STATISTICS IN MEDICINE, VOL. 16, 1965–1982 (1997)

META-ANALYSIS FOR THE EVALUATION OF POTENTIAL SURROGATE MARKERS

MICHAEL J. DANIELS[1]* AND MICHAEL D. HUGHES[2,3]

[1] *Department of Statistics, Carnegie Mellon University, U.S.A.*
[2] *Department of Biostatistics, Harvard School of Public Health, U.S.A.*
[3] *Medical Statistics Unit, London School of Hygiene and Tropical Medicine, U.K.*

SUMMARY

We describe a meta-analysis approach for the evaluation of a potential surrogate marker. Surrogate markers are useful in helping to identify therapeutic mechanisms of action and disease pathogenesis, and for selecting therapies to take forward from phase II to phase III clinical trials. They have also become increasingly important for regulatory purposes by providing a basis for preliminary approval of drugs pending clinical outcome studies. Methodology for evaluating surrogate markers has focused on determining the difference in the effects of two treatments on clinical outcome in an individual clinical trial, and then estimating the proportion of this difference explained by the treatment's effects on the potential marker. Studies are, however, frequently underpowered or cease before they accumulate sufficient evidence to draw strong conclusions about the value of a potential surrogate marker using this approach, and there are also some technical difficulties with the approach. Consideration of the association between the difference in treatment effects on the clinical outcome and the difference in treatment effects on the potential marker over a range of trials provides an alternative means to evaluate a potential marker. We describe a meta-analysis approach using Bayesian methods to model this association. Importantly, this approach enables one to obtain prediction intervals for the true difference in clinical outcome for a given estimated treatment difference in the effect on the potential marker. We illustrate the methodology by applying it to results from studies of the AIDS Clinical Trials Group to assess the value of CD4 T-lymphocyte cell count as a potential surrogate marker for the treatment effects on the development of AIDS or death. © 1997 by John Wiley & Sons, Ltd.

Statist. Med., **16**, 1965–1982 (1997)
No. of Figures: 1 No. of Tables: 5 No. of References: 43

1. INTRODUCTION

In any clinical trial, the selection of an appropriate endpoint is an important issue. In some trials, however, the main endpoint of interest, for example death, is rare and/or takes a long period of time to reach. In such trials, there would be benefit in finding a more proximate endpoint to determine more quickly the effect of an intervention. These types of intermediate endpoints are often called *surrogate markers*. A surrogate marker (or surrogate endpoint) is defined as 'an endpoint measured in lieu of some other so-called true endpoint'.[1] Prentice[2] defined it more

* Correspondence to: Michael J. Daniels, Department of Statistics, Carnegie Mellon University, 232J Baker Hall, Pittsburgh, PA 15213, U.S.A.

Contract grant sponsor: Howard Hughes Medical Institute Predoctoral Fellowship, the National Institutes of Health
Contract grant number: AI 24643

CCC 0277–6715/97/171965–18$17.50

Received June 1996
Revised November 1996

formally as 'a response variable for which a test of the null hypothesis of no relationship to the treatment groups under comparison is also a valid test of the corresponding null hypothesis based on the true endpoint'. As well as providing information about possible mechanisms of action of drugs and of disease pathogenesis, one might use a good marker in phase I/II trials to help determine the value of further pursuit of therapies for clinical endpoint (phase III) studies. In addition, and particularly important at present, regulatory bodies might use a good marker to allow for preliminary approval ('accelerated' approval) of a drug on the basis of marker effects, pending completion of longer-term clinical outcome trials. If, however, an endpoint is not a good marker, misleading answers concerning interventions might result.

For convenience, we refer to the potential surrogate marker as the response variable and to the true endpoint as the clinical outcome. Prentice[2] proposed an approach to the validation of a response variable as a marker that involves three criteria: the variable must be prognostic of disease progression and be affected by treatment, and the effect of the treatment on the variable should mediate the effect of the treatment on clinical outcome. Typically, these criteria have been investigated with data from individual clinical trials. The usual approach is to fit a model that describes some measure of clinical outcome as the dependent variable with treatment group as a covariate, then to determine whether inclusion of the response variable as a covariate in the model fully explains the difference in the treatment's effects on the clinical outcome. If not fully explained, then one might evaluate the 'proportion of treatment effect explained' by the response variable.[3] For example, such models have been used to investigate the CD4 cell count as a potential surrogate marker in clinical trials that involve subjects with the human immunodeficