [12, 13].

Random effects meta-regression allows for residual heterogeneity, usually by assuming the underlying effects follow a normal distribution around the effects predicted by the covariates [14]:

$$y_i \sim \mathrm{N}(\theta_i, v_i)$$

$$\theta_i \sim \mathrm{N}(\mathbf{X}_i\boldsymbol{\beta}, \tau^2)$$

The heterogeneity variance parameter, τ^2, can be estimated in several ways. A non-iterative estimate may be obtained through a method of moments approach, and iterative estimates include a restricted maximum likelihood (REML) estimate, a maximum likelihood estimate and an empirical Bayes estimate. The REML estimate has been advocated [12], and is the default method in the command 'metareg' that has been written for STATA [15], although Berkey *et al.* prefer the empirical Bayes method [14].

A statistical test for evidence of effect of a particular covariate is available via a Wald test using the statistic

$$T = \frac{\hat{\boldsymbol{\beta}}_j}{\mathrm{SE}(\hat{\boldsymbol{\beta}}_j)} \tag{1}$$

An unresolved issue is the distribution of T under the null hypothesis in random effects meta-regression. It is commonly referred to the standard normal distribution, as in fixed effect meta-regression, although this ignores the fact that the heterogeneity variance, τ^2, has been estimated. Berkey *et al.* [14] experimented with t-distributions with various degrees of freedom. For their particular example they concluded that $k-m-4$ degrees of freedom gave the most appropriate coverage.

More recently, Knapp and Hartung have shown the superior properties of alternative estimates of variance from random effects meta-regression [16]. They multiply the usual standard errors $\mathrm{SE}(\hat{\boldsymbol{\beta}}_j)$ by $\max\{1, \sqrt{q}\}$ where

$$q = \frac{1}{k-m-1}\sum \frac{(y_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})^2}{v_i^*}$$

and $v_i^* = v_i + \hat{\tau}^2$ is the marginal variance of y_i. The modified Wald statistics are compared with a t-distribution with $k-m-1$ degrees of freedom.

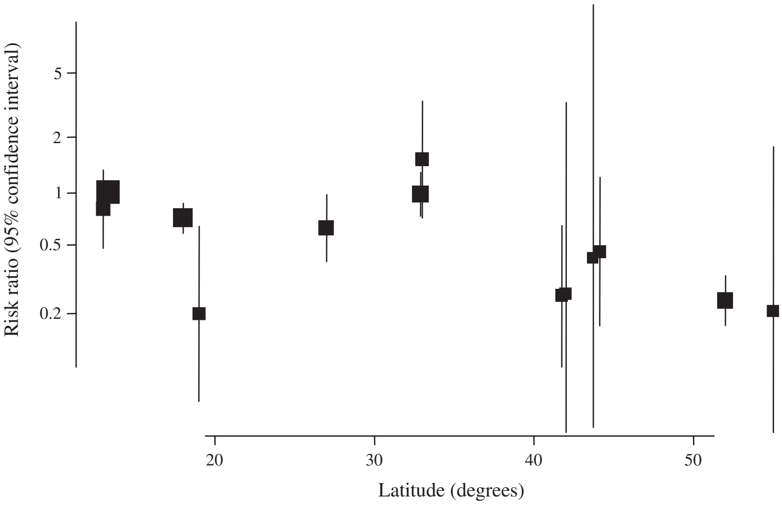


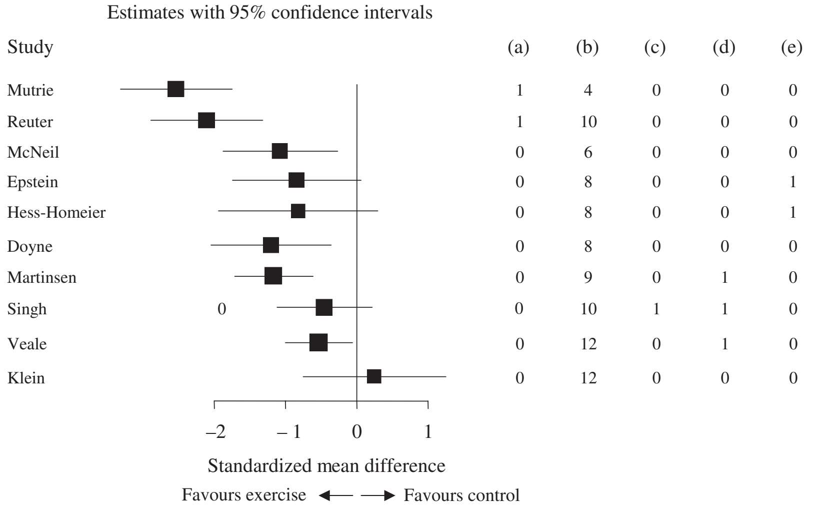
Figure 1. Results from 13 trials of BCG vaccines for preventing tuberculosis [14], plotted against absolute latitude in degrees.

When several covariates are investigated, some would advocate an adjustment for multiple comparisons to control false-positive findings, although opinions differ as to whether such techniques are appropriate [17, 18]. A Bonferroni adjustment to the nominal significance level might be applied but this relies on knowledge of the true type I error rate for a single covariate. Furthermore, colinearity among covariates in meta-regression is common [4], and a standard Bonferroni approach may be inappropriately conservative.

2.2. *Examples*

We introduce two examples of published meta-regression analyses, and return to them later in the paper. The first is that with which Berkey and colleagues introduced methodology for random effects meta-regression analysis, being a collection of 13 trials of BCG vaccines [1, 14] (Figure 1). The outcome is mortality, with vaccine efficacy measured as the ratio of mortality risks in the treated and control groups. The covariate of interest is (absolute) geographical latitude of the study, a proxy for several potentially important variables including climate and environmental mycobacteria. Berkey *et al.* obtain an estimate of -0.0268 (SE 0.011) for the regression coefficient of log risk ratio on absolute latitude in degrees using a random effects approach with empirical Bayes estimate of τ^2, and -0.0282 (SE 0.004) using a fixed effect approach.

