STATISTICS IN MEDICINE
*Statist. Med.* **17**, 2537–2550 (1998)

# META-ANALYSIS OF MULTIPLE OUTCOMES BY REGRESSION WITH RANDOM EFFECTS

C. S. BERKEY[1,2,*], D. C. HOAGLIN[2,3], A. ANTCZAK-BOUCKOMS[2], F. MOSTELLER[2,4] AND G. A. COLDITZ[1,2]

[1]*Channing Laboratory, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, 181 Longwood Avenue, Boston, MA 02115, U.S.A.*
[2]*Technology Assessment Group, Department of Health Policy and Management, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, U.S.A.*
[3]*Abt Associates Inc., 55 Wheeler Street, Cambridge, MA 02138, U.S.A.*
[4]*Department of Statistics, Science Center, Harvard University, Cambridge, MA 02138, U.S.A.*

## SUMMARY

Earlier work showed how to perform fixed-effects meta-analysis of studies or trials when each provides results on more than one outcome per patient and these multiple outcomes are correlated. That fixed-effects generalized-least-squares approach analyzes the multiple outcomes jointly within a single model, and it can include covariates, such as duration of therapy or quality of trial, that may explain observed heterogeneity of results among the trials. Sometimes the covariates explain all the heterogeneity, and the fixed-effects regression model is appropriate. However, unexplained heterogeneity may often remain, even after taking into account known or suspected covariates. Because fixed-effects models do not make allowance for this remaining unexplained heterogeneity, the potential exists for bias in estimated coefficients, standard errors and *p*-values. We propose two random-effects approaches for the regression meta-analysis of multiple correlated outcomes. We compare their use with fixed-effects models and with separate-outcomes models in a meta-analysis of periodontal clinical trials. A simulation study shows the advantages of the random-effects approach. These methods also facilitate meta-analysis of trials that compare more than two treatments.
© 1998 John Wiley & Sons, Ltd.

## INTRODUCTION

When each of the published research papers addressing a particular topic reports results on two or more endpoints or outcomes that have been measured on each patient, meta-analysts often perform a separate synthesis of the results for each endpoint. Because such multiple endpoints are usually correlated, a simultaneous analysis that takes their correlation into account should add efficiency and accuracy.

Raudenbush *et al.*[1] showed how to analyze *effect sizes* (a standardized or scale-free estimate of treatment difference; see p. 10 of Laird and Mosteller[2]) for two or more outcomes jointly in

* Correspondence to: C. S. Berkey, Channing Laboratory, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, 181 Longwood Avenue, Boston, MA 02115, U.S.A. E-mail: catherine.berkey@channing.harvard.edu

Contract/grant sponsor: Upjohn
Contract/grant sponsor: AHCPR

CCC 0277–6715/98/222537–14$17.50
© 1998 John Wiley & Sons, Ltd.
*Received November 1996*
*Accepted February 1998*

a fixed-effects generalized-least-squares (GLS) multiple-regression model that allows adjustment for study-level covariates. This method and others to be described here are relevant to the meta-analysis of studies assessing continuous multiple endpoints measured on each subject, such as systolic and diastolic blood pressures. Gleser and Olkin[3] illustrate the Raudenbush method, along with univariate (separate outcomes) approaches and composite effect sizes, for the synthesis of correlated effect sizes using data from studies that evaluated coaching for improving math and verbal scores on the Scholastic Aptitude Test (SAT).

A recent generalization by Berkey *et al.*[4,5] for the regression-synthesis of multiple outcomes keeps the measured outcomes in their original units (rather than creating standardized effect sizes) to simplify the interpretation of the results (see Mosteller and Colditz,[6] p. 8). This newer method also provides other benefits, such as allowing the inclusion of trials that compare three or more treatments or various subsets of the treatments considered by the meta-analysis (see Berkey *et al.*[5]).

These fixed-effects GLS regression models analyze the multiple outcomes jointly within a single model, and they allow inclusion of covariates, such as duration of therapy or quality of trial, that may explain observed heterogeneity of results among the trials.[1,4,5] Sometimes the covariates explain all the among-trial variation, and the fixed-effects multiple-outcomes regression model is appropriate, but unexplained heterogeneity may often remain, even after exploration of the contributions of all known or suspected covariates. Fixed-effects models do not take this unexplained heterogeneity into account, and the potential exists for bias in estimated coefficients, their standard errors and *p*-values. Thus, inferences can be misleading.

A report by the National Research Council[7] (see p. 185) recommends random-effects models as preferable to fixed-effects models for meta-analysis. Thus, we propose two random-effects approaches for the regression meta-analysis of multiple correlated outcomes. We apply these two approaches in an example and compare their results with results from our fixed-effects multiple-outcomes model, with separate-outcomes fixed-effects regression, and with separate-outcomes random-effects regression.

## METHODS

### Data

A recent meta-analysis[8] by Antczak-Bouckoms *et al.* located five randomized controlled trials that compared a surgical (S) procedure (modified Widman flap) with a non-surgical (NS) procedure (scaling and root planning or curettage with anaesthesia) for the treatment of moderate periodontal disease. All these studies used a split-mouth design, where segments of each patient's mouth were randomly allocated to different treatment procedures. Each patient had at least one section treated surgically and at least one other section treated non-surgically. The two outcomes assessed on each patient were (pre- to post-treatment *mm* changes in) probing depth (PD) and attachment level (AL). The goal of treatment is to decrease probing depths and to increase attachment levels around the teeth. All trials were of similar quality; neither the patients nor the observers could be adequately blinded to treatment allocation. Mean ages of patients were similar across the five trials (ranging from 40 to 48 years), and all trials included both males and females. All studies provided results for patients evaluated at one year after treatment, and all provided results on patients with periodontal disease of medium severity (baseline probing depths between 4 and 6 *mm*). Thus, the design of this study[8] has already addressed many of the factors that one

© 1998 John Wiley & Sons, Ltd. 

might consider as sources of heterogeneity in a meta-analysis. In spite of this, considerable heterogeneity is present among the trials' results for both PD (estimated among-trial variance $\tau^2 = 0{\cdot}0102$, test for homogeneity[9] $Q = 12{\cdot}82$, $p = 0{\cdot}012$) and AL ($\hat{\tau} = 0{\cdot}0573$, $Q = 112{\cdot}08$, $p < 0{\cdot}001$).

