STATISTICS IN MEDICINE, VOL. 15, 537-557 (1996)

MULTIPLE-OUTCOME META-ANALYSIS OF CLINICAL TRIALS

C. S. BERKEY*

Technology Assessment Group, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, U.S.A.

J. J. ANDERSON

Arthritis Center, Boston University School of Medicine, 80 E. Concord Street, Boston, MA 02118, U.S.A.

AND

D. C. HOAGLIN

Abt Associates Inc., 55 Wheeler Street, Cambridge, MA 02138, U.S.A.

SUMMARY

When several clinical trials report multiple outcomes, meta-analyses ordinarily analyse each outcome separately. Instead, by applying generalized-least-squares (GLS) regression, Raudenbush *et al.* showed how to analyse the multiple outcomes jointly in a single model. A variant of their GLS approach, discussed here, can incorporate correlations among the outcomes within treatment groups and thus provide more accurate estimates. Also, it facilitates adjustment for covariates. In our approach, each study need not report all outcomes nor evaluate all treatments. For example, a meta-analysis may evaluate two or more treatments (one 'treatment' may be a control) and include all randomized controlled trials that report on any subset (of one or more) of the treatments of interest. The analysis omits other treatments that these trials evaluated but that are not of interest to the meta-analyst. In the proposed fixed-effects GLS regression model, study-level and treatment-arm-level covariates may be predictors of one or more of the outcomes. An analysis of rheumatoid arthritis data from trials of second-line drug treatments (used after initial standard therapies prove unsatisfactory for a patient) motivates and applies the method. Data from 44 randomized controlled trials were used to evaluate the effectiveness of injectable gold and auranofin on the three outcomes tender joint count, grip strength, and erythrocyte sedimentation rate. The covariates in the regression model were quality and duration of trial and baseline measures of the patients' disease severity and disease activity in each trial. The meta-analysis found that gold was significantly more effective than auranofin on all three treatment outcomes. For all estimated coefficients, the multiple-outcomes model produced moderate changes in their values and slightly smaller standard errors, to the three separate outcomes models.

INTRODUCTION

Often investigators conducting a randomized clinical trial have an opportunity to report a variety of outcome measures and thus to compare several aspects of the treatments. Consequently, the actual reporting varies among studies. Some report relatively many outcome measures, others report only a few, and still others construct a single composite measure that combines several variables.

[*] Also at the Channing Laboratory, Harvard Medical School and Brigham and Women's Hospital, Boston, MA 02115, U.S.A.

CCC 0277-6715/96/050537-21

Received July 1994
Revised March 1995

This diversity of outcome measures poses a challenge for the meta-analyst who later wishes to combine the results of the studies. If all trials compare the same two groups, such as a particular treatment with a control or alternatively two specific treatments, it may be appropriate to express the difference between the two groups on each outcome measure as an effect size and then apply the method of Raudenbush *et al.*[1] An effect size is standardized (scale-free) mean difference: the ratio of the difference between treatment and control group means to the standard deviation in the control group.[2] Raudenbush *et al.*[1] use generalized-least-squares (GLS) regression[3] to take advantage of correlation, within treatment groups, among the effect sizes for the multiple outcomes. Their approach also handles data from studies that report different subsets of the outcomes.

In addition to not reporting the same outcomes, many studies do not include the same treatments. For example, one group of clinical trials in rheumatoid arthritis compares injectable gold to a placebo, whereas another group of clinical trials compares auranofin to a placebo, and yet other groups compare injectable gold to auranofin, with or without placebo controls. If we wish to include more than two treatment types in a single meta-analysis, and no single treatment or control group appears in every study to serve as the common group for the computation of effect sizes, then a GLS analysis in terms of effect sizes becomes substantially more complicated.

Along these lines, Dear[4] presents a method for the meta-analysis of survival data, in which the multiple outcomes are survival proportions reported at multiple time points. He uses GLS regression to fit models with covariates to data that can include single- and multi-arm trials, as well as two-arm comparative trials.

Working within the framework of randomized trials, we present a similar model for two or more distinct measurable outcomes recorded for each patient. For example, previous work in clinical trials in rheumatoid arthritis showed that tender joint count, erythrocyte sedimentation rate and grip strength (all frequently reported measures) reflect important aspects of disease state.[5]

Taking the multiple-outcomes effect-size regression model of Raudenbush *et al.*[1] as our starting point, we develop a multiple-outcomes model that does not use effect sizes. This approach involves fairly standard multivariate theory,[6] which we recapitulate in the present context. Both approaches incorporate the correlations between the multiple outcomes, providing more accurate estimates. Thus, we present a fixed-effects model for GLS regression analysis in the meta-analysis setting for outcomes in their original units (not effect sizes), thereby simplifying the interpretation of the results. An initial application of our method compared the improvements in two correlated outcomes (probing depth and attachment level) in patients treated both surgically and non-surgically for periodontal disease.[7]

Our model can include more than two treatment groups from multi-arm trials (one arm for each separate treatment or control therapy) and can also include a single arm from randomized trials that include any one of the treatments. Users of this approach may additionally require that each study have a common placebo. If the studies all report on a common pair of treatments, our method can directly analyse the within-trial comparisons.[7] The approach of Raudenbush *et al.*[1] would compute effect sizes, whereas our model would compute differences (in the original units for each outcome) between the two treatments. If each study reports changes (from before to after treatment) in outcomes for both treatment groups, then we could analyse the within-trial differences of these changes. Thus, when the data for the meta-analysis permit, our method can directly analyse the within-trial comparisons. However, as our application illustrates, it handles a broader range of situations.

Similar to Raudenbush and to Dear, each study need not report all outcomes, and, similar to Dear's model in the survival data setting, this method can combine more diverse trials and more than two treatments into a single meta-analysis, thus achieving a more powerful analysis.

The GLS method requires estimates of the correlations among the various outcomes, which the individual trials do not generally report. Estimates of these correlations often are available from some external source. Alternatively, we may treat the impact of assumed values as a problem in sensitivity analysis. In our example, we estimate treatment-specific correlations among three outcomes from the patient-level data of one of the studies included in the meta-analysis. Although estimates obtained in this manner and applied to all trials may seem problematic, our view is that estimated correlations drawn from one trial are often closer to the unreported correlations for the remaining trials than is zero (the usual separate single-outcome analyses assume zero correlation).

GENERALIZED-LEAST-SQUARES MODEL

Many randomized clinical trials evaluate the progress of patients by comparing each patient's value on one or more outcome measures to the corresponding baseline value. Thus, we focus on designs that have baseline values (measurements) and measurements after treatment for each treatment group. (The method, however, does not require baseline measurements.) Because analyses of before and after values arise often in clinical research, including the application that motivated this work, we assume this structure throughout the development that follows.

We consider K studies, each evaluating the effects of one or more treatments, and possibly also a placebo, on one or more of P outcome variables. The total number of treatments and placebos evaluated is $T_{\max}$. In study i and treatment group t let y_{itjpb} denote the score for patient j on the pth outcome variable before (b) the start of treatment (or placebo). Similarly, y_{itjpa} is the same patient's score after (a) treatment. The sample size in study i for treatment group t is n_{it}.

Assume that in treatment group $t = 1, \dots, T_i$ of study $i = 1, \dots, K$,

$$E(y_{itjpb}) = \mu_{itpb} \quad \text{and} \quad \operatorname{var}(y_{itjpb}) = \sigma^2_{itpb}$$

$$E(y_{itjpa}) = \mu_{itpa} \quad \text{and} \quad \operatorname{var}(y_{itjpa}) = \sigma^2_{itpa}$$

for subjects $j = 1, \dots, n_{it}$ and outcomes $p = 1, \dots, P_i$. We discuss correlations among these y's later. (We may label the outcomes differently in different studies, so that outcome $p = 1$ need not be the same measure for all i, and similarly for treatment $t = 1$.) For some inferences, it may be useful to assume that the y_{itjpb} and the y_{itjpa} are Gaussian. One might also assume a common before-treatment σ^2_{ipb} or a common after-treatment σ^2_{ipa} or even a common before- and after-treatment σ^2_{ip}. We retain different values when the published studies provide separate estimates, before and after, by treatment group. The outcome measures, or dependent variables, for analysis are

$$\bar{y}_{itpb} - \bar{y}_{itpa},$$

the difference, for study i and treatment t, between the means on outcome p before and after treatment. For other applications, one may find after-minus-before differences preferable. These calculations apply only to patients measured both before and after treatment. We drop the subject subscript j hereafter, except where needed for clarity, because published reports do not ordinarily provide subject-level data.