Our second example is from a systematic review of the effectiveness of exercise as an intervention in the management of depression [19] (Figure 2). A standardized mean difference was used to compare results across different depression scales. The authors listed eight potential sources of heterogeneity for investigation using random effects meta-regression based on 10 studies. Two of these covariates (publication in conference abstracts rather than journals or theses, and length of follow-up) were found to be statistically significantly related to the magnitude of benefit from exercise.



Study	(a)	(b)	(c)	(d)	(e)
Mutrie	1	4	0	0	0
Reuter	1	10	0	0	0
McNeil	0	6	0	0	0
Epstein	0	8	0	0	1
Hess-Homeier	0	8	0	0	1
Doyne	0	8	0	0	0
Martinsen	0	9	0	1	0
Singh	0	10	1	1	0
Veale	0	12	0	1	0
Klein	0	12	0	0	0

Figure 2. Results from 10 trials of exercise in the management of depression [19]. Covariates are as follows: (a) published as abstract; (b) length of follow-up; (c) intention-to-treat analysis; (d) randomized allocation adequately concealed; and (e) published as PhD thesis. Dichotomous covariates are coded as 1 = yes, 0 = no.

3. SIMULATION STUDY

Previous studies by Berkey *et al.* [14] and Knapp and Hartung [16] have assessed properties of random effects meta-regression in simulations based specifically on the BCG data. Properties of fixed effect methods for identifying causes of heterogeneity commonly used in the social sciences have also been evaluated [20, 21]. However, neither the type I error rates in the presence of heterogeneity nor the general properties of random effects methods have received attention.

3.1. *Objectives*

Here we aim to evaluate, through Monte Carlo simulation under the null hypothesis of no true association, some properties of methods for meta-regression. In particular, we address the following questions:

1. How is the chance of a false-positive regression coefficient in meta-regression influenced by (i) the true extent of heterogeneity, (ii) the number of studies, (iii) the relative weights awarded to studies, (iv) the number of covariates investigated and (v) the extent of correlation between covariates?
2. How is the chance of a false-positive regression coefficient in meta-regression influenced by the method used?

3.2. Methods

For the purposes of this paper we compare the following methods:

(F) fixed effect meta-regression (equivalent to weighted regression with adjusted standard errors), relating the statistic (1) to the standard normal distribution,

(R) random effects meta-regression, with REML estimate of heterogeneity variance, relating the statistic (1) to the standard normal distribution, as in common practice,

(R_{KH}) random effects meta-regression, with REML estimate of heterogeneity variance, relating the modification of Knapp and Hartung to the t-distribution with $k-m-1$ degrees of freedom.

We simulate data under the null hypothesis of no association between effect estimates and any covariate, yet with an unexplained component of heterogeneity according to the standard random effects meta-analysis model. Without loss of generality we assign the average effect to 0:

$$\theta_i \sim \mathrm{N}(0,\tau^2)$$

$$y_i \sim \mathrm{N}(\theta_i, v_i) \quad \text{for } i = 1,\ldots,k$$

Covariates are simulated from a multivariate (standard) normal distribution so that correlation ρ may be imposed between pairs of covariates:

$$\mathbf{X} = (\mathbf{1}\ \tilde{\mathbf{X}})$$

$$\tilde{\mathbf{X}} \sim \mathrm{MVN}\left(\mathbf{0}, \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \ddots & \vdots \\ \vdots & \ddots & 1 & \rho \\ \rho & \cdots & \rho & 1 \end{pmatrix}\right)$$

Parameters that may be varied across simulation runs are the number of studies (k), the number of covariates (m), the correlation between covariates (ρ), the extent of genuine heterogeneity (τ^2) and the set of weights ($w_i = 1/v_i$). We investigate $k = 5, 10$ and 100 studies; $m = 1, 3$ and 5 covariates; correlations $\rho = 0, 0.5$ and 0.9 (when $m = 3$ or 5); and heterogeneity variances $\tau^2 = 0, 1$ and 5. Taking $\tau^2 = 0$ corresponds to a fixed effect model.

For the within-study variances, $\{v_i\}$, we select three patterns and calculate numerical values in relation to the heterogeneity variance. The patterns are:

- *Equal within-study variances.*
- *Variable within-study variances*: To create variances reflecting a plausible real-life situation we considered five randomized trials with sample sizes 20, 40, 80, 160 and 320,

respectively, with odds ratio 0.7 and control group event rate 0.3. For $k=5$, we assign variances proportional to those yielded by these hypothetical trials. The variances, based on the sum of reciprocals of rounded expected cell counts in a 2×2 table, are then proportional to the set $\{1.10, 0.50, 0.26, 0.13, 0.06\}$. For $k=10$ and $k=100$, we replicate the variances as required.

- *Unbalanced within-study variances*: We assign these so that four-fifths of the trials have identical within-study variances and the remaining fifth are 'mega-trials' with $\frac{1}{20}$ of the variance.

We determine numerical values for the variances by making use of the 'typical' within-study variance suggested by Higgins and Thompson [22]:

$$s^2 = \frac{\sum w_i(k-1)}{(\sum w_i)^2 - \sum w_i^2},$$

where $w_i = 1/v_i$. We fix $s^2=1$ in all cases so that the 'average' size of a study is similar across all simulated meta-analyses. This also allows us to quantify the heterogeneity using a statistic, $I^2 = \tau^2/(\tau^2+s^2)$, that describes the proportion of total variation in effect estimates attributable to heterogeneity rather than within-study variability [22]. For $\tau^2=0$, we obtain $I^2=0$ per cent; for $\tau^2=1$ we obtain $I^2=50$ per cent; for $\tau^2=5$, we obtain $I^2=83$ per cent.

For each of the 189 unique combinations of parameter values we repeat the following, within S-plus, 1000 times.