A potential factor that remains is year of publication of the study results, a proxy for time of conduct of a trial. With more surgical experience, the efficacy of a surgical procedure may improve over time, so that later studies might demonstrate greater benefit. When this occurs, we might expect the present-day benefit to be better than the 'average benefit' computed from studies completed between 10 and 20 years ago. Thus, if the year of the trial is correlated with the trial results, there may be important clinical implications.

For these reasons, we wish to fit a meta-regression model that includes year of publication as a covariate. Our earlier work[4] on these data focused on fixed-effects estimation of the main effects of treatment while taking into account the correlation between the two outcomes PD and AL. It turns out that year of publication explains some, but not all, of the among-trial variation in the results, and thus the fixed-effects regression method may provide misleading inferences.

Table I presents the data that we need for model fitting: from each trial, the estimated benefit of surgical treatment over non-surgical treatment for PD and for AL (positive values indicate that surgery provides the better patient outcome); the within-trial covariance matrix $S_i$ of the two outcomes (the square roots of the diagonal elements of $S_i$ are the standard errors of the mean outcomes), and the year of publication of each trial.

## Models

In the case of two correlated outcomes, all the methods that we consider provide a separate regression model for predicting each outcome from the covariates. The same covariates need not appear in both regression models, but our example uses the same covariate. Thus, a particular covariate might suggest improvement in one outcome and at the same time deterioration in the other, or it may suggest benefit (or harm) to both outcomes. However, because the two outcomes are correlated, the two regression models are not independent and should not be estimated as if they were. Thus, the methods below all provide a separate regression model for each outcome, but the estimated regression coefficients and their standard errors may differ among methods, depending upon the particular statistical assumptions made.

One could apply other methods (summary statistics) that we do not consider here to meta-analysis of multiple outcomes.[10] Those methods have the benefit of providing a single overall probability statement concerning treatment differences, but they assume that the direction of the effect of a particular treatment is the same for all outcomes. This is not a suitable assumption for our example, where surgery is better for one outcome (PD) but non-surgical treatment is better for the other outcome (AL).

The *general random-effects* (*RE*) *multiple-outcomes meta-regression model* is

$$\mathbf{y}_i = X_i\boldsymbol{\beta} + \boldsymbol{\delta}_i + \mathbf{e}_i \tag{1}$$

where: $\mathbf{y}_i$ is a vector of $p$ outcomes (means) reported by trial $i$ (most estimation methods below do not require that all $p$ outcomes be reported by trial $i$); $X_i$ is a matrix containing the observed trial-level covariates for trial $i$ – its pattern of entries indicates what outcome appears in each row of $\mathbf{y}_i$, and to which outcome each covariate relates; $\boldsymbol{\beta}$ is the vector of regression coefficients to estimate – it may include a separate intercept for each outcome and a separate slope for each

© 1998 John Wiley & Sons, Ltd. 

Table I. Results from five published trials comparing surgical and non-surgical treatments for medium--severity periodontal disease, one year after treatment. Patient outcomes are improvements (mm) in probing depth (PD) and attachment level (AL). Positive surgical minus non-surgical values (S − NS) indicate that surgery is better. $S_i$ is the within-trial covariance matrix of the two outcomes (means) in trial $i$

| Trial | Publication year | Number of patients | Improvement in Probing depth S − NS | Improvement in Attachment level S − NS | $S_i$ PD | $S_i$ AL |
|---|---|---|---|---|---|---|
| 1 | 1983 | 14 | +0·47 | −0·32 | 0·0075<br>0·0030 | 0·0030<br>0·0077 |
| 2 | 1982 | 15 | +0·20 | −0·60 | 0·0057<br>0·0009 | 0·0009<br>0·0008 |
| 3 | 1979 | 78 | +0·40 | −0·12 | 0·0021<br>0·0007 | 0·0007<br>0·0014 |
| 4 | 1987 | 89 | +0·26 | −0·31 | 0·0029<br>0·0009 | 0·0009<br>0·0015 |
| 5 | 1988 | 16 | +0·56 | −0·39 | 0·0148<br>0·0072 | 0·0072<br>0·0304 |

outcome against each corresponding covariate; $\boldsymbol{\delta}_i$ is a vector of $p$ random effects associated with trial $i$. These represent, for each outcome, the trial's true deviation from the true mean of all trials having the same covariate values (specified in $X_i$). The $p \times p$ $\text{cov}(\boldsymbol{\delta}_i) = D$ is to be estimated – it represents the among-trial covariance that is unexplained by the regression. The assumption that the $\boldsymbol{\delta}_i$ arise from a multivariate Gaussian distribution MVN $(\mathbf{0}, D)$ allows us to compute probabilities, but most of the analysis may proceed without this assumption (it is not required for estimating the $\beta$'s and their standard errors); and $\mathbf{e}_i$ (independent of $\boldsymbol{\delta}_i$) is the vector of random sampling errors within trial $i$, having $p \times p$ covariance matrix $S_i$, which we assume is known (but is customarily estimated/reported by the individual trials). If each $n_i$ is sufficiently large, then the vector $\mathbf{e}_i$ is approximately MVN $(\mathbf{0}, S_i)$. Thus

$$\text{cov}(\mathbf{y}_i) = D + S_i \quad \text{and} \quad \mathbf{y}_i \sim \text{MVN}(X_i\boldsymbol{\beta}, D + S_i).$$

In the *random-effects separate-outcomes model*, we assume that the off-diagonal elements of each $S_i$ and the off-diagonal elements of $D$ are zero.

In the *fixed-effects multiple-outcomes model*, all the $\boldsymbol{\delta}_i = \mathbf{0}$ and the matrix $D = 0$.

In the *fixed-effects separate-outcomes model*, all the $\boldsymbol{\delta}_i = \mathbf{0}$, the matrix $D = 0$, and the off-diagonal elements of each $S_i$ are zero.

## Estimation

We observe the $\mathbf{y}_i$, $X_i$, and $S_i$, and we desire to estimate $\boldsymbol{\beta}$ and $D$. Aside from the different assumptions outlined above, the three random-effects methods (to be described in Sections 3 to 5) each estimate the components of $D$ in a slightly different manner. First we review estimation for the GLS fixed-effects multiple-outcomes model.

© 1998 John Wiley & Sons, Ltd. 

*1. GLS fixed-effects multiple-outcomes regression model*

The *GLS fixed-effects multiple-outcomes* regression model[4,5] is estimated by univariate GLS regression with stacked response vectors. The method begins by stacking the $p$ outcomes from the $k$ studies. For example, for the $k = 5$ periodontal trials with $p = 2$ outcomes (Table I), we stack the $2 \times 1$ $\mathbf{y}_i$ into the $10 \times 1$ ($kp \times 1$) outcome vector $\mathbf{y}$:

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_5 \end{bmatrix} = \begin{bmatrix} \text{study 1 surgical} - \text{non-surgical PD} \\ \text{study 1 surgical} - \text{non-surgical AL} \\ \vdots \\ \text{study 5 surgical} - \text{non-surgical PD} \\ \text{study 5 surgical} - \text{non-surgical AL} \end{bmatrix}. \tag{2}$$

The $10 \times 10$ ($kp \times kp$) block-diagonal matrix $S$

$$S = \text{diag}(S_1, S_2, S_3, S_4, S_5) \tag{3}$$

is the fixed-effects estimate of the covariance matrix of $\mathbf{y}$.

In the fixed-effects linear model

$$\mathbf{y} = X\boldsymbol{\beta} + \mathbf{e} \tag{4}$$

the full-rank matrix $X$ contains values of the predictor variables of the regression model (including study-level covariates), $\boldsymbol{\beta}$ is a column vector of $r$ regression coefficients to estimate, and $\mathbf{e}$ is a vector of random errors.

The $10 \times 4$ ($kp \times r$) design matrix $X$ for the present application includes for each outcome an intercept and a slope term for the covariate (year of publication $-$ 1983); the $10 \times 4$ matrix (five $2 \times 4$ matrices, one for each study, stacked) is

$$X = \begin{bmatrix} X_1 \\ \cdot \\ X_i \\ \cdot \\ X_5 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ & \cdot & & \\ & \cdot & & \\ & \cdot & & \\ 1 & 0 & 5 & 0 \\ 0 & 1 & 0 & 5 \end{bmatrix}. \tag{5}$$

To fit this FE model by GLS regression, we calculate

$$\hat{\boldsymbol{\beta}} = (X^{\mathrm{T}} S^{-1} X)^{-1} X^{\mathrm{T}} S^{-1} \mathbf{y} \tag{6}$$

and

$$\widehat{\text{cov}}(\hat{\boldsymbol{\beta}}) = (X^{\mathrm{T}} S^{-1} X)^{-1}. \tag{7}$$

Seber[11] (p. 60) reviews the underlying theory for GLS estimation.

The $4 \times 1$ vector $\hat{\boldsymbol{\beta}}$ in this example thus contains summary estimates of the surgical minus non-surgical effects for probing depth and for attachment level, and the effects of year of trial on each outcome. The square roots of the diagonal elements of $\widehat{\text{cov}}(\hat{\boldsymbol{\beta}})$ are the standard errors of these estimates (and the off-diagonals are their covariances). Thus, we may use the estimated

© 1998 John Wiley & Sons, Ltd. 

$\boldsymbol{\beta}$ and $\mathrm{cov}(\boldsymbol{\beta})$ to test the null hypothesis, $H_0: \boldsymbol{\beta} = \mathbf{0}$, that surgical and non-surgical therapies are equally effective for both probing depth and attachment level, regardless of year of trial. We can also test the significance of other multivariate hypotheses or of individual $\beta$'s.

### 2. *Separate-outcomes fixed-effects regression model*

If we set the off-diagonal elements of the matrix $S$ above to zero in the calculation of both $\hat{\boldsymbol{\beta}}$ and $\widehat{\mathrm{cov}}(\hat{\boldsymbol{\beta}})$, we obtain estimates for the (independent or uncorrelated) *separate-outcomes fixed-effects model*. Alternatively, we can estimate this model by weighted least squares (WLS) regression, performed separately for each outcome, where the weights are the reciprocal variances $1/S_{ij}$ ($S_{ij}$, which appears on the diagonal of the $S_i$ matrix, is the variance of (mean) outcome $j$ in study $i$).

### 3. *GLS random-effects multiple-outcomes regression model*

We propose a *GLS random-effects multiple-outcomes* regression model for meta-analysis that is a simple extension of the fixed-effects GLS multiple-outcomes model.[4,5] Basically, we replace the within-trial covariance matrices, $S_i$, wherever they appear in the estimation for our non-iterative fixed-effects multiple-outcomes method, by $D + S_i$, where $D$ is the current estimate of the among-trial covariance (which is unexplained by the regression model). We iterate between estimating the $p \times p$ covariance matrix $D$ and estimating the regression coefficient vector $\boldsymbol{\beta}$.