For study i and treatment t, the $P_i \times 1$ vector of outcome changes due to treatment has expectation and covariance matrix

$$E(\bar{\mathbf{y}}_{it\Box b} - \bar{\mathbf{y}}_{it\Box a}) = \boldsymbol{\mu}_{it\Box b} - \boldsymbol{\mu}_{it\Box a},$$

$$\operatorname{cov}(\bar{\mathbf{y}}_{it\Box b} - \bar{\mathbf{y}}_{it\Box a}) = \textstyle\sum_{it}.$$

The covariance reflects the correlation among the different outcome measurements and the pairing of the before and after measurements. The bold $\bar{\mathbf{y}}$ are vectors of means of outcomes, and the $\square$ in the subscript maintains the place of the outcome subscript, p, which the vectors now incorporate.

Under current publication practices, reports of studies commonly give $\bar{y}_{itpb}$ and $\bar{y}_{itpa}$ and the standard errors of these means. A report may also give $\bar{y}_{itpb} - \bar{y}_{itpa}$, but it seldom includes either the standard error of this difference or the covariances or correlations between the differences for the individual outcomes. Because we need the covariance matrix $\sum_{it}$ of $\bar{\mathbf{y}}_{it\square b} - \bar{\mathbf{y}}_{it\square a}$ for the GLS analysis, we must either obtain it from the investigators responsible for the study or estimate it from a combination of statistics given in the report and values (standard errors and correlations) that we assume. If the studies for meta-analysis typically report the standard error of each before–after difference, then we would take advantage of that information in estimating $\sum_{it}$. Regrettably, most reports do not give either these standard errors or all the ingredients needed to calculate them. Even the $\text{SE}(\bar{y}_{itpb})$ and $\text{SE}(\bar{y}_{itpa})$ are not always reported. For the one-outcome situation, Follmann *et al.*[8] provide heuristic suggestions for imputing variances from partial variance information. If possible, we obtain a separate estimate of the correlation matrix for each treatment. One may use study-specific correlations if available, or one may pool them.[9] Even though it would be useful, we do not ordinarily expect the published report of a study to include these correlations separately by treatment group. Thus, we must often estimate the correlations from available external data and check that the resulting correlation matrix is positive-definite. In our example, we use sensitivity analysis to assess how the meta-analytic conclusions respond to variation in these correlations. In what follows we construct the necessary covariance matrix from the standard errors of the $\bar{y}_{itpb}$ and the $\bar{y}_{itpa}$ and the correlations between the corresponding patient-level variables.

We estimate the covariance matrix $\sum_{it}$ by $S_{it} = C_i \, \text{SE}_{it} \, \text{CORR}_{it} \, \text{SE}_{it} \, C_i^{\text{T}}$, (for a matrix C and a vector $\mathbf{y}$, Morrison[10] (page 84: eqns 23 and 26) derives the covariance of $C\mathbf{y}$, a set of linear combinations of the elements of a random vector) where the $P_i \times 2P_i$ matrix

$$C_i = \begin{bmatrix} 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & -1 & \cdots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & -1 \end{bmatrix}$$

and the diagonal matrix

$$\text{SE}_{it} = \text{diag}(\text{SE}(\bar{y}_{it1b}), \text{SE}(\bar{y}_{it1a}), \ldots, \text{SE}(\bar{y}_{itP_ib}), \text{SE}(\bar{y}_{itP_ia}))$$

contains the standard errors of the means of each outcome in study i, before and after treatment t. CORR_{it} is the ith study's within-treatment-group-t population correlation matrix of the patient-level variables,

$$y_{itj1b}, y_{itj1a}, y_{itj2b}, y_{itj2a}, \ldots, y_{itjP_ib}, y_{itjP_ia},$$

the P_i outcomes measured before and after treatment t.

Raudenbush *et al.*[1] use univariate GLS regression[3] with stacked response vectors instead of classical multivariate regression. This avoids the missing values that multivariate regression would involve when not all studies report the same outcome variables. Even if all outcome variables were present in all trials, we would still need to weight multivariate regression because the response variables are means (rather than individual observations) with different sample sizes.

We therefore proceed in the same manner as Raudenbush *et al.*[1] by stacking all the study-and-treatment-specific $P_i \times 1$ vectors of outcome changes into a single vector:

$$\mathbf{d} = \begin{array}{l} \left.\begin{array}{c} \bar{\mathbf{y}}_{11\Box b} - \bar{\mathbf{y}}_{11\Box a} \\ \bar{\mathbf{y}}_{12\Box b} - \bar{\mathbf{y}}_{12\Box a} \\ \vdots \\ \bar{\mathbf{y}}_{1T_1\Box b} - \bar{\mathbf{y}}_{1T_1\Box a} \end{array}\right\} \text{from study } i = 1 \\ \left.\begin{array}{c} \bar{\mathbf{y}}_{21\Box b} - \bar{\mathbf{y}}_{21\Box a} \\ \vdots \\ \bar{\mathbf{y}}_{2T_2\Box b} - \bar{\mathbf{y}}_{2T_2\Box a} \\ \vdots \end{array}\right\} \text{from study } i = 2 \\ \left.\begin{array}{c} \bar{\mathbf{y}}_{K1\Box b} - \bar{\mathbf{y}}_{K1\Box a} \\ \vdots \\ \bar{\mathbf{y}}_{KT_K\Box b} - \bar{\mathbf{y}}_{KT_K\Box a} \end{array}\right\} \text{from study } i = K. \end{array}$$

The length of $\mathbf{d}$ is $m = \sum_{i=1}^{K} T_i P_i$, the total number of outcomes reported for all treatment groups by the K studies. If every study reported all P outcomes for all $T_{\max}$ treatment arms, then $m = KT_{\max}P$. (If all studies reported on the same 1 outcome and compared the same two treatments, this $\mathbf{d}$ vector reduces to the pre-post differences, alternating by treatment. This $\mathbf{d}$ vector, if included in a model with a suitable design matrix, allows us to estimate the difference between the two treatments' pre-post differences.)

We form the block-diagonal $m \times m$ matrix S, which is the covariance matrix of $\mathbf{d}$,

$$S = \operatorname{diag}(S_{11}, \dots, S_{1T_1}, \dots, S_{i1}, \dots, S_{iT_i}, \dots, S_{K1}, \dots, S_{KT_K}).$$

Each diagonal block of S is $P \times P$ for each study i that reports all P outcomes. For studies reporting $P_i < P$ outcomes, the non-singular S_{it} is $P_i \times P_i$.

We create a design matrix X that contains dummy variables to indicate type of outcome, as well as the usual dummy predictor variables to indicate type of treatment and other categorical covariates, and we create separate columns for each covariate for each outcome. For example, if there are three separate outcomes, then we may estimate the overall mean for each outcome by stacking $I_{3\times 3}$ matrices as the first three columns of X. Furthermore, we represent the effect of a particular treatment-arm-level covariate $\mathbf{x}_1$, assumed to have a different effect on each of the three outcomes, by stacking the diagonal matrices $\mathbf{x}_{1it} I_{3\times 3}$ to obtain the next three columns of the design matrix. Thus, a model that includes this 6-column X matrix would have six regression coefficients (β's): two coefficients for each outcome variable, an intercept and a coefficient for the covariate $\mathbf{x}_1$. If a second covariate $\mathbf{x}_2$ were thought to modify only one of the three outcomes, say the first one, then we would include it by stacking vectors $\mathbf{x}_{2it}[1\ 0\ 0]^{\mathrm{T}}$ to obtain another column for X. For any study that does not report all P outcomes or all $T_{\max}$ treatments considered in the meta-analysis, we do not include the corresponding rows of $\mathbf{d}$ and X and the corresponding rows and columns of S. The Appendix gives the details of $\mathbf{d}$ and X for two simple examples.

To fit by GLS the linear model

$$\mathbf{d} = X\boldsymbol{\beta} + \mathbf{e}$$

where $\text{cov}(\mathbf{e}) = S$ and the full-rank matrix X contains the predictor variables (treatment-group variables and other study-specific or treatment-arm-specific covariates), we calculate

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

and

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

Seber[3] (p. 60) discusses the underlying theory for GLS estimation.

Hypothesis testing proceeds as usual, as we now briefly summarize. We evaluate the goodness of fit of the model, assuming $\mathbf{d}$ is Gaussian, by comparing the generalized residual sum of squares

$$H_E = (\mathbf{d} - X\hat{\boldsymbol{\beta}})^{\mathrm{T}}S^{-1}(\mathbf{d} - X\hat{\boldsymbol{\beta}}),$$

with the χ^2_{m-r} distribution, where m equals the dimension of $\mathbf{d}$ and r equals the dimension of $\boldsymbol{\beta}$. Under the null hypothesis, the only source of unexplained variation in each of the P outcome variables (the before-minus-after-treatment changes) is sampling error. A large H_E indicates lack of fit that we could reduce by including more treatment-group and study-level covariates when available, or by using a random-effects model. Testing the significance of the full model (H_0: $\boldsymbol{\beta} = \mathbf{0}$) requires comparison of

$$H_R = \mathbf{d}^{\mathrm{T}}S^{-1}\mathbf{d} - H_E$$

with the χ^2_r distribution. We evaluate individual β's by dividing each estimated coefficient by its standard error to obtain a z statistic. We also may test a composite linear hypothesis, H_0: $\Gamma\boldsymbol{\beta} = \mathbf{0}$, by computing $\text{cov}(\Gamma\hat{\boldsymbol{\beta}}) = \Gamma\text{cov}(\hat{\boldsymbol{\beta}})\Gamma^{\mathrm{T}}$. Then division of each linear contrast by its standard error yields a z statistic.

Readers may obtain the SAS/IML code (on a disk and on paper) from the first author.[11] The disk includes a small example dataset and code for fitting a GLS model with treatment variables and one continuous covariate.

APPLICATION

Data

Efficacy of second-line drugs in rheumatoid arthritis (that is, drugs used after initial standard therapies are unsuccessful) is evaluated on a variety of measures, often including tender joint count, erythrocyte sedimentation rate (ESR) and grip strength. Tender joint count is the number of joints (out of a total of 68 evaluated) that are tender on examination.[12] Patients with the most disease activity may have 35 or more tender joints, whereas those with least disease activity may have only 1 or 2. The ESR, measured in mm per hour, is a laboratory measure of inflammation; it is generally low (<11) when the disease is relatively inactive, but it may exceed 92 in very active disease. Grip strength, measured in mmHg using a sphygnamometer, is less than 300 mmHg in most people, but may be 50 mmHg or lower when the disease is severe. We illustrate the fixed-effects GLS regression model for multiple-outcome meta-analysis by summarizing results from 44 clinical trials of treatments for rheumatoid arthritis, in which we focus on these three outcomes.

These 44 studies are a subset of the clinical trials included in a meta-analysis that compared seven agents and placebo.[13] We analyse the results of treatment arms from placebo-controlled and comparative studies of two treatments, injectable gold and auranofin. Placebo-controlled studies include 4 of gold (with a total of 316 patients followed in the placebo and gold arms), 4 of

Table I. Unweighted mean ($\bar{x}$) of study-level changes (before- minus after-treatment means) in each outcome by treatment group. The (unweighted) standard deviations (SD) of these study-level changes and the number of studies reporting them, K, are shown*

Change (before – after) in outcome		Treatment		
		Gold	Auranofin	Placebo
Tender joint count (number of joints)	$\bar{x}$	9·77	8·18	6·03
	SD	4·37	3·04	4·24
	K	23	22	9
Grip strength (mmHg)	$\bar{x}$	– 34·04	– 19·04	– 13·83
	SD	27·19	11·31	9·32
	K	20	19	9
ESR (mm/hour)	$\bar{x}$	21·34	9·15	1·69
	SD	10·93	9·46	2·50
	K	25	23	7

* For tender joint count and ESR, positive $\bar{x}$ implies improvement after treatment; for grip strength, negative $\bar{x}$ implies improvement after treatment. Baseline means (ranges) are 21·88 (12 to 40) joints for tender joint count, 127·65 mmHg (22·8 to 285) for grip strength, and 44·88 mm/hour (22 to 75) for ESR

auranofin (761 patients), and 2 with gold and auranofin in the same study (187 patients). Studies without placebos include 7 comparative trials of gold versus auranofin (162 patients), 12 auranofin treatment arms (819 patients) from comparative trials that compared auranofin with a treatment other than gold, and 15 gold treatment arms (270 patients) from trials that compared gold with a treatment other than auranofin, for a total of 63 treatment arms (that is, 10 placebo, 25 auranofin, and 28 gold) and 2515 patients. Of the 44 trials, only 6 did not report tender joint count, 11 did not report grip strength, and 4 did not report ESR.

Table I summarizes the reported changes in each outcome (before-treatment minus after-treatment means) for each treatment. These are simple unweighted means of the trial-arm summary values. The initial fit of our GLS model to the data revealed some questionable data values (of these 12 values, three were subsequently corrected, two ESR's and one grip strength, and the remainder were confirmed). The results in Table I and hereafter are based on the corrected values, obtained by re-checking the original publications.

We estimated treatment-specific correlations (Table II) among the multiple outcomes from the patient-level data of a trial that compared auranofin, injectable gold, and placebo.[14] Our analysis applied these treatment-specific correlations among the changes (before- minus after-treatment) in the three outcomes to all studies. We derived the correlations shown from the CORR_{it} matrices described in the Model section. We used treatment-specific standard deviations computed from this particular trial for any other trial that did not report this information.

Because study participants in these RCTs were randomly allocated to treatment arm, our GLS analysis regards treatment arms as independent units. Therefore, we do not expect any substantial correlation among the outcomes between arms within a study.

Model with treatment variables

Our initial model contains treatment variables only (Table III), that is, no study-level or arm-level covariates. Table III presents the estimated β, SE($\hat{\beta}$), and associated significance levels. The dependent variable is change (before minus after) in outcome (joint, grip and ESR) during the

Table II. Correlations among the three outcomes (before- minus- after-treatment changes in tender joint count, grip strength and ESR) for subjects in three treatment groups (data from Ward *et al.*[14]). Also shown are correlations of each outcome before treatment with that outcome after treatment, and the baseline standard deviations

Before – after treatment changes in outcome	Joint	Grip	ESR	Before with after correlation	Baseline SD
Injectable gold (N = 54)					
Tender joint count	1			0·65	16·3
Grip strength	– 0·21	1		0·77	60·1
ESR	+ 0·11	– 0·21	1	0·65	28·7
Auranofin (N = 64)					
Tender joint count	1			0·67	15·5
Grip strength	– 0·42*	1		0·77	60·6
ESR	+ 0·34*	– 0·10	1	0·76	30·9
Placebo (N = 43)					
Tender joint count	1			0·64	14·7
Grip strength	– 0·36*	1		0·77	53·1
ESR	– 0·04	– 0·39*	1	0·49	23·0

* $p < 0·05$

Table III. GLS multiple-outcome model for comparing drug treatments (but without allowing for any covariates). The estimated β's represent the mean change (before minus after treatment) in the specified outcome (joint, grip, ESR) for patients in each treatment group relative to the placebo

		β	SE(β)	β/SE(β)	p-value	
Placebo effects	joint	5·61	0·42	13·32	0·0000	
	grip	– 10·28	2·00	– 5·15	0·0000	
	ESR	2·17	1·01	2·14	0·0323	
Auranofin effects	joint	1·69	0·51	3·31	0·0009	auranofin over placebo effects
	grip	– 9·42	2·35	– 4·01	0·0001	
	ESR	7·47	1·13	6·61	0·0000	
Gold effects	joint	3·50	0·56	6·22	0·0000	gold over placebo effects
	grip	– 20·07	2·82	– 7·11	0·0000	
	ESR	18·48	1·30	14·21	0·0000	

$H_E = 569·9$ 148 d.f. $p \approx 0$

$H_R = 2185·8$ 9 d.f. $p \approx 0$

Composite hypothesis test: gold (G) versus auranofin (A) β's

	G – A	SE (G – A)	p-value	
Joint	1·81	0·47	0·0000	if positive, gold better
Grip	– 10·65	2·35	0·0000	if negative, gold better
ESR	11·00	0·96	0·0000	if positive, gold better

trial. The first three coefficients shown in Table III are intercept terms, which represent the mean changes among placebo subjects. For tender joint count and ESR, a positive change indicates improvement, whereas a negative change in grip strength represents improvement. We represent each of the treatments auranofin and gold with an indicator variable (1 for the specified treatment and 0 otherwise), so that, for example, the auranofin coefficients estimate the effect of auranofin on changes in joint, grip and ESR beyond those changes observed in placebo subjects. From Table III we estimate that placebo subjects have 5·61 fewer tender joints after the trials, auranofin-treated subjects have 7·30 fewer tender joints (5·61 + 1·69), and gold-treated subjects have 9·11 (5·61 + 3·50) fewer tender joints after the trials. The p-values indicate that not only did auranofin- and gold-treated subjects experience significant improvement in tender joint count, grip strength and ESR beyond the placebo effects, but the placebo subjects also experienced significant improvement in all three outcomes. The statistic H_R for testing $\boldsymbol{\beta} = \mathbf{0}$ is highly significant, so that, not surprisingly, at least one of the nine estimated coefficients is non-zero.