- Generate data as described above;
- apply fixed and random effects methods of meta-regression, collecting the coefficients, standard errors and the adjusted standard errors proposed by Knapp and Hartung;
- perform the statistical tests (F, R, R_{KH}), based on all regression coefficients (not including the intercept) and their standard error using a nominal 5 per cent significance level;
- collect the results of the significance tests.

From these we obtain the observed false-positive rate over 1000 simulations for each covariate using each method. (A 5 per cent significance level based on 1000 simulations has a standard error of 0.7 per cent.) In addition, we collect a composite false-positive rate for the 'most significant' covariate being statistically significant.

3.3. Results

We concentrate on three particular aspects of our findings: (i) full results for false-positive rates from meta-regression with a single covariate; (ii) the effect of increasing numbers of covariates; (iii) the effect of correlation between covariates.

(i) *A single covariate*: Figure 3 illustrates false-positive rates for meta-regressions with a single covariate. We note first that the fixed effect method, while achieving its correct 5 per cent significance level in the absence of heterogeneity, gives unacceptably high false-positive rates in the presence of heterogeneity, irrespective of the number of studies or pattern of the variances. The observed false-positive rate is typically 20 per cent for a moderate amount of heterogeneity ($I^2=50$ per cent), and can exceed 50 per cent for extreme heterogeneity ($I^2=83$ per cent). We do not show further results for the fixed effect method.

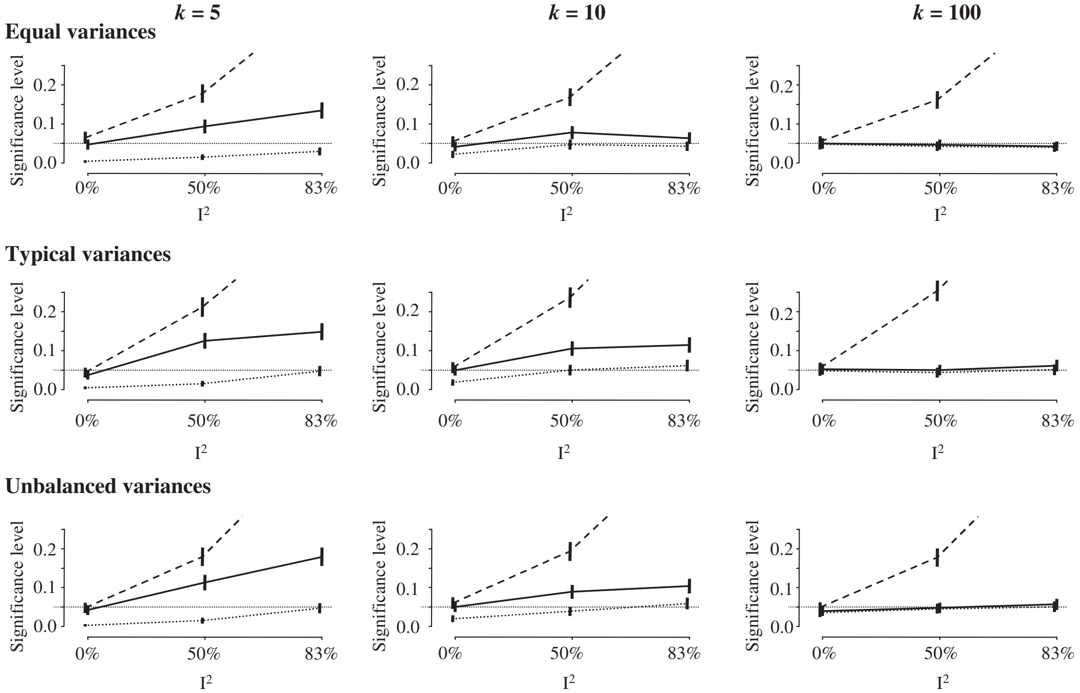


Figure 3. False-positive rates (with 95 per cent confidence intervals) for fixed effect and random effects meta-regression analysis with a single covariate. Based on 1000 simulations and REML estimation used in the random effects analyses. The nominal 5 per cent significance level is shown as a horizontal dotted line. Dashes $= F$, solid line $= R$, dots $= R_{\mathrm{KH}}$.

For the random effects methods, in the absence of any heterogeneity only the standard normal test (R) achieves the correct 5 per cent level. In the presence of heterogeneity, the rate of false-positive findings using the standard normal distribution increases substantially above 5 per cent for small numbers of studies. The t-test of and Knapp and Hartung (R_{KH}) gives false-positive rates below 5 per cent, although achieves the desired level approximately when there is substantial heterogeneity.

When there are many studies ($k = 100$) both random effects methods approximately achieve their nominal significance level. False-positive rates are similar for the three patterns of within-study variances (study weights). Although we do not present results, we also found other random effects estimation methods (e.g. method of moments, empirical Bayes) gave very similar findings to the REML approach.

(ii) *Effect of numbers of covariates*: When looking at multiple covariates we examine the rates of at least one covariate being statistically significant (at a nominal 5 per cent level). This is equivalent to assessing the statistical significance of the 'most significant' covariate. If several uncorrelated covariates are examined in separate single-variable meta-regressions then, as expected, the results described above are compounded. For example, a 5 per cent false-positive rate for a single covariate yields a $(1 - 0.95^3) = 14.3$ per cent false-positive rate when three covariates are examined and a 22.6 per cent false-positive rate when five are examined.