Iteration 1 fits the GLS fixed-effects multiple-outcomes model as described above (equations (2)–(6)). We then compute the $kp \times 1$ vector of residuals $(\mathbf{y} - X\hat{\boldsymbol{\beta}})$ and rearrange these residuals into a $k \times p$ matrix $(\mathbf{y} - X\hat{\boldsymbol{\beta}})_{k \times p}$, where row $i$ contains the residuals for trial $i$ and column $p$ contains the residuals for outcome $p$. Our initial estimate of the covariance matrix $D$ is

$$\hat{D} = (k-2)^{-1}(\mathbf{y} - X\hat{\boldsymbol{\beta}})^{\mathrm{T}}_{p \times k}(\mathbf{y} - X\hat{\boldsymbol{\beta}})_{k \times p} - k^{-1} \sum_{i=1}^{k} S_i. \tag{8}$$

(We use the factor $(k-2)^{-1}$ because for each outcome we estimate two parameters, an intercept and a slope.) To begin iteration 2, we replace $S_i$ by $\hat{D} + S_i$ in the fixed-effects multiple-outcomes estimators (equation (3)), to obtain random-effects weights for each trial, $(\hat{D} + S_i)^{-1}$. Using these revised weights, we calculate a new estimate of $\boldsymbol{\beta}$ (equation (6)), then new residuals $(\mathbf{y} - X\hat{\boldsymbol{\beta}})$ and a new $\hat{D}$ (equation (8)). We iterate between estimating $D$ from residuals and estimating the regression coefficients $\boldsymbol{\beta}$, until the estimates converge (our convergence criterion is that each element of $\hat{D}$ change, between iterations, by less than 0·00001 (tolerance)).

This general approach, as well as the fixed-effects approach described above, does not require that all included trials provide results on both (or all) outcomes, although every trial in our example did so; nor does it require that all trials present results from the same set of treatment groups (but all five trials have compared the same two treatments). Some parts of the calculation, particularly the process summarized in equation (8), become more complicated, however.

### 4. *Random-effects separate-outcomes regression model*

We here fit the *random-effects separate-outcomes regression model* for the analysis of the two outcomes PD and AL separately, using the estimation procedure of Berkey *et al.*[12] The model looks like the general random-effects multiple-outcomes model in equation (1), except that now $\mathbf{y}_i$, $\boldsymbol{\delta}_i$ and $\mathbf{e}_i$ are scalars rather than vectors, and $X_i$ is now a vector rather than a matrix and contains only those covariates that affect the one outcome. Estimation is by iteratively reweighted

© 1998 John Wiley & Sons, Ltd. 

least squares, where we estimate the among-trial variance $D_j$ of outcome $j$ as suggested by Morris[13]

$$\hat{D}_j = \frac{\sum_i (\hat{D}_j + S_{ij})^{-1}\{(k/k - r_j))(y_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})^2 - S_{ij}\}}{\sum_i (\hat{D}_j + S_{ij})^{-1}} \tag{9}$$

where $S_{ij}$ is the variance of mean outcome $j$ reported by trial $i$, and $r_j$ is the number of regression parameters to estimate in the model for outcome $j$.

Alternatively, one could set the off-diagonal elements of the $S_i$ and of $D$ to zero in the RE multiple-outcomes procedure of the previous section, but then we would estimate the among-trial variance of each outcome differently than shown here.

### 5. *MML random-effects multiple-outcomes regression model*

We also propose a *random-effects multiple-outcomes* model that we estimate by multivariate maximum likelihood (MML), relying upon results in Dempster *et al.*,[14] as described in the appendix in Berkey and Laird.[15] We can use this approach only when all the trials in the meta-analysis provide results for all the outcomes considered (or else, omit from the synthesis any trial that does not report on all $p$ outcomes).

We also estimate this model by iterating between estimating the among-trial covariance $D$ and estimating $\boldsymbol{\beta}$. Assuming that $\mathbf{y}_i \sim \text{MVN}(X_i\boldsymbol{\beta}, S_i + D)$ as before

$$\hat{\boldsymbol{\beta}} = (\Sigma_i X_i^{\mathrm{T}}(S_i + \hat{D})^{-1} X_i)^{-1} \Sigma_i X_i^{\mathrm{T}}(S_i + \hat{D})^{-1}\mathbf{y}_i \tag{10}$$

and

$$\hat{D} = \Sigma_i[\hat{D}(S_i + \hat{D})^{-1}(\mathbf{y}_i - X_i\hat{\boldsymbol{\beta}})(\mathbf{y}_i - X_i\hat{\boldsymbol{\beta}})^{\mathrm{T}}(S_i + \hat{D})^{-1}\hat{D} + \hat{D} - \hat{D}(S_i + \hat{D})^{-1}\hat{D}]/k \tag{11}$$

where $\hat{D}$ on the right hand side is the estimate from the previous iteration. After convergence, we estimate the covariance matrix of $\hat{\boldsymbol{\beta}}$ by

$$\widehat{\text{cov}}(\hat{\boldsymbol{\beta}}) = (\Sigma_i X_i^{\mathrm{T}}(S_i + \hat{D})^{-1} X_i)^{-1}. \tag{12}$$

Illustrative code written in SAS PROC IML,[16] for fitting each of these five models to the periodontal disease data in Table I, is available from the first author.

## Simulation Study

Because the main purpose of this paper is to propose that one should also use random-effects methods for multiple-outcomes meta-analysis, we compare the properties of the fixed-effects and random-effects versions of the GLS multiple-outcomes analyses. We also use simulation to investigate the relative properties of the separate-outcomes and multiple-outcomes analyses. So that our comparisons reflect only the two specific issues that concern us here (univariate or multivariate analysis, fixed-effects or random-effects analysis), we use the following four GLS methods on the same simulated data:

(i) fixed-effects multiple-outcomes;
(ii) fixed-effects separate-outcomes (obtained here by setting off-diagonal entries of $S_i$ to zero in the GLS FE multiple-outcomes method);
(iii) random-effects multiple-outcomes;
(iv) random-effects separate-outcomes (obtained here by setting off-diagonal entries of both $S_i$ and $D$ to zero in the GLS random-effects multiple-outcomes method).

© 1998 John Wiley & Sons, Ltd. 