The bottom panel of Table III tests whether the auranofin coefficients differ significantly from the gold coefficients – composite linear hypothesis. We estimate that gold-treated patients had a 1·8 greater reduction in tender joint count than auranofin patients. Gold was also significantly better than auranofin in terms of grip strength and ESR.

Although the model of Table III produces firm conclusions concerning the relative efficacy of injectable gold and auranofin, the generalized residual sum of squares H_E is quite large, indicating that significant variation beyond sampling error remains. We next consider incorporation of covariates to account for more of the variation between studies, and to adjust for study or treatment variables that might introduce bias into our estimated coefficients.

Model with covariates and treatment variables

Felson *et al.*[13,15] examined study-level covariates for their association, both theoretical and empirical, with the outcomes under consideration. We use an *a priori* set of covariates found important in the larger data set. Five covariates available in all studies and having an expected association with the outcomes included the duration of the trial (range 12 to 52 weeks, mean 26·2 weeks in these 44 studies) and the mean disease duration for the treatment group (range 1 to 15 years, mean 5·7 years in these 63 treatment arms). For tender joint count only, initial tender joint count (range 12 to 40, mean 21·9 in these arms, with a theoretical range of 0 to 68 for an individual patient) and blindedness (66 per cent of the 44 studies blinded the evaluators to treatment allocation) had significant associations with outcome. There have also been quality scores (theoretical range from 0 to 20, observed range from 9 to 20, mean = 14) assigned to each study[15] which provide the fifth covariate. The quality scoring system, a modification of a conventional scoring instrument,[16] evaluated 10 different characteristics of trial reports (one of the 10 being the blindedness indicator), with attention focussed on completeness of reporting of the 10 items.

Detsky *et al.*[17] discuss four approaches for taking quality of trials (quality scores) into account in meta-analyses of RCTs. One approach[13] incorporates them into weights. Second and third approaches use a threshold quality score as an inclusion/exclusion criterion and sequentially combine trials based upon quality score. Detsky's fourth approach plots treatment effect versus quality score (or, similarly, includes quality score as a covariate in a regression model). Along these lines, we include quality score as a covariate in the GLS model. We centred each continuous covariate at its mean value. We coded blindedness 1 for blinded studies, 0 otherwise.

Prior to our analyses, we hypothesized that patients in longer trials would experience greater improvement. Also, we anticipated that patients with longer disease duration would be more resistant to improvements. On the other hand, patients with a higher tender joint count at baseline could experience greater improvement in tender joint count. Finally, we expected that studies with blinded evaluators and studies with higher quality scores would provide estimates of improvement in rheumatoid arthritis outcomes nearer to the truth, whatever the truth might be.

Table IV. GLS multiple-outcome model for comparing drug treatments, adjusting for covariates used by Felson *et al.* The covariates are quality score (QS-14·5) centred at the mean score of 14·5, years of disease prior to trial (centred at mean; disyr-5·7), weeks of the trial (centred; weeks-26), baseline tender joint count (centred; start-22), and blinding status of the trial (blinded: 0 = evaluators not blinded, 1 = blinded to treatment allocation). The estimated $\hat{\beta}$'s represent changes (before minus after treatment) in the particular outcome (joint, grip or ESR)

		$\hat{\beta}$	SE($\hat{\beta}$)	$\hat{\beta}$/SE($\hat{\beta}$)	p-value	
Placebo effects	joint	5·32	0·78	6·79	0·0000	
	grip	– 8·31	2·12	– 3·91	0·0000	
	ESR	1·17	1·06	1·10	0·2710	
Auranofin effects	joint	2·69	0·54	4·95	0·0000	auranofin over placebo effects
	grip	– 10·61	2·60	– 4·08	0·0000	
	ESR	8·57	1·19	7·20	0·0000	
Gold effects	joint	4·48	0·59	7·55	0·0000	gold over placebo effects
	grip	– 21·96	2·88	– 7·64	0·0000	
	ESR	19·16	1·32	14·49	0·0000	
QS-14·5	joint	0·42	0·11	3·96	0·0000	
	grip	– 0·81	0·29	– 2·76	0·0058	
	ESR	0·10	0·12	0·81	0·4179	
Disyr-5·7	joint	– 0·37	0·08	– 4·53	0·0000	
	grip	2·84	0·50	5·69	0·0000	
	ESR	– 0·77	0·16	– 4·90	0·0000	
Weeks-26	joint	0·19	0·07	2·82	0·0047	
	grip	0·10	0·27	0·37	0·7090	
	ESR	0·11	0·16	0·66	0·5108	
Start-22	joint	0·22	0·05	4·83	0·0000	
Blinded	joint	– 0·75	0·81	– 0·92	0·3562	

$H_E = 439{\cdot}1$ 137 d.f. $p \approx 0$
$H_R = 2316{\cdot}6$ 20 d.f. $p \approx 0$

Composite hypothesis test: gold (G) versus auranofin (A) $\hat{\beta}$'s

	G – A	SE (G – A)	p-value	
Joint	1·79	0·50	0·0004	if positive, gold better
Grip	– 11·35	2·49	0·0000	if negative, gold better
ESR	10·59	0·98	0·0000	if positive, gold better

Correlations among the five covariates ranged from – 0·26 (blindedness with years of disease duration, indicating that subjects in blinded trials had not been ill as long, on average) to 0·54 (blindedness with quality score, indicating that blinded studies tended to have better scores for quality of reporting). The length of the study was unrelated (r = – 0·05) to quality score.

Table IV summarizes the GLS model with treatment variables and these covariates as predictor variables. As expected, subjects in longer trials experienced greater average improvement in tender joint count. Subject groups with higher initial tender joint count also experienced greater improvement during the trials. (Further analyses suggested greater benefit for patients treated with gold than with auranofin, but practically nil for placebo patients.) Patients with longer average disease duration showed less improvement on all three outcomes. On study quality variables, studies with better scores for quality of reporting showed more improvement in both tender joint count and grip strength.

Comparing Table IV with Table III, we see that inclusion of the covariates quality score, years of disease duration, weeks of trial, baseline tender joint count and blindedness does affect the magnitude of the nine treatment coefficients, but, aside from the auranofin and gold effects on change in tender joints, they generally do not move by more than one standard error. In individual clinical trials, Beach and Meier[18] demonstrated that adjustment for covariates often reduces the precision of the main treatment effect. In our meta-analysis of many trials, the standard errors of our main effects are larger for the model with covariates (Table IV versus III). The full model (with these covariates) provided the same conclusions as Table III, concerning the superiority of gold over auranofin on all three outcomes, from the composite hypothesis tests.

Unfortunately, this model, with $H_E = 439{\cdot}1$ on 137 d.f. ($p < 0{\cdot}01$), still does not explain as much variation as we would wish. For the fixed-effects model, the residual variation, measured by H_E, should be comparable to the within-study variation. An examination of the residuals from the full model revealed three unusual residuals for tender joint count, three for grip strength, and three for ESR. These nine outliers arose from nine different studies. Eight of the nine residuals had corresponding observed measurements near the extremes of their distributions (roughly in the extreme 5 per cent). The ninth residual (for tender joint count in a trial that has a fairly typical mean among these patients) arose from the shortest trial (12 weeks), with the lowest quality score (9) and smallest treatment group (6 subjects). All nine outliers arose from very small treatment groups (n_{it} = 6, 8, 10, 11, 12, 12, 13, 13 and 16), and were checked to confirm that data extraction errors were not responsible.