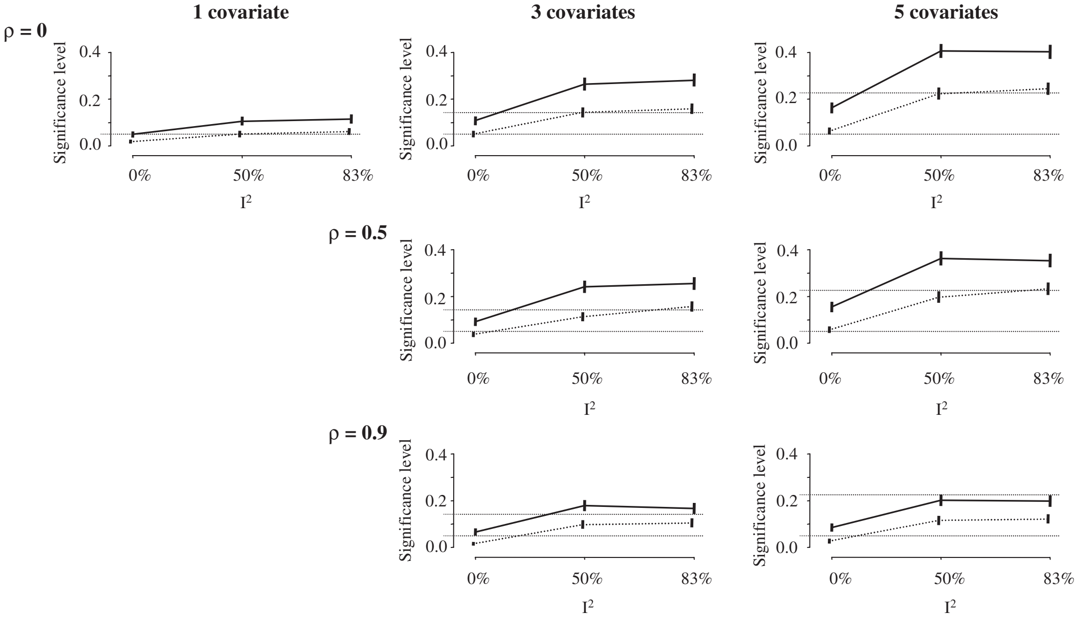


Figure 4. False-positive rates (with 95 per cent confidence intervals) for random effects meta-regression analysis examining the effects of increasing number of covariates ($m = 1, 3, 5$) and increasing correlation between covariates ($\rho = 0, 0.5, 0.9$). All analyses include 10 studies with a 'typical' pattern of within-study variances (see text). Based on 1000 simulations and REML estimation. The nominal 5 per cent significance level and expected level from independent tests are shown as horizontal dotted lines. The graph in the top left is the graph in the centre of Figure 3. Solid line $= R$, dots $= R_{\mathrm{KH}}$.

The top row of Figure 4 illustrates this for random effects (REML) analyses of $k = 10$ studies with a 'typical' variance pattern. Similar results were obtained for other variance patterns and numbers of studies.

(iii) *Effect of positive correlation between covariates*: Consider first the use of separate single covariate meta-regression analyses. Figure 4 (looking vertically) illustrates the effect of increasing correlation between each pair of covariates ($m = 3$ or 5) to $\rho = 0.5$ and 0.9. False-positive rates decrease as correlation increases. This is to be expected. For instance, consider a correlation of 1 when all meta-regressions are identical and so the true false-positive rate should be 5 per cent. However, when the correlation is as high as $\rho = 0.9$, the 5 per cent nominal level is exceeded by all methods in the presence of moderate or extreme heterogeneity ($I^2 = 50$ or 83 per cent).

In contrast, we found false-positive rates from multiple meta-regression analyses to be largely unaffected by correlation between covariates. This arises because multiple regression evaluates the 'independent' or 'partial' effects of multiple covariates. This is the only factor we found to substantially affect false-positive rates when there are many studies. Figure 5 illustrates false-positive rates in large data sets ($k = 100$ studies) and $m = 5$ covariates comparing single with multiple meta-regressions as correlation between covariates increases.

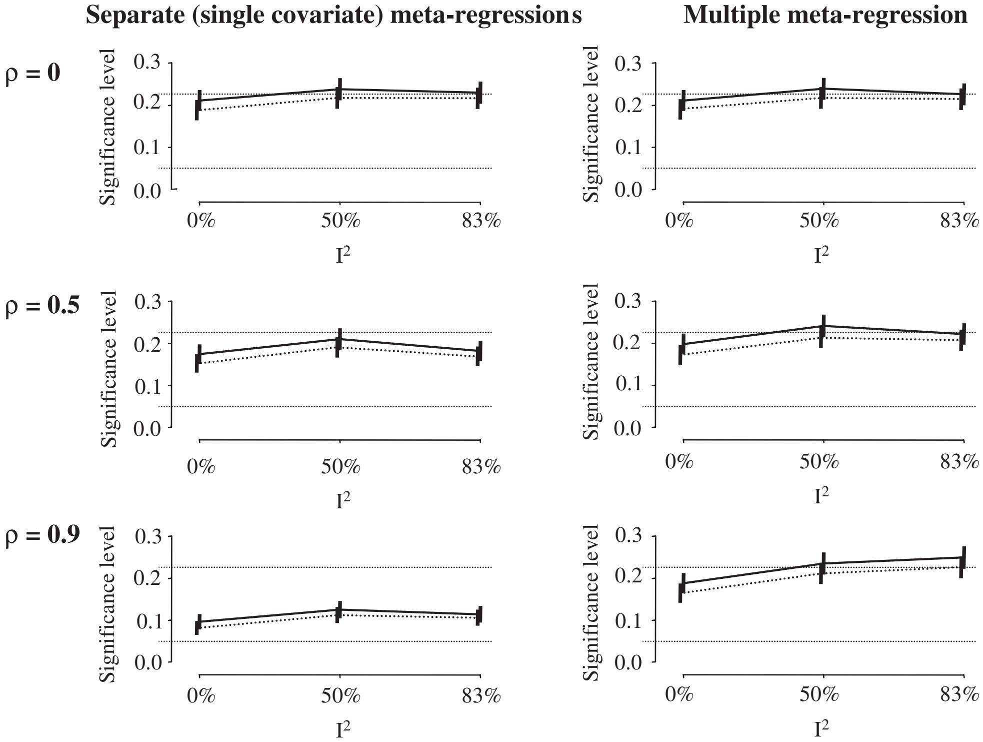


Figure 5. False-positive rates (with 95 per cent confidence intervals) for random effects meta-regression analysis comparing single-variable regressions versus multiple regression with increasing correlation between covariates ($\rho = 0, 0.5, 0.9$). All analyses include 100 studies and five covariates, with a ‘typical’ pattern of within-study variances (see text). Based on 1000 simulations and REML estimation. The nominal 5 per cent significance level and expected significance level from five independent tests with uncorrelated covariates are shown as horizontal dotted lines. Solid line $= R$, dots $= R_{\text{KH}}$.

3.4. Summary and conclusions

Key causes of *increased* false-positive rates in meta-regression are:

- use of fixed effect meta-regression when heterogeneity is present,
- using the standard normal test for random effects meta-regression when there are few studies and heterogeneity is present,
- testing several covariates without adjusting for multiplicity.

A key cause of *decreased* false-positive rates in meta-regression is:

- using the t-test described by Knapp and Hartung when there are few studies and little heterogeneity.

We have observed high rates of false-positive findings from meta-regression as it is typically practised. In particular, fixed effect meta-regressions have unacceptably high type I error rates in the presence of heterogeneity. In random effects meta-regressions, the problem may partly be avoided using tests based on a t-distribution for the test statistic. However, such alternatives can be unreasonably conservative when the number of studies is small. The true distributions of the test statistics are complicated functions of the number of studies, the extent of heterogeneity, the extent of colinearity among the covariates and, more subtly, the pattern of weights across the studies. In the light of this awkward situation, we propose the non-parametric alternative of using a permutation test.

4. A PERMUTATION TEST FOR META-REGRESSION

Permutation tests are well established as a means of calculating significance levels [23]. They are especially appropriate when the data may not be considered randomly sampled from a defined population [24, 25], as is typically the case in meta-analysis. A permutation test may be constructed in the same way as in standard linear regression [26], that is by randomly shuffling the rows of the design matrix and re-assigning them to the response vector. In our situation, we take the 'response' to be the linked pair of an effect estimate and its variance. The covariance structure of the covariates is maintained.

On each random re-allocation of covariates to responses, a test statistic is computed. The true significance level for the relationship between response and a particular covariate is determined by comparing an observed test statistic for the original data set with the distribution of test statistics across random re-allocations. For example, if in 12 out of 1000 re-allocations the original test statistic is equalled or exceeded, then the permutation test p-value is 0.012.

Since the parametric distribution of the test statistic in a permutation test is unimportant, the choice among alternative test statistics is generally wider. Two obvious choices for meta-regression are the T-statistic in (1) and the regression coefficient itself. These are not equivalent choices as they are in unweighted regression [24]. We choose the traditional T-statistic, since this allows easier handling of comparisons across multiple covariates. We do not make the modifications to the standard error proposed by Knapp and Hartung [16].

As an artificial example that illustrates the tempering effect of the permutation test, consider three trials with effect estimates -5, 0 and 5, and each with estimated variance 4, such that approximate 95 per cent confidence intervals for the trials are $(-9,-1)$, $(-4,4)$ and $(1,9)$. Consider a covariate taking values $-1,0$ and 1, respectively. A meta-regression of the effect estimates on the covariate yields coefficient 5, with standard error $\sqrt{2}$, using any of the above methods. Comparing the ratio $5/\sqrt{2}$ to a standard normal distribution produces a p-value of 0.0004. In this situation, there are exactly six possible permutations for the covariate, of which two yield a T-statistic with absolute value at least as high as $5/\sqrt{2}$ (in fact equal to this). The p-value from the permutation test is then $2/6=0.33$. This should lead to considerably more caution than the p-value of 0.0004.

The algorithm for the permutation test with a single covariate is easily implemented:

1. Perform the meta-regression on the original data and store T_{orig}, the test statistic T in (1)
2. Repeat N times:

2.1. Randomly re-allocate the pairs $\{y_i, v_i\}$ to covariate values. This is readily achieved by randomly permuting the indices $i = 1, \ldots, k$ in the pairs $\{y_i, v_i\}$.
2.2. Perform the meta-regression and collect the test statistic T in (1).
3. Determine n, the number of statistics $|T|$ that equal or exceed $|T_{\text{orig}}|$. The permutation test p-value for the meta-regression is n/N.

The method may be extended as follows to assess several covariates taking into account multiple tests, for either multiple regression or a series of single variance regressions. Suppose there are m covariates. Let us define the 'most significant' to be the covariate yielding the largest T-statistic (in absolute value), and the 'least significant' to be that yielding the smallest T-statistic (in absolute value):

1. Perform the meta-regression on the original data, and store $T_{\text{orig}}^{(1)}$, $T_{\text{orig}}^{(2)}, \ldots, T_{\text{orig}}^{(m)}$, the test statistics ordered by magnitude so that $T_{\text{orig}}^{(1)}$ refers to the most significant covariate and $T_{\text{orig}}^{(m)}$ the least significant covariate.
2. Repeat N times:
   2.1. Randomly re-allocate the pairs $\{y_i, v_i\}$ to sets of covariates.
   2.2. Perform the meta-regression and collect the ordered test statistics $T^{(1)}$, $T^{(2)}, \ldots, T^{(m)}$
3. Compare $|T_{\text{orig}}^{(1)}|$ with the collection of $|T^{(1)}|$s, $|T_{\text{orig}}^{(2)}|$ with the collection of $|T^{(2)}|$s and so on, to produce a p-value for each covariate.

All the permutation test results that follow are based on $N = 1000$.

5. APPLICATION OF PERMUTATION TESTS

5.1. Single covariate: BCG data

We use the same data as Berkey *et al.* (using their smoothed estimator for the variances of the log risk ratios), which comprise 13 trials of BCG vaccines (Figure 1). The heterogeneity among these studies may be quantified as $I^2 = 86$ per cent and so is close to the most extreme heterogeneity we evaluated in our simulations (see Figure 3, central graph). We assume that a single covariate, latitude of study, is being investigated as a potential explanation for the heterogeneity. Application of the permutation test to these data is illustrated in Table I. The p-values from the permutation tests are less extreme than from the parametric tests. Of particular note are the p-values from the fixed effect meta-regression. From the parametric test this is grossly small, but from the permutation test it is in line with the results of random effects meta-regressions.