The Berkey *et al.*[12] separate-outcomes RE regression method and the proposed MML RE[15] approach estimate the among-trial covariances quite differently than do the GLS RE approaches. We do not include them in these simulations because differences in the simulated results might be attributable to the particular estimator of $D$ used rather than to the two issues of primary interest.

We simulated 5000 meta-analyses. Each meta-analysis consisted, as in our periodontal meta-analysis, of $k = 5$ trials that report results on two correlated outcomes. We assumed a linear covariate that corresponds to year of trial, centred at 1983; we assumed for each trial one of the covariate values observed in the five real trials. The regression model parameters and among-trial covariance matrix $D$ used to generate the data were the estimates from the GLS multiple-outcomes model on the real data, and the five within-trial covariance matrices were the $S_i$ listed in Table I.

The random numbers for the simulations came from the SAS[17] function RANNOR, a Gaussian random number generator. We use the bivariate normal distribution to generate correlated errors; $\delta_{1i}$ is generated as $N(0, D_{11})$, and then $\delta_{2i}$ is randomly generated using the conditional distribution of $\delta_2$ given $\delta_1 \sim N(\rho\delta_1\sqrt{(D_{22}/D_{11})}, (1-\rho^2)D_{22})$, where $\rho$ is the correlation between $\delta_2$ and $\delta_1$. Correlated $\mathbf{e}_i$'s are similarly generated, independent of the $\boldsymbol{\delta}_i$'s, and assuming the bivariate normal distribution ($\mathbf{e}_i \sim N(\mathbf{0}, S_i)$).

For each of the four methods, we evaluated bias in the estimated parameters and coverage of their 95 per cent confidence intervals, and also the results of testing two multivariate hypotheses: (i) that the treatment effects are both zero ($H_0$: $\beta_1 = \beta_2 = 0$); (ii) that the covariate $x$ has no effect on the two outcomes ($H_0$: $\beta_3 = \beta_4 = 0$).

## RESULTS

Each trial individually suggests that surgical treatment results in better probing depths (PD) one year after treatment, but non-surgical treatment results in better attachment levels (AL) by one year after treatment (Table I). The matrices $S_i$ embody a correlation between the two outcomes of approximately 0·40. Although lacking power for a sample size of 5, tests failed to reject normality for either outcome ($p = 0{\cdot}80$, $p = 0{\cdot}79$).

For each method, Table II shows the assumed values (zeros) or the estimated values for the among-trial covariance matrix $D$. The within-trial covariance matrices $S_i$ (Table I) are used in all the multiple-outcomes approaches (both FE and RE); in the separate-outcomes approaches their off-diagonal elements are ignored (assumed zero). Because all three random-effects methods (separate outcomes, multiple-outcomes GLS, multiple-outcomes MML) use iterative estimation, we show the number of iterations required to achieve convergence. The estimated regression coefficients with their standard errors, from each of the five methods, are shown across the bottom half of Table II.

The year of publication ($x$) was not an important predictor of either PD or AL outcome, although it approached significance in the FE separate-outcomes analysis of AL (because of the low standard error) (Table II). The direction of the coefficient of year (from all five methods) might suggest that more recent trials report greater benefit of non-surgical treatment for outcome AL.

The most notable finding in Table II is that all the random-effects standard errors are considerably larger than the corresponding fixed-effects standard errors. This example might suggest that going from a FE separate-outcomes model to a FE multiple-outcomes model has virtually no effect on the estimated standard errors; the impact is more on the estimated regression coefficients. For the RE methods, the differences between the separate-outcomes and

© 1998 John Wiley & Sons, Ltd. 

Table II. Five meta-regression methods fit to periodontal trial results, with covariate year of publication ('$x$' is the year of publication minus 1983). $D$ is the among-trial covariance that is unexplained by the regression model. All elements of $D$ are assumed to be zero by FE methods. Multiple-outcomes methods use the full $2 \times 2$ $S_i$ matrices; separate-outcomes methods ignore (effectively set to 0) the off-diagonal elements of $D$ and $S_i$

| Fixed effects | | | Random effects | | |
|---|---|---|---|---|---|
| Separate outcomes | Multiple outcomes GLS | | Separate outcomes | Multiple outcomes GLS | Multiple outcomes MML |
| Iterations = 1 (tolerance = 0·00001) | Iterations 1 | | Iterations 5 for PD 6 for AL | Iterations 5 | Iterations 25 |
| $D$ matrix | | | | | |
| PD $\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$ | $\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$ | | $\begin{bmatrix} 0{\cdot}020 & 0 \\ 0 & 0{\cdot}036 \end{bmatrix}$ | $\begin{bmatrix} 0{\cdot}022 & 0{\cdot}013 \\ 0{\cdot}013 & 0{\cdot}028 \end{bmatrix}$ | $\begin{bmatrix} 0{\cdot}008 & 0{\cdot}009 \\ 0{\cdot}009 & 0{\cdot}025 \end{bmatrix}$ |
| Outcome PD models | | | | | |
| $Y = \underset{(\mathrm{SE}(\beta_1))}{\beta_1} + \underset{(\mathrm{SE}(\beta_3))}{\beta_3 x}$ | | | | | |
| PD = 0·345 − 0·008$x$ (0·029) (0·008) | 0·304 − 0·005$x$ (0·029) (0·008) | | 0·363 + 0·005$x$ (0·073) (0·022) | 0·359 + 0·005$x$ (0·075) (0·022) | 0·348 + 0·001$x$ (0·052) (0·015) |
| Outcome AL models | | | | | |
| $Y = \underset{(\mathrm{SE}(\beta_2))}{\beta_2} + \underset{(\mathrm{SE}(\beta_4))}{\beta_4 x}$ | | | | | |
| AL = −0·394 − 0·012$x$ (0·019) (0·007) | −0·394 −0·009$x$ (0·019) (0·007) | | −0·340 −0·014$x$ (0·092) (0·028) | −0·336 −0·011$x$ (0·083) (0·026) | −0·335 −0·011$x$ (0·079) (0·024) |

multiple-outcomes estimates (coefficients and standard errors) were fairly minor. It appears that differences between fixed-effects and random-effects methods have a greater overall impact (via the standard errors) than differences between single-outcomes and multiple-outcomes estimation. We explore this more fully later.