Analyses assuming independent outcomes

As a basis for comparison with the common practice of analysing each outcome variable separately, we refitted the GLS model after setting the between-outcome correlations to zero. In this way we obtained separate analyses for joint, grip and ESR while taking into account (through weights) the substantial differences in precision among the results of the individual studies. Table V summarizes these estimates. Generally, ignoring the correlations between the outcomes led to moderate changes in a number of the coefficients and slight to moderate increases in their standard errors. The comparisons of gold versus auranofin were unaffected. The correlations assumed in fitting the Table IV model, however, tended to be small (shown in Table II).

Sensitivity analysis

To investigate further the influence of the between-outcome correlations on the estimated model, we refitted the model of Table IV using sets of alternative correlations obtained from patient-level data.[14,19,20] As our analysis focuses on before-minus-after changes in three outcomes, we searched those data for triplets of any change-variables that exhibited a range of correlations. For example, from the external patient-level data, we noted that before-to-after changes in the three measurements 'tender joint severity, 'patient assessment of disease severity', and 'physician assessment of disease severity' tended to have moderate correlation (+ 0·41, + 0·46, and + 0·59), We thus constructed a 6 × 6 correlation matrix (tender joint score before, tender joint score after, ...) and used it (for each t) as our CORR_{it} matrix for estimation of $\sum_{it}$, after changing the signs of the relevant correlations to match the positive or negative associations among joint, grip and ESR. The newly fitted model shows the effect of larger between-outcome correlations than those observed for joint, grip and ESR.

We also refitted the model with all between-outcome correlations (but not the before-with-after correlations within outcome) set to zero. Comparison of these two newly fitted models provides

Table V. GLS model, with between-outcome correlations set to zero, for comparing drug treatments, adjusting for covariates used by Felson *et al.* This analysis is equivalent to fitting three separate WLS models

		$\hat{\beta}$	SE($\hat{\beta}$)	$\hat{\beta}$/SE($\hat{\beta}$)	*p*-value	
Placebo effects	joint	4·67	0·83	5·64	0·0000	
	grip	− 6·96	2·22	− 3·14	0·0017	
	ESR	0·83	1·07	0·78	0·4377	
Auranofin effects	joint	2·92	0·55	5·32	0·0000	auranofin over placebo effects
	grip	− 11·90	2·71	− 4·39	0·0000	
	ESR	8·55	1·20	7·12	0·0000	
Gold effects	joint	4·65	0·60	7·79	0·0000	gold over placebo effects
	grip	− 22·18	2·97	− 7·46	0·0000	
	ESR	19·26	1·33	14·47	0·0000	
QS-14·5	joint	0·37	0·11	3·31	0·0009	
	grip	− 0·94	0·30	− 3·09	0·0020	
	ESR	0·15	0·12	1·20	0·2290	
Disyr-5·7	joint	− 0·34	0·08	− 4·08	0·0000	
	grip	3·18	0·53	6·00	0·0000	
	ESR	− 0·71	0·16	− 4·46	0·0000	
Weeks-26	joint	0·17	0·07	2·48	0·0133	
	grip	− 0·07	0·27	− 0·27	0·7862	
	ESR	− 0·08	0·17	− 0·45	0·6499	
Start-22	joint	0·25	0·05	5·19	0·0000	
Blinded	joint	− 0·10	0·87	− 0·11	0·9097	

Composite hypothesis test: gold (G) versus auranofin (A) $\hat{\beta}$'s

	G − A	SE (G − A)	*p*-value	
Joint	1·74	0·51	0·0006	if positive, gold better
Grip	− 10·28	2·55	0·0001	if negative, gold better
ESR	10·71	0·98	0·0000	if positive, gold better

further insight on the effect of ignoring moderately large between-outcome correlations, as if we were fitting three separate WLS models for joint, grip and ESR.

Similarly, we constructed other 6 × 6 correlation matrices, so that in all we used five 6 × 6 matrices. The largest between-outcome-change correlation in each of these 5 triplets was (ignoring signs) 0·24, 0·47, 0·59, 0·77 and 0·87 (see Table VI(*a*)). We re-estimated the GLS model of Table IV using each of these 6 × 6 correlation matrices and, as described above, for each 6 × 6 CORR_{it} matrix we subsequently set all between-outcome correlations to zero, leaving only the within-outcome correlations, to assess the impact of ignoring these between-outcome correlations (see Table VI(*b*)).

Table VI presents the results of this sensitivity analysis. We focus here on the main hypothesis of interest: whether gold is more effective than auranofin, as measured by tender joint count, grip strength and ESR. In Table VI(*a*), moving downward from slightly correlated change variables to highly correlated change variables, we note a steady increase in the size of the estimated treatment effect (gold minus auranofin) for tender joint count, but for grip strength the effect moves up and down, and for ESR the effect is nearly constant. The standard errors and associated *p*-values follow no obvious pattern.

Table VI. Sensitivity of GLS – MO model to various levels of correlation, between and within outcomes. The results of composite hypothesis tests comparing gold versus auranofin are given when the full model (Table IV) is estimated, with the correlations shown below substituted

	Between-outcome change correlations		Before with after correlations	Estimate of treatment effect (gold β-auranofin β)	SE	p-value
	joint	grip				
(a) Between-outcome change correlations are shown						
Joint			0·30	1·80	0·72	0·0122
Grip	– 0·16		0·77	– 11·03	2·53	0·0000
ESR	+ 0·21	– 0·24	0·61	10·74	1·09	0·0000
Joint			0·63	1·86	0·52	0·0003
Grip	– 0·47		0·58	– 10·80	3·25	0·0009
ESR	+ 0·41	– 0·28	0·30	10·77	1·44	0·0000
Joint			0·29	1·91	0·71	0·0071
Grip	– 0·59		0·30	– 11·11	4·00	0·0055
ESR	+ 0·46	– 0·41	0·63	10·74	1·06	0·0000
Joint			0·76	2·00	0·41	0·0000
Grip	– 0·77		0·77	– 10·79	2·22	0·0000
ESR	+ 0·34	– 0·33	0·66	10·76	1·02	0·0000
Joint			0·58	2·04	0·54	0·0001
Grip	– 0·87		0·66	– 10·39	2·52	0·0000
ESR	+ 0·56	– 0·54	0·57	10·71	1·13	0·0000
(b) Between-outcome change correlations are all set to zero						
Joint			0·30	1·72	0·72	0·0164
Grip			0·77	– 10·28	2·57	0·0001
ESR			0·61	10·71	1·09	0·0000
Joint			0·63	1·72	0·53	0·0011
Grip			0·58	– 10·26	3·44	0·0029
ESR			0·30	10·77	1·45	0·0000
Joint			0·29	1·72	0·72	0·0172
Grip			0·30	– 10·25	4·42	0·0202
ESR			0·63	10·70	1·07	0·0000
Joint			0·76	1·72	0·42	0·0000
Grip			0·77	– 10·28	2·56	0·0001
ESR			0·66	10·69	1·02	0·0000
Joint			0·58	1·72	0·56	0·0022
Grip			0·66	– 10·27	3·07	0·0008
ESR			0·57	10·72	1·15	0·0000

Because both between-outcome and within-outcome correlations change as we go down Table VI(*a*), we also examine Table VI(*b*), in which we set to zero all between-outcome-change correlations, so that only within-outcome (before with after) correlations differ as we move down the table. The treatment effect (gold minus auranofin) changes little as we move down Table VI(*b*), indicating minor influences on coefficients due to within-outcome correlations, whose impact is instead on the standard errors. Comparing panels (*a*) and (*b*) of Table VI indicates that ignoring the between-outcome correlations increases the standard errors only slightly (they are more sensitive to the within-outcome correlations), but the coefficients are more highly affected.

In summary, the sensitivity analysis, based on a diverse range of observed correlations, indicates that: (1) accounting for the between-outcome correlations may alter the estimated coefficients; but (2) the within-outcome correlations, rather than the between-outcome correlations, affect the size of the standard errors of the coefficients and have minimal impact on the coefficients themselves; (3) neither the clinical nor the statistical conclusions, however, were substantially affected by the assumed correlations. Therefore, we conclude that, in this particular application, obtaining external estimates of the correlation matrices CORR_{it} from a single database is acceptable.