5.2. Multiple covariates analysed separately: exercise and depression data

For illustration we select five of the covariates considered by Lawlor and Hopker (Figure 2) for the potential explanation of heterogeneity among 10 trials of exercise for the management of depression (multiple meta-regression on all eight covariates yields a singular matrix). The heterogeneity is such that $I^2 = 75$ per cent. The correlation matrix for the five covariates is

Table I. Results of meta-regression relating log(risk ratio) of mortality to absolute latitude (degrees) of study; data from Berkey *et al.* [14]. Results for the empirical Bayes method are included for comparison with those presented in the original paper.

Method	Coefficient (SE)	Meta-regression p-value (method)	Permutation test p-value
Fixed effect meta-regression	-0.0282 (0.004)	$<10^{-10}$ (F)	0.031
Random effects meta-regression (REML estimate of τ^2)	-0.0268 (0.011)	0.012 (R) 0.037 (Berkey) 0.032 (R_{KH})	0.054
Random effects meta-regression (empirical Bayes estimate of τ^2)	-0.0268 (0.011)	0.015 (R) 0.041 (Berkey) 0.033 (R_{KH})	0.059

as follows ((a)–(e) defined as in Figure 2).

$$\begin{array}{cc}
 & \begin{array}{ccccc} (a) & (b) & (c) & (d) & (e) \end{array} \\
\begin{array}{c} (a) \\ (b) \\ (c) \\ (d) \\ (e) \end{array} & \begin{pmatrix} 1 & -0.36 & -0.17 & -0.33 & -0.25 \\ \cdot & 1 & 0.18 & 0.45 & -0.15 \\ \cdot & \cdot & 1 & 0.51 & -0.17 \\ \cdot & \cdot & \cdot & 1 & -0.33 \\ \cdot & \cdot & \cdot & \cdot & 1 \end{pmatrix}
\end{array}$$

Results from a series of single covariate meta-regression analyses appear in Table II. Again the fixed effect analysis is seen to be highly anticonservative. Among random effects analyses, test results using the approach of Knapp and Hartung (R_{KH}) are generally in good agreement with the permutation test p-values for this data set, both being considerably more conservative than the test based on a standard normal distribution for the statistic T (R). This is to be expected since the data would lie towards the right of the centre right graph in Figure 4 where R_{KH} is seen to have an acceptable type I error rate. The two covariates identified as being statistically significantly related to the effect size by Lawlor and Hopker (publication as an abstract and length of follow-up) are statistically significant for all methods, although for publication as an abstract the p-value is less extreme by a factor of 10 when using the permutation test compared with the Knapp test.

The above analyses do not address multiplicity, so the final column in Table II illustrates a non-parametric adjustment for multiple testing, describing the degree of 'surprise' one might have about the observed results. Given the five particular covariates investigated, the first p-value describes the likelihood that the observed most significant covariate (publication as an abstract) is compatible with chance. Among the random re-allocations of covariates to responses in the permutation test, the T-statistic for the most significant covariate exceeded the observed T-statistic for publication as an abstract 4 per cent of the time. When examining the second most significant covariate among the permutations, this rarely gave as large a T-statistic as length of follow-up (0.3 per cent). This suggests a relationship exists between effectiveness of exercise and the length of follow-up, as well as with publication as an abstract.

5.3. Multiple covariates analysed together: exercise and depression data

In Table III, we provide results of multiple meta-regression analyses in which five regression coefficients are estimated simultaneously. Regression coefficients for publication as an abstract, follow-up and use of intention-to-treat analysis are similar to the single covariate meta-regressions. The coefficient for allocation concealment is reversed in direction. This covariate is most closely correlated with other variables (see matrix above). Here the Knapp–Hartung test and the permutation test yield very similar results. Again, these results do not address multiplicity.

The final column of the table gives the p-value for the most- to least-significant variables, an analysis that addresses multiplicity. When assessing the most significant variables after adjusting for other variables in the analysis, the findings are now more compatible with chance effects resulting from multiple testing.

Table II. Results of single covariate meta-regression analyses with permutation tests applied to the exercise and depression data of Lawlor and Hopker [19]. Standardized mean differences are related separately to the five covariates, listed in order of increasing p-value.

Covariate	i	Model	Coefficient (SE)	Meta-regression p-value (method)	Permutation test p-value	Permutation test p-value for ith most significant covariate
(a) Published as abstract	1	Fixed effect Random effects	−1.57 (0.31) −1.56 (0.34)	$<10^{-5}$ (F) $<10^{-5}$ (R) 0.002 (R_{KH})	0.02 0.02	0.03 0.04
(b) Follow-up (weeks)	2	Fixed effect Random effects	0.20 (0.05) 0.21 (0.08)	$<10^{-5}$ (F) 0.008 (R) 0.03 (R_{KH})	0.02 0.03	0.002 0.003
(c) Allocation concealment	3	Fixed effect Random effects	0.61 (0.23) 0.52 (0.48)	0.008 (F) 0.31 (R) 0.35 (R_{KH})	0.44 0.36	0.05 0.21
(d) Intention-to-treat	4	Fixed effect Random effects	0.64 (0.36) 0.68 (0.81)	0.08 (F) 0.40 (R) 0.43 (R_{KH})	0.22 0.36	0.11 0.10
(e) Published as PhD thesis	5	Fixed effect Random effects	0.20 (0.38) 0.27 (0.69)	0.59 (F) 0.69 (R) 0.70 (R_{KH})	0.80 0.73	0.25 0.12

Table III. Results of a multiple meta-regression analysis with permutation tests applied to the exercise and depression data of Lawlor and Hopker [19]. Covariates listed in order of increasing p-value. In the random effects multiple meta-regression, the estimated amount of heterogeneity is $\hat{\tau}^2 = 0$, so fixed and random effects meta-regressions coincide. However, p-values from the permutation tests differ according to the meta-regression method applied to the randomly re-arranged data sets.