We replicated this exercise (not shown), with no covariates in the model (having only intercept terms) and using the DerSimonian and Laird[9] random-effects method instead of random-effects regression.[12] We reached essentially the same conclusions, most notably the underestimation of standard errors by fixed-effects methods.

## Adequacy of Simulation

The 25,000 $\boldsymbol{\delta}$'s (= 5 trials × 5000 meta-analyses) that we generated had variance 0·022 for the first outcome (alias PD), variance 0·028 for the second outcome (alias AL), and covariance 0·013. These correspond perfectly to the covariance matrix $D$ that we assumed for the simulation study.

The 5000 $\mathbf{e}_1$ vectors that we generated for trial 1, assuming the covariance $S_1$ (Table I), had a covariance matrix that differed from $S_1$ by less than 0·0002 in each of the 4 cells. The covariance matrix of the 5000 $\mathbf{e}_2$ vectors for trial 2 was also within 0·0002 of the assumed $S_2$. The 5000 $\mathbf{e}_3$ for trial 3 (and similarly for trial 4) had a covariance matrix that was within 0·0001 of $S_3$ ($S_4$). For the fifth trial, the covariance of the generated $\mathbf{e}_5$ vectors was within 0·0003 of $S_5$ (in Table I).

© 1998 John Wiley & Sons, Ltd. 

## Simulation results

### *Fixed effects versus Random effects*

Table III summarizes the results from fitting the GLS fixed-effects and random-effects multiple-outcomes methods to the simulated data of 5000 meta-analyses with a covariate. Table IV shows the corresponding results for the separate-outcomes methods.

For the random-effects methods, we used a convergence criterion (for the changes between iterations) of less than 0·0001 (tolerance) in each of the elements of the estimated matrix $D$. For this simulation study, we stopped at 20 iterations if convergence had not yet been reached. Over half the RE meta-analyses converged to solutions by the fourth iteration, and only 114 of 5000 random-effects multiple-outcomes models had not yet converged by iteration 20.

As indicated in Table III, the FE method assumes that the among-trial covariance matrix $D$ is zero everywhere; for the RE method, the mean $\hat{D}$ computed from the 5000 estimates of $D$ suggests that the random-effects estimator of $D$ has little bias.

Both FE and RE multiple-outcomes methods appeared to provide unbiased estimates of the regression coefficients ($\hat{\beta}$'s in Table III). The standard deviations of the 5000 FE and RE estimates were also somewhat comparable (the standard deviation of 5000 $\hat{\beta}_1$ was 0·084 for the FE estimates and 0·077 for the RE estimates). The major difference lies in the standard errors reported by the methods (the mean of 5000 FE standard errors of $\hat{\beta}_1$ was 0·029 versus the RE mean standard error of 0·071; similarly for the other betas). For all four betas, the mean of the 5000 RE reported standard errors corresponds better (than FE) to the observed standard deviation of the estimated $\beta$. As a result, significance tests and 95 per cent confidence intervals of the FE and RE estimates differ greatly, with the RE method providing the more truthful results. Using the $t$-distribution on 3 (5 trials minus 2 estimated parameters per outcome) degrees of freedom, the RE method comes close to providing confidence intervals with 95 per cent coverage (Table III).

We evaluated the multivariate hypothesis that the treatment effects on both outcomes are null ($\chi^2_{2\text{df}}$ test of the hypothesis $H_0$: $\beta_1 = \beta_2 = 0$). Because the simulated treatment effects were so large, all 5000 meta-analyses, both FE and RE, rejected this hypothesis (Table III). We finally evaluated the multivariate hypothesis that the covariate $x$ had no effect on either outcome ($\chi^2_{2\text{df}}$ test of the hypothesis $H_0$: $\beta_3 = \beta_4 = 0$). The fixed-effects method rejected the null hypothesis 82 per cent of the time, considerably more often than the RE method (36 per cent), due to the FE bias of standard errors towards zero (Table III).

The comparison of fixed-effects and random-effects separate-outcomes methods (Table IV) provided conclusions very similar to those from Table III. Neither FE nor RE separate-outcomes method appeared to provide biased estimates of $\boldsymbol{\beta}$, but the random-effects estimators again provided more appropriate standard errors, confidence intervals and hypothesis tests.

### *Separate-outcomes versus multiple-outcomes*

To compare separate-outcomes versus multiple-outcomes estimation, we compare Table III with Table IV, focusing in turn on the FE results and the RE results. The only important difference is in the performance of the multivariate test ($\chi^2_{2\text{df}}$) of the hypothesis $H_0$: $\beta_3 = \beta_4 = 0$. The multivariate RE method (Table III), which takes into account the correlation between $\hat{\beta}_3$ and $\hat{\beta}_4$, is more efficient and rejects the null hypothesis more often (36 per cent of the time for the RE multiple-outcomes method versus 26 per cent for the RE separate-outcomes method). As noted

© 1998 John Wiley & Sons, Ltd. 