Exclusion of single arms from comparative trials

Meta-analyses do not customarily include trials that contain only one of the treatments that they study. We consider it a strength that our multiple-outcomes model can include single treatment arms from randomized trials in a meta-analysis. This flexibility allows us to draw on the additional information available on those treatments, and thus it enhances power.

For comparative purposes, however, we omitted all those trials that provided only a single treatment group and we re-estimated the full model (Table VII). By comparing Tables VII and IV, we notice that the quality score of the trial is no longer important, and the coefficient for baseline tender joint count (start) is nearly doubled. The other difference is that duration of trial (weeks) becomes important for ESR in Table VII, but its negative coefficient is in the opposite direction from what we would expect. The discrepancy of greatest interest is that the difference between the effects of gold and auranofin somewhat diminish (when we delete trials that provide data on only one of these treatments), to the point that gold is no longer significantly better (1·17 joints, $p = 0{\cdot}11$) than auranofin in reducing the number of tender joints. One could argue that this is due to lower power in Table VII, because the standard error of the gold-minus-auranofin β for tender joint count is 50 per cent greater than in Table IV. The β in Table VII (1·17), however, is also 35 per cent smaller than that in Table IV.

To ascertain whether the gold-only and auranofin-only trials drove the Table IV results, we looked at unweighted means of improvement in tender joint count by type of trial. Trials of gold versus auranofin versus placebo reported gold better than auranofin by 2·45 joints, on average, whereas gold versus auranofin trials indicated gold better by only 0·52 joints. Because the gold-only trials provided means of only 0·81 joints better than auranofin-only trials, it does not appear that these trials provided the bulk of the evidence in favour of gold treatment for tender joints.

A further observation of possible relevance is that the set of trials in the Table IV analysis included some patients possibly more ill than the subset in Table VII. They had more tender joints at baseline (21·9 versus 20·9) and more years of disease prior to the trial (5·7 versus 4·8 years). We therefore conclude that inclusion of single arms from comparative trials in the Table IV analysis provides additional information on patients more ill, and because the standard errors in Table IV (relative to Table VII) are considerably smaller, also provides enhanced power for our finding of gold's superiority over auranofin in reducing tender joints.

We conducted a final GLS analysis restricted further to just those 9 trials that evaluated both gold and auranofin. Seven trials of gold versus auranofin contained 162 patients, whereas two trials of gold, auranofin and a placebo included 187 patients. In this model the dependent variable **d** is a vector of within-trial differences of the treatment effects ((gold before minus after) minus (auranofin before minus after)), with corresponding changes in the S and X matrices. None of the covariates considered earlier was significant (all $p > 0{\cdot}18$), so we present the fitted model that did not include the covariates (Table VIII). The estimated β's in this version of the model have the

Table VII. GLS multiple-outcome model on 17 trials, each providing data on at least two of the three treatment arms of interest to us. Covariates are centred at the means computed from these 17 trials

		$\hat{\beta}$	SE($\hat{\beta}$)	$\hat{\beta}$/SE($\hat{\beta}$)	p-value	
Placebo effects	joint	5·10	1·35	3·78	0·0002	
	grip	−11·13	2·27	−4·89	0·0000	
	ESR	1·35	1·10	1·23	0·2193	
Auranofin effects	joint	3·16	0·64	4·97	0·0000	auranofin over placebo effects
	grip	−13·12	2·79	−4·70	0·0000	
	ESR	8·27	1·28	6·49	0·0000	
Gold effects	joint	4·33	0·69	6·29	0·0000	gold over placebo effects
	grip	−20·46	3·37	−6·07	0·0000	
	ESR	14·62	1·57	9·34	0·0000	
QS-15	joint	0·19	0·18	1·03	0·3038	
	grip	−0·31	0·63	−0·50	0·6180	
	ESR	−0·28	0·31	−0·93	0·3540	
Disyr-4·8	joint	−0·71	0·18	−3·94	0·0001	
	grip	2·73	0·67	4·07	0·0000	
	ESR	−1·27	0·24	−5·21	0·0000	
Weeks-26·5	joint	0·34	0·09	3·68	0·0002	
	grip	−0·06	0·37	−0·17	0·8620	
	ESR	−0·74	0·32	−2·30	0·0213	
Start-21	joint	0·40	0·09	4·21	0·0000	
Blinded	joint	−0·12	1·28	−0·09	0·9624	

$H_E = 191·5$ 71 d.f. $p \approx 0$
$H_R = 1361·5$ 20 d.f. $p \approx 0$

Composite hypothesis test: gold (G) versus auranofin (A) $\hat{\beta}$'s

	G − A	SE (G − A)	p-value	
Joint	1·17	0·74	0·1129	if positive, gold better
Grip	−7·34	3·17	0·0207	if negative, gold better
ESR	6·35	1·37	0·0000	if positive, gold better

Table VIII. GLS multiple-outcome model for comparing drug treatments in 9 trials that compared gold and auranofin. Within-trial differences of gold and auranofin treatment effects are used here. No covariates were significant (all $p > 0·18$), so they were dropped from the model

		$\hat{\beta}$	SE($\hat{\beta}$)	$\hat{\beta}$/SE($\hat{\beta}$)	p-value
Gold versus auranofin effects	joint	0·82	1·21	0·67	0·5006
	grip	−4·93	4·08	−1·21	0·2271
	ESR	5·49	1·95	2·81	0·0049

$H_E = 31·2194$ 20 d.f. $p \approx 0·0524$
$H_R = 8·4188$ 3 d.f. $p \approx 0·0381$

same intended interpretation as the composite hypothesis tests at the bottom of Tables III, IV, V and VII. Relative to the earlier tables, the estimated treatment advantage of gold over auranofin (Table VIII) is smaller for all three outcomes, and the standard errors are larger in each instance, not surprising because this is much smaller set of trials. We find gold continues as significantly ($p = 0·005$) better than auranofin for the outcome ESR.

DISCUSSION

We have illustrated an application of GLS regression to the meta-analysis of multiple outcomes that appropriately takes into account correlations among the outcomes of interest. This method is suitable when treatment effects are given by two or more continuously measured variables recorded on each patient in randomized controlled trials. Many multiple-outcome situations are not addressed by this methodology; for example, trials that report the two outcomes 'the proportion of patients who experienced MI' and 'proportion of patients who died from MI'. Dear[4] considers the meta-analysis of survival proportions reported at multiple years.

One strength of this methodology is that it allows for adjustment of study-level and treatment-group-level covariates when evaluating treatment effectiveness, and, as our application illustrates, every study need not report on all outcomes (tender joint count, grip strength and ESR), nor must every study include all treatment groups (auranofin, injectable gold, placebo). The flexibility of this approach thus allows the meta-analysis to use more of the available data.

Standard approaches to meta-analysis ordinarily make all comparisons within trials, so as to free them from differences among trials. Thus, a placebo or common treatment must appear in all trials. When the available trials have varied subsets of treatments, restriction of the analysis to within-trial comparisons sacrifices much of the data. In such situations our GLS method considers each treatment group as an independent observation. The method, however, uses some context-specific information in the form of study- and treatment-level covariates. Thus, when the fixed-effects model is adequate, differences among trials are accounted for, and so the lack of within-trial comparisons should be less of an issue.

However, as we illustrated, we can also model a direct within-trial comparison that takes into account between-outcome correlations when all trials compare the same two treatments. In another example, our GLS meta-analysis of treatments for periodontal disease used paired within-patient comparisons of two treatments.[7] In the present context we focus on within-trial comparisons by defining $\mathbf{y}$ as the within-trial treatment effects (for example, within each trial, compute gold-minus-auranofin group means for each outcome) and having X contain study-level covariates, but no treatment variables or treatment-arm-level covariates. We have illustrated this here (Table VIII and Appendix, Example 2), but we could include only the 9 studies that evaluated both treatments (out of 44 studies). The gold and auranofin treatment groups from these 9 trials included just 298 (12 per cent) of the subjects who completed the 44 trials. The small size of this subset of the randomized trials raises concerns about bias. On the other hand, those who prefer to use only within-trial comparisons would be concerned about the possibility of bias in including single arms from other randomized trials. Meta-analysts should consider the tension between these two approaches. In the present example the subset of 9 trials had on average somewhat lower baseline tender joint counts (19·9 versus 21·9). The results obtained were also somewhat different. The apparent difference in effectiveness between gold and auranofin was approximately halved, compared with the result in the 44-trial analysis. To ascertain whether the gold versus auranofin treatment effect depends on the baseline joint count, which would explain the discrepancies between the two analyses, we fitted a model to the 44 trials that included a treatment by baseline-joint-count interaction. This model suggested a marginally significant ($p = 0{\cdot}089$) tendency for gold's benefit over auranofin to increase further as the baseline count rises. The magnitude of the association, however, was not sufficiently strong to explain fully the absence of a gold versus auranofin effect on tender joints in the subset of 9 trials (Table VIII).