Covariate	i	Model	Coefficient (SE)	Meta-regression p-value (method)	Permutation test p-value	Permutation test p-value for ith most significant covariate
(a) Published as abstract	1	Fixed effect	−1.34 (0.39)	0.0006 (F)	0.07	0.35
		Random effects	−1.34 (0.39)	0.0006 (R)	0.02	0.11
				0.03 (R_{KH})		
(b) Follow-up (weeks)	2	Fixed effect	0.16 (0.06)	0.01 (F)	0.23	0.33
		Random effects	0.16 (0.06)	0.01 (R)	0.06	0.09
				0.06 (R_{KH})		
(c) Intention-to-treat	3	Fixed effect	0.46 (0.39)	0.24 (F)	0.59	0.69
		Random effects	0.46 (0.39)	0.24 (R)	0.32	0.25
				0.30 (R_{KH})		
(d) Allocation concealment	4	Fixed effect	−0.41 (0.35)	0.25 (F)	0.61	0.33
		Random effects	−0.41 (0.35)	0.25 (R)	0.34	0.10
				0.31 (R_{KH})		
(e) Published as PhD thesis	5	Fixed effect	−0.01 (0.44)	0.98 (F)	0.99	0.94
		Random effects	−0.01 (0.44)	0.98 (R)	0.98	0.88
				0.98 (R_{KH})		

6. DISCUSSION

Heterogeneity in meta-analyses should be investigated to increase the scientific and clinical relevance of their results [27]. Advice to systematic reviewers who wish to explore heterogeneity using statistical techniques is often to minimize the number of covariates investigated, to select those justified through scientific rationale and to specify them in advance. However, we have observed that in practice reviewers often fail to follow these guidelines and may be tempted to examine numerous covariates in an attempt to explain observed heterogeneity [28]. *Post hoc* explorations may yield covariates that turn out to be important when investigated further, but in some fields—particularly meta-analysis of clinical trials—false-positive findings that are mis-interpreted may have damaging consequences in clinical care. In the case of multiple covariates, we agree with Westfall and Young's sentiment that 'subject matter experts need to have both raw and multiplicity adjusted p-values to help them make more informed judgements on the repeatability of their multiple outcome experiments' [29].

We aimed to explore the impact of numbers of covariates, and of other aspects of meta-analytic data sets, on the likelihood of false-positive results in meta-regression. We found fixed effect meta-regression to be unacceptable in the presence of heterogeneity. Random effects meta-regression referring the standard T-statistic to a normal distribution is also highly anticonservative when the number of studies is small. These problems are compounded when multiple covariates are assessed. Knapp and Hartung have recently proposed an alternative t-test for regression coefficients in meta-regression [16]. This has much more appropriate false-positive rates than the standard normal test. All random effects meta-regression methods performed well on single covariates when the number of studies is large, since the standard T-statistic is asymptotically normal.

Since the false-positive rates of various meta-regression methods cannot be satisfactorily summarized analytically, we have proposed a permutation test approach to assess statistical significance in meta-regression. We recommend its use particularly when there are not many studies in a meta-analysis, a situation very common in practice [28]. A particular advantage of the permutation test approach is that it may be used to temper findings for the 'most significant' covariate of several. In an example, we found that $p < 10^{-5}$ for a single 'most significant' covariate among five in separate regressions increased to $p = 0.11$ in a permutation test in which the distribution of 'most significant' findings from multiple regression was examined. Researchers need to be honest about how many covariates were considered as possible explanations of heterogeneity in a meta-analysis, as this fundamentally affects the interpretation of the 'significant' associations found. We note that reliable information on the number of covariates investigated is rarely available in publications. Even in our examples in this paper, we have assumed that only one covariate was ever considered in the BCG trials analysis, and that only five covariates were investigated in the depression trials, both assumptions we know to be incorrect [1, 19]. We stress that we do not advocate the *post hoc* examination of multiple potential causes of heterogeneity in a fishing expedition for significant associations. We align ourselves with those who would discourage meta-regression analyses in the absence of plausible hypotheses [30] and a reasonable volume of reliable data.

An alternative approach to dealing with multiple covariates is a random coefficient model, in which the coefficients are assumed to follow a random effects distribution. Bayes or empirical

Bayes estimates of the coefficients are shrunk towards the mean of the distribution [31, 32]. The shrinkage reduces the risk of false-positive conclusions from the covariates with the strongest observed relationships. Such analyses require an assumption of exchangeability across coefficients for the various covariates, a situation unlikely in a meta-regression context. A more recent suggestion involves random coefficient modelling of observed covariates in addition to fixed effect modelling of selected covariates derived from them [33]. An example is given of a single case-control study investigating the relationship between breast cancer and intake of 35 nutrients. The nutrients are derived from reported intake of 87 food items, and these are themselves included as random effects in the model to allowing for residual associations with breast cancer. Such an approach may sometimes be appropriate for meta-regression, for example by including random coefficients for several components of study quality, while investigating the relationship between effect size and a derived quality score. Such methods are novel suggestions, and have not yet undergone any empirical investigation in the context of meta-regression.

The permutation test evaluates only statistical significance. Confidence intervals for the regression coefficients may be obtained using a similar approach [24], or using a bootstrap procedure. However, this would not address the biases caused by multiplicity when several covariates are investigated, and only the most significant relationships reported.