Table III. Fixed-effects and random-effects multiple-outcomes GLS regression analysis in 5000 simulated meta-analyses (of five trials). The 'true' $D$ and regression coefficients used in generating the $\mathbf{y}_i$ are shown on the left. The $S_i$ assumed for the five trials appear in the Table I. The $\beta_3$ and $\beta_4$ are coefficients for the study-level covariate $x$

| True model | | Fixed-effects | Random-effects |
|---|---|---|---|
| | | (assumed) | (means) |
| $D = \begin{bmatrix} 0{\cdot}022 & 0{\cdot}013 \\ 0{\cdot}013 & 0{\cdot}028 \end{bmatrix}$ | | $\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$ | $\begin{bmatrix} 0{\cdot}024 & 0{\cdot}013 \\ 0{\cdot}013 & 0{\cdot}029 \end{bmatrix}$ |
| First outcome ('PD') | | | |
| $\beta_1 = 0{\cdot}359$ | mean $\hat{\beta}_1$ (SD) | 0·359 (0·084) | 0·359 (0·077) |
| | (mean of SE) | (0·029) | (0·071) |
| $\beta_3 = 0{\cdot}005$ | mean $\hat{\beta}_3$ (SD) | 0·005 (0·025) | 0·005 (0·023) |
| | (mean of SE) | (0·008) | (0·021) |
| Second outcome ('AL') | | | |
| $\beta_2 = -0{\cdot}336$ | mean $\hat{\beta}_2$ (SD) | −0·338 (0·099) | −0·337 (0·084) |
| | (mean of SE) | (0·019) | (0·074) |
| $\beta_4 = -0{\cdot}011$ | mean $\hat{\beta}_4$ (SD) | −0·011 (0·029) | −0·011 (0·026) |
| | (mean of SE) | (0·006) | (0·023) |
| Coverage of 95% confidence interval ($t$ 3 d.f.): | | | |
| $\beta_1$ | | 0·73 | 0·96 |
| $\beta_2$ | | 0·46 | 0·92 |
| $\beta_3$ | | 0·84 | 0·95 |
| $\beta_4$ | | 0·53 | 0·93 |
| $H_0$: $\beta_1 = \beta_2 = 0$ proportion rejected: | | 1·0 | 1·0 |
| $H_0$: $\beta_3 = \beta_4 = 0$ proportion rejected: | | 0·82 | 0·36 |

earlier, the fixed-effects methods both reject $H_0$ far too often (82 per cent and 80 per cent) because they ignore the unexplained among-trial variation.

## DISCUSSION

This work demonstrates that fixed-effects regression models for correlated outcomes (and also for single outcomes) may seriously underestimate the standard errors of regression coefficients when the regression model does not explain all the among-trial heterogeneity. The result is that 95 per cent confidence intervals for $\hat{\boldsymbol{\beta}}$ may not be wide enough, $p$-values may be biased downward, and incorrect conclusions for hypothesis tests may be reported. The two random-effects multiple-outcomes models that we proposed provided estimated models and inferences that are more realistic than are the fixed-effects estimates, which assume that the components of $D$ are all zero.

When we compared the RE separate-outcomes to the multiple-outcomes models in a simulation study, we observed that the multiple-outcomes methods provided more efficient multivariate hypotheses tests. We believe that these multivariate tests are more sensitive (than other aspects of these models) to the strength of the correlations between the outcomes. We did not note any biases or difference in efficiencies for univariate tests (on single regression coefficients).

Our simulations on the RE methods also suggested that using the $t$-distribution for inferences worked well, providing coverages near 95 per cent for all four $\hat{\beta}$'s. Earlier work on fixed-effects

© 1998 John Wiley & Sons, Ltd. 

Table IV. Fixed-effects and random-effects separate-outcomes GLS regression analysis of 5000 simulated meta-analyses (of five trials). The 'true' $D$ and regression coefficients used in generating the $\mathbf{y}_i$ are shown on the left. The $S_i$ assumed for the five trials appear in the Table I. The $\beta_3$ and $\beta_4$ are coefficients for the study-level covariate $x$

| True model | | Fixed-effects | Random-effects |
|---|---|---|---|
| | | (assumed) | (diagonal means) |
| $D = \begin{bmatrix} 0{\cdot}022 & 0{\cdot}013 \\ 0{\cdot}013 & 0{\cdot}028 \end{bmatrix}$ | | $\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$ | $\begin{bmatrix} 0{\cdot}024 & 0{\cdot}000 \\ 0{\cdot}013 & 0{\cdot}029 \end{bmatrix}$ |
| First outcome ('PD') | | | |
| $\beta_1 = 0{\cdot}359$ | mean $\hat{\beta}_1$ (SD) | 0·359 (0·083) | 0·359 (0·077) |
| | (mean of SE) | (0·029) | (0·071) |
| $\beta_3 = 0{\cdot}005$ | mean $\hat{\beta}_3$ (SD) | 0·005 (0·025) | 0·005 (0·023) |
| | (mean of SE) | (0·008) | (0·021) |
| Second outcome ('AL') | | | |
| $\beta_2 = -0{\cdot}336$ | mean $\hat{\beta}_2$ (SD) | −0·337 (0·098) | −0·337 (0·084) |
| | (mean of SE) | (0·019) | (0·074) |
| $\beta_4 = -0{\cdot}011$ | mean $\hat{\beta}_4$ (SD) | −0·011 (0·029) | −0·011 (0·026) |
| | (mean of SE) | (0·007) | (0·023) |
| Coverage of 95% confidence interval ($t$ 3 d.f.): | | | |
| $\beta_1$ | | 0·73 | 0·96 |
| $\beta_2$ | | 0·48 | 0·92 |
| $\beta_3$ | | 0·84 | 0·95 |
| $\beta_4$ | | 0·53 | 0·93 |
| $H_0$: $\beta_1 = \beta_2 = 0$ proportion rejected: | | 1·0 | 0·99 |
| $H_0$: $\beta_3 = \beta_4 = 0$ proportion rejected: | | 0·80 | 0·26 |

multiple-outcomes regression[1,4,5] used the normal distribution. Other investigators have explored using a $t$-distribution, rather than the normal distribution, in the framework of the DerSimonian and Laird method.[9]

Our simulation study was limited, in that we made assumptions that corresponded to the actual meta-analysis that inspired this work. In other words, we assumed a meta-analysis of five trials with two outcomes, with the $S_i$, $D$, and $\boldsymbol{\beta}$ that we observed in the periodontal trials. Because our findings regarding the comparison of fixed-effects with random-effects methods are consistent with findings in other situations,[9,12] we expect that our conclusions (regarding underestimation of the standard errors of FE regression model parameters) generalize. However, our conclusions regarding the benefits (greater efficiency of multivariate hypothesis tests) of the multivariate versus univariate analysis may not generalize so readily. Also, the lack of bias in $\hat{\boldsymbol{\beta}}$ that we noted in the separate-outcomes models may not generalize.

The methods presented here are applicable to multiple continuous outcomes measured on each subject (for example, systolic and diastolic pressures). Alternatively, Berlin *et al.*[18] discuss methods for meta-analysing dose–response studies, a different type of multiple endpoint that cannot be analysed by our methods, in which each study provides a series of dose-specific relative risks, with one category serving as the common reference group. Those authors consider both fixed- and random-effects models for epidemiologic dose–response studies. Another type of multiple endpoint arises when each trial or study reports survival proportions at a series of time

© 1998 John Wiley & Sons, Ltd. 

points. Dear[19] extended the method of Raudenbush *et al.*[1] to survival analysis, showing how to estimate the correlations among the serial survivorships and allowing the survival proportions reported at multiple times by the trials to be analysed together in a fixed-effects model.

The application of regression methods to meta-analysis can identify important sources of heterogeneity, provide answers to questions not yet envisioned at the start of the individual trials, and describe/adjust/control for covariate differences among the studies. Our primary purpose in initiating this work was to explore the feasibility of fitting random-effects (rather than fixed-effects) models for multiple-outcomes regresssion meta-analysis. A disadvantage of the MML method is that it requires that all studies report on all the outcomes, whereas the proposed random-effects GLS method does not, but we are encouraged by the broad similarities among the estimates (regression coefficients and their standard errors) of all three random-effects methods in Table II (a separate-outcomes method and two multiple-outcomes methods).

Our example presented a very simple situation in which all trials compared the same two treatments (surgical and non-surgical), and all trials reported results for each of the multiple outcomes (PD and AL), but these methods are also applicable to more-complicated meta-analyses, such as a meta-analysis of three treatments and four outcomes, in which any trial may consider only a subset of the treatments or outcomes (see Berkey *et al.*[5]). We have shown that random-effects models for multiple-outcomes regression meta-analysis are feasible and worthy of further research.

ACKNOWLEDGEMENTS

Grants from AHCPR and Upjohn to the Harvard School of Public Health provided funds for this methodological work. We thank Dr. Bruce Pihlstrom for his contributions to preliminary phases of this research, and the reviewers for their very helpful suggestions. The authors dedicate this paper to the memory of Dr. Anthony Bouckoms and Christopher Bouckoms.

## REFERENCES

1. Raudenbush, S., Becker, B. and Kalaian, H. 'Modeling multivariate effect sizes', *Psychological Bulletin*, **103**, 111–120 (1988).
2. Laird, N. and Mosteller, F. 'Some statistical methods for combining experimental results', *International Journal of Technology Assessment in Health Care*, **6**, 5–30 (1990).
3. Gleser, L. and Olkin, I. 'Stochastically dependent effect sizes', *in* Cooper, H. and Hedges, L. (eds), *The Handbook of Research Synthesis*, Russell Sage Foundation, New York, 1994, pp. 339–355.
4. Berkey, C., Antczak-Bouckoms, A., Hoaglin, D., Mosteller, F. and Pihlstrom, B. 'Multiple-outcomes meta-analysis of treatments for periodontal disease', *Journal of Dental Research*, **74**, 1030–1039 (1995).
5. Berkey, C., Anderson, J. and Hoaglin, D. 'Multiple-outcome meta-analysis of clinical trials', *Statistics in Medicine*, **15**, 537–557 (1996).
6. Mosteller, F. and Colditz, G. 'Understanding research synthesis (meta-analysis)', *Annual Review of Public Health*, **17**, 1–23 (1996).
7. National Research Council. *Combining Information*: *Statistical Issues and Opportunities for Research*, National Academy Press, Washington DC, 1992.
8. Antczak-Bouckoms, A., Joshipura, K., Burdick, E. and Tulloch, J. F. C. 'Meta-analysis of surgical versus non-surgical method of treatment for periodontal disease', *Journal of Clinical Periodontology*, **20**, 259–268 (1993).
9. DerSimonian, R. and Laird, N. 'Meta-analysis in clinical trials', *Controlled Clinical Trials*, **7**, 177–188 (1986).
10. O'Brien, P. 'Procedures for comparing samples with multiple endpoints', *Biometrics*, **40**, 1079–1087 (1984).
11. Seber, G. A. F. *Linear Regression Analysis*, Wiley, New York, 1977.

© 1998 John Wiley & Sons, Ltd. 

12. Berkey, C., Hoaglin, D., Mosteller, F. and Colditz, G. 'A random-effects regression model for meta-analysis', *Statistics in Medicine*, **14**, 395–411 (1995).
13. Morris, C. N. 'Parametric empirical Bayes inference: theory and applications', *Journal of the American Statistical Association*, **78**, 47–55 (1983).
14. Dempster, A., Laird, N. and Rubin, D. 'Maximum likelihood with incomplete data via the EM algorithm', *Journal of the Royal Statistical Society*, *Series B*, **39**, 1–38 (1977).
15. Berkey, C. and Laird, N. 'Nonlinear growth curve analysis: estimating the population parameters', *Annals of Human Biology*, **13**, 111–128 (1986).
16. SAS Institute, Inc. *SAS/IML User's Guide*, Version 5 Edition, SAS Institute, Inc., Cary, NC, 1985.
17. SAS Institute, Inc. *SAS User's Guide*: *Basics*, Version 5 Edition, SAS Institute Inc. , Cary, NC, 1985.
18. Berlin, J. A., Longnecker, M. P. and Greenland, S. 'Meta-analysis of epidemiologic dose-response data', *Epidemiology*, **4**, 218–228 (1993).
19. Dear, K. B. G. 'Iterative generalized least squares for meta-analysis of survival data at multiple times', *Biometrics*, **50**, 989–1002 (1994).

© 1998 John Wiley & Sons, Ltd. 