One difficulty in implementing the general approach is the need for estimates of correlations among the multiple outcomes. When none of the published trials provides the needed information, the meta-analyst may be able to obtain it from one or more investigators who have the original patient-level data. Alternatively, one can use sensitivity analysis to see how different magnitudes of correlations affect the substantive conclusions. The sensitivity analysis in our application did not indicate that our fitted model was sensitive (enough to modify conclusions) to the specific correlations assumed in model fitting. Thus we would not have reached different overall conclusion, regarding rheumatoid arthritis therapy, had we obtained moderately different correlations from a different source. Follmann *et al.*[8] found, as we did, that point estimates were generally insensitive to assumptions about the before-with-after correlations of a single outcome.

The particular outcomes used in the current analysis are only moderately and not always consistently correlated. In fact, they were selected as outcomes in the larger meta-analysis[13] because they were commonly available representatives of three aspects of outcome in rheumatoid arthritis that are conceptually distinct and that empirically have only partial overlap.[5] In another application, however, GLS models for meta-analysis could be more sensitive to the specified correlations.

Studies included in a meta-analysis may use more than one version of an outcome measure, or they may include two closely related outcomes. In practice, meta-analysts may have relied on the existence of strong correlations between these versions to justify choosing just one of them. For example, in rheumatology, tender joint count and swollen joint count are related and strongly correlated. The method given here allows for inclusion of both related outcomes, taking their correlations into consideration. For example, we could add swollen joint count as a fourth outcome throughout our application.

Although, in this particular dataset, inclusion of covariates in the model did not affect the basic conclusions about competing arthritis therapies, several covariates were associated with improvements in rheumatoid arthritis symptoms. Briefly, treatment groups whose patients had more tender joints at baseline showed greater declines in tender joint count during the trials, and longer trials also yielded greater declines in tender joint count, but trials where patients had had rheumatoid arthritis for longer periods of time prior to the trial produced less improvement on all three outcomes. Of methodologic interest is the observation that quality score was correlated with two of the outcomes, so that better-documented studies reported more improvement in tender joint count and grip strength. This finding is unusual in that other studies have found little association between treatment effects and quality of RCTs,[21] though three recent studies observed an increase in treatment effect with increasing quality score (Imperiale and McCullough,[22] Berlin and Colditz,[23] Colditz *et al.*[24]) and a fourth observed a decrease in treatment effect with improving study quality (Eisenberg *et al.*[25]). Prior to 1990, there were larger associations found when looking across designs.[26] Our estimated treatment effect of gold versus auranofin, however, was similar, whether or not we adjusted for quality score (Tables III and IV). We observed greater differences in meta-analysis results from dropping studies rather than from covariates such as quality scores. Greenland[27] argues that a focus on the specific elements of a quality score can often be more helpful than use of overall quality scores, though one empirical study failed to find evidence to support this.[21]

In the full analysis of all relevant trials, gold was better than auranofin by 1·8 tender joints. Table IV implies that a patient with the usual 21·9 tender joints at baseline is expected to have 12·1 tender joints after treatment with injectable gold versus 13·9 tender joints after treatment with auranofin. A patient with a baseline grip strength of 127·7 mmHg is expected to improve to 157·9 mmHg if treated with gold, but only to 146·6 mmHg if treated with auranofin, and the patient with baseline ESR equal to 44·9 mm/hour expects a favourable decline to 24·5 after treatment with gold or to 35·1 after treatment with auranofin. For tender joint count, the magnitude of the placebo effect (5·3 tender joints according to Table IV, 6·0 by Table I) is greater than the additional effects of either experimental therapy (2·7 for auranofin or 4·5 for gold).

If we had a multiple-outcome random-effects GLS regression model for meta-analysis, we would find it informative to see how much further we could reduce the H_E statistic in our application. We would expect such a model to provide larger p-values and wider (but more realistic) confidence intervals, as in our single-outcome random-effects regression model for meta-analysis.[28]

Debate on whether or not to include single arms from comparative trials in a meta-analysis will continue. We justify it here on the basis of randomization to treatment groups, outcomes being within-patient change scores (so that, in a sense, patients provide their own controls), and the inclusion of study-level covariates to adjust for study differences. Users of our method can, with no modification in the methodology, either include or exclude trials that have only one of the treatments considered in the meta-analysis.

Our GLS multiple-outcomes regression model for meta-analysis provides another alternative to the usual approach of a separate meta-analysis for each of the outcomes reported by a group of clinical trials. We applied our model successfully to a meta-analysis of two therapies for rheumatoid arthritis evaluated in randomized trials, which provided us with the opportunity to consider many interesting methodologic and clinical issues. Although the individual trials reported no significant differences in the efficacy of gold and auranofin, the meta-analysis found both clinically significant and statistically significant differences between them on all three outcomes.

APPENDIX

Here we illustrate how to set up the data for analysis by the multiple-outcomes GLS regression model.

Example 1

Suppose there are three randomized controlled trials, denoted by subscripts 1, 2 and 3. We include two particular treatments (subscripts 1 = placebo and 2 = experimental) in the analysis. We denote two correlated outcomes with subscripts 1 and 2. Trial 2 has only one treatment group (no placebo), and Trial 3 reports only the first outcome. The model we fit considers a continuous covariate, say mean age of subjects in each treatment group, that we think modifies outcome 1, and a second covariate, say quality score (QS) of the trial, that we think modifies outcome 2. We assume here that both continuous covariates are centred. We denote the means reported for each outcome in each treatment group by each trial as $\bar{y}_{\text{trial},\text{trt},\text{outcome}}$ and arrange these means in a vector. For the X matrix, as shown below, we define the entries in the first column to be 1 for outcome 1 and 0 otherwise; entries in the second column of X are 1 for outcome 2 and 0 otherwise. Third-column entries are defined as 1 for treatment 2 *and* outcome 1, and the fourth column is 1 for treatment 2 *and* outcome 2 (0 otherwise). Column 5 equals the age for outcome 1 and 0 for outcome 2, and column 6 equals QS for outcome 2 (0 otherwise).

The β's for estimation using the X matrix shown below are:

β_1 estimates the placebo effect on outcome 1;
β_2 estimates the placebo effect on outcome 2;
β_3 estimates the experimental minus placebo treatment effect on outcome 1;

β_4 estimates the experimental minus placebo treatment effect on outcome 2;
β_5 estimates the effect of age on outcome 1;
β_6 estimates the effect of QS on outcome 2.

d vector	X matrix					
$\bar{y}_{111}$	1	0	0	0	age$_{11}$	0
$\bar{y}_{112}$	0	1	0	0	0	QS$_1$
$\bar{y}_{121}$	1	0	1	0	age$_{12}$	0
$\bar{y}_{122}$	0	1	0	1	0	QS$_1$
$\bar{y}_{221}$	1	0	1	0	age$_{22}$	0
$\bar{y}_{222}$	0	1	0	1	0	QS$_2$
$\bar{y}_{311}$	1	0	0	0	age$_{31}$	0
$\bar{y}_{321}$	1	0	1	0	age$_{32}$	0

$$S = \begin{bmatrix} S_{11} & & & & \\ & S_{12} & & & \\ & & S_{22} & & \\ & & & S_{31} & \\ & & & & S_{32} \end{bmatrix}.$$

S_{11}, S_{12} and S_{22} are 2×2, S_{31} and S_{32} are 1×1, and all other entries in S are zero. Each submatrix is specific to a treatment arm within a trial, and it reflects the covariances between outcomes in the same treatment group.