We have not addressed the power of our meta-regression methodology to detect genuine relationships as statistically significant. Our principal motivation, based on observations of meta-analyses of clinical trials, is to protect researchers from drawing spurious positive conclusions that may adversely affect patient care. We use the same statistic T for the permutation test that is used in model-based random effects meta-analysis. Hence, for moderate or large samples, this has approximately the same power as a parametric method when the model assumptions underlying the parametric test hold [24]. For instance, in our example of a single covariate in Table I, the significance levels are similar. While adjustment for multiple tests does of course reduce power, it has been argued that 'without adjustment, it is too easy to reject null hypotheses, thereby reaching potentially erroneous conclusions' [29]. Meta-regression analyses in practice are also compounded with other difficulties [7] arising from aggregation of patient-level covariates, confounding of study-level covariates, small numbers of studies and selection of scientifically relevant characteristics [28]. The interpretation of detected relationships will always remain hard.

Although our simulation study addressed only continuous covariates we expect similar findings with other types of covariate. Our second example contained a mixture of continuous and binary covariates, and observations were similar irrespective of the nature of the covariate. One factor we have not considered is a dependence of the extent of heterogeneity on the value of a covariate. We have adopted the usual assumption in random effects meta-regression that the heterogeneity variance is constant, although for some covariates this may not be appropriate.

We have demonstrated important limitations of meta-regression analyses in the circumstances in which they are commonly used. Statistically significant results are highly likely if the method does not properly account for the presence of heterogeneity. Examining multiple covariates obviously increases the risk of a false-positive conclusion. In the light of these findings we propose that a permutation test be performed before claiming that a covariate is statistically significant.

ACKNOWLEDGEMENTS

We are indebted to Debbie Lawlor for providing unpublished data from the meta-analysis of exercise for depression, and to referees for helpful comments. JH was supported in part by MRC Grant G9815466.

REFERENCES

1. Colditz GA, Brewer TF, Berkey CS, Wilson ME, Burdick E, Fineberg HV, Mosteller F. Efficacy of BCG vaccine in the prevention of tuberculosis: meta-analysis of the published literature. *Journal of the American Medical Association* 1994; **271**:698–702.
2. Law MR, Wald NJ, Thompson SG. By how much and how quickly does reduction in serum cholesterol concentration lower risk of ischaemic heart disease? *British Medical Journal* 1994; **308**:367–373.
3. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JHP, Bossuyt PMM. Empirical evidence of design-related bias in studies of diagnostic tests. *Journal of American Medical Association* 1999; **282**:1061–1066.
4. Berlin JA, Antman EM. Advantages and limitations of metaanalytic regressions of clinical trials data. *Online Journal of Current Clinical Trials* 1994; **4**:Doc No. 134.
5. Davey Smith G, Egger M, Phillips AN. Meta-analysis: beyond the grand mean? *British Medical Journal* 1997; **315**:1610–1614.
6. Lau J, Ioannidis JP, Schmid CH. Summing up evidence: one answer is not always enough. *The Lancet* 1998; **351**:123–127.
7. Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine* 2002; **21**:1559–1574.
8. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. *Journal of Clinical Epidemiology* 1995; **48**:1503–1510.
9. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology* 1996; **49**:1373–1379.
10. Huque MF, Dubey SD. A meta-analysis methodology for utilizing study-level covariate information from clinical trials. *Communications in Statistics—Theory and Methods* 1994; **23**:377–394.
11. Stram DO. Meta-analysis of published data using a linear mixed-effects model. *Biometrics* 1996; **52**:536–544.
12. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in Medicine* 1999; **18**:2693–2708.
13. Greenland S. Quantitative methods in the review of epidemiologic literature. *Epidemiologic Reviews* 1987; **9**:1–30.
14. Berkey CS, Hoaglin DC, Mosteller F, Colditz GA. A random-effects regression model for meta-analysis. *Statistics in Medicine* 1995; **14**:395–411.
15. Sharp S. Meta-analysis regression. *Stata Technical Bulletin* 1998; **42**:16–22.
16. Knapp G, Hartung J. Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine* 2003; **22**:2693–2710.
17. Hochberg J, Tamhane A. *Multiple Comparison Procedures*. Wiley: New York, 1987.
18. Rothman KJ. No adjustments are needed for multiple comparisons. *Epidemiology* 1990; **1**:43–46.
19. Lawlor DA, Hopker SW. The effectiveness of exercise as an intervention in the management of depression: systematic review and meta-regression analysis of randomised controlled trials. *British Medical Journal* 2001; **322**:1–8.
20. Sánchez-Meca J, Marín-Martínez F. Testing continuous moderators in meta-analysis: a comparison of procedures. *British Journal of Mathematical and Statistical Psychology* 1998; **51**:311–326.
21. Marín-Martínez F, Sánchez-Meca J. Testing for dichotomous moderators in meta-analysis. *Journal of Experimental Education* 1998; **67**:69–81.
22. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* 2002; **21**:1539–1558.
23. Edgington ES. *Randomization Tests*. Marcel Dekker: New York, 1995.
24. Manly BFJ. *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Chapman & Hall: London, 1997.
25. Sprent P. *Data Driven Statistical Methods*. Chapman & Hall: London, 1998.
26. Noreen EW. *Computer-Intensive Methods for Testing Hypotheses: an Introduction*. Wiley: New York, 1989.
27. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. *British Medical Journal* 1994; **309**:1351–1355.
28. Higgins J, Thompson S, Deeks J, Altman D. Statistical heterogeneity in systematic reviews of clinical trials: a critical appraisal of guidelines and practice. *Journal of Health Service Research Policy* 2002; **7**:51–61.
29. Westfall PH, Young SS. *Resampling-Based Multiple Testing*: *Examples and Methods for P-Value Adjustment*. Wiley: New York, 1993.

30. Goodman SN. Multiple comparisons, explained. *American Journal of Epidemiology* 1998; **147**:807–812.
31. DuMouchel WH, Harris JE. Bayes methods for combining the results of cancer studies in humans and other species. *Journal of the American Statistical Association* 1983; **78**:293–308.
32. Greenland S, Robins JM. Empirical Bayes adjustments for multiple comparisons are sometimes useful. *Epidemiology* 1991; **2**:244–251.
33. Greenland S. When should epidemiologic regressions use random coefficients? *Biometrics* 2000; **56**:915–921.