Example 2

Suppose there are two randomized controlled trials, denoted by subscripts 1 and 2. Both trials include two treatments, one of which may be placebo (this formulation of the model does not allow missing treatments). We denote two correlated outcomes by subscripts 1 and 2. Trial 2 reports only the second outcome. The model we fit includes a continuous covariate, the quality score (QS) of each trial, that we think modifies both outcomes, though perhaps differently. We here assume that QS is centred at the mean of the trials. We denote the outcome-specific within-trial difference between the two treatment means as $\bar{y}_{\text{trial, outcome}}$ and arrange these in a vector. We define the entries in the first column of X to be 1 for outcome 1 (0 otherwise), and second-column entries to be 1 for outcome 2 (0 otherwise). The third column equals the trial's quality score for outcome 1 (0 otherwise), and the fourth column equals QS for outcome 2 (0 otherwise).

The β's for estimation using the X matrix shown below are:

β_1 estimates the treatment difference on outcome 1;
β_2 estimates the treatment difference on outcome 2;
β_3 estimates the impact of QS on treatment difference for the first outcome;
β_4 estimates the impact of QS on treatment difference for the second outcome.

$$\begin{array}{l|cccc} \mathbf{d}\text{ vector} & & X\text{ matrix} & & \\ \bar{y}_{11} & 1 & 0 & QS_1 & 0 \\ \bar{y}_{12} & 0 & 1 & 0 & QS_1 \\ \bar{y}_{22} & 0 & 1 & 0 & QS_2 \end{array}$$

$$S = \begin{bmatrix} S_1 & \\ & S_2 \end{bmatrix}.$$

S_1 is 2×2, S_2 is 1×1, and all other entries in S are zero. Each submatrix of S is specific to a trial.

ACKNOWLEDGEMENTS

This research was supported by grant HS05936 from the Agency for Health Care Policy and Research to Harvard University and by NIH-NIAMS Multipurpose Arthritis Center Grant P60AR20613 to Boston University. The authors thank Frederick Mosteller, Graham Colditz, Jesse Berlin, Alexia Antczak-Bouckoms, the editor, and reviewers for their comments and suggestions at various stages of the work, David Felson, who made the data available to us, and Jeff Knight, who prepared the manuscript.

REFERENCES

1. Raudenbush, S. W., Becker, B. J. and Kalaian, H. 'Modeling multivariate effect sizes', *Psychological Bulletin*, **103**, 111–120 (1988).
2. Glass, G. V. 'Primary, secondary, and meta-analysis of research', *Educational Researcher*, **5**, 3–8 (1976).
3. Seber, G. A. F. *Linear Regression Analysis*, Wiley, New York, 1977.
4. Dear, K. B. G. 'Iterative generalized least squares for meta-analysis of survival data at multiple times', *Biometrics*, **50**, 989–1002 (1994).
5. Anderson, J. J., Felson, D. T., Meenan, R. F. and Williams, H. J. 'Which traditional measures should be used in rheumatoid arthritis clinical trials?', *Arthritis and Rheumatism*, **32**, 1093–1099 (1989).
6. Johnson, R. A. and Wichern, D. W. *Applied Multivariate Statistical Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1982.
7. Berkey, C. S., Antczak-Bouckoms, A., Hoaglin, D. C., Mosteller, F. and Pihlstrom, B. L. 'Multiple-outcome meta-analysis of treatments for periodontal disease', *Journal of Dental Research*, **74**, 1030–1039 (1995).
8. Follmann, D., Elliott, P., Suh, I. and Cutler, J. 'Variance imputation for overviews of clinical trials with continuous response', *Journal of Clinical Epidemiology*, **45**, 769–773 (1992).
9. Hedges, L. V. and Olkin, I. *Statistical Methods for Meta-Analysis*, Academic Press, New York, 1985.
10. Morrison, D. F. *Multivariate Statistical Methods*, McGraw-Hill, New York, 1976.
11. SAS Institute Inc. *SAS/IML User's Guide, Release* 6.03 *Edition*, SAS Institute Inc., Cary, NC, 1988.
12. *Dictionary of the Rheumatic Diseases, Vol. I: Signs and Symptoms*, NY Contact Associates Informational Ltd., New York, 1982.
13. Felson, D. T., Anderson, J. J. and Meenan, R. F. 'Use of short-term efficacy/toxicity tradeoffs to select second-line drugs in rheumatoid arthritis: A meta-analysis of published clinical trials', *Arthritis and Rheumatism*, **35**, 1117–1125 (1992).
14. Ward, J. R., Williams, H. J., Egger, M. J., Reading, J. C., Boyce, E., Altz-Smith, M., Samuelson, C. O. Jr., Willkens, R. F., Solsky, M. A., Hayes, S. P., Blocka, K. L., Weinstein, A., Meenan, R. F., Guttadauria, M., Kaplan, S. B. and Klippel, J. 'Comparison of auranofin, gold sodium thiomalate, and placebo in the treatment of rheumatoid arthritis: a controlled clinical trial', *Arthritis and Rheumatism*, **26**, 1303–1315 (1983).
15. Felson, D. T., Anderson, J. J. and Meenan, R. F. 'The comparative efficacy and toxicity of second-line drugs in rheumatoid arthritis', *Arthritis and Rheumatism*, **33**, 1449–1461 (1990).
16. DerSimonian, R., Charette, L. J., McPeek, B. and Mosteller, F. 'Reporting on methods in clinical trials', *New England Journal of Medicine*, **306**, 1332–1337 (1982).

17. Detsky, A. S., Naylor, C. D., O'Rourke, K., McGeer, A. J. and L'Abbé, K. A. 'Incorporating variations in the quality of individual randomized trials into meta-analysis', *Journal of Clinical Epidemiology*, **45**, 255–265 (1992).
18. Beach, M. L. and Meier, P. 'Choosing covariates in analysis of clinical trials', *Controlled Clinical Trials*, **10**, 161S–175S (1989).
19. Williams, H. J., Ward, J. R., Reading, J. C., Egger, M. J., Grandone, J. T., Samuelson, C. O., Jr., Furst, D. E., Sullivan, J. M., Watson, M. A., Guttadauria, M., Cathcart, E. S., Kaplan, S. B., Halla, J. T., Weinstein, A. and Plotz, P. H. 'Low-dose D-penicillamine therapy in rheumatoid arthritis: a controlled, double-blind clinical trial', *Arthritis and Rheumatism*, **26**, 581–592 (1983).
20. Williams, H. J., Willkens, R. F., Samuelson, C. O., Jr., Alarcón, G. S., Guttadauria, M., Yarboro, C., Polisson, R. P., Weiner, S. R., Luggen, M. E., Billingsley, L. M., Dahl, S. L., Egger, M. J., Reading, J. C. and Ward, J. R. 'Comparison of low-dose oral pulse methotrexate and placebo in the treatment of rheumatoid arthritis: a controlled clinical trial', *Arthritis and Rheumatism*, **28**, 721–730 (1985).
21. Emerson, J. D., Burdick, E., Hoaglin, D. C., Mosteller, F. and Chalmers, T. C. 'An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials', *Controlled Clinical Trials*, **11**, 339–352 (1990).
22. Imperiale, T. F. and McCullough, A. J. 'Do corticosteroids reduce mortality from alcoholic hepatitis?', *Annals of Internal Medicine*, **113**, 299–307 (1990).
23. Berlin, J. A. and Colditz, G. A. 'A meta-analysis of physical activity in the prevention of coronary heart disease', *American Journal of Epidemiology*, **132**, 612–628 (1990).
24. Colditz, G. A., Brewer, T. F., Berkey, C. S., Wilson, M. E., Burdick, E., Fineberg, H. V. and Mosteller, F. 'Efficacy of BCG vaccine in the prevention of tuberculosis: Meta-analysis of the published literature', *Journal of the American Medical Association*, **271**, 698–702 (1994).
25. Eisenberg, D. M., Delbanco, T. L., Berkey, C. S., Kaptchuk, T. J., Kupelnick, B., Kuhl, J. and Chalmers, T. C. 'Cognitive behavioral techniques for hypertension: Are they effective?', *Annals of Internal Medicine*, **118**, 964–972 (1993).
26. Miller, J. N., Colditz, G. A. and Mosteller, F. M. 'How study design affects outcomes in comparisons of therapy. II: Surgical', *Statistics in Medicine*, **8**, 455–466 (1989).
27. Greenland, S. 'Quality scores are useless and potentially misleading', *American Journal of Epidemiology*, **140**, 300–301 (1994).
28. Berkey, C. S., Hoaglin, D. C., Mosteller, F. and Colditz, G. A. 'A random-effects regression model for meta-analysis'. *Statistics in Medicine*, **14**, 395–411 (1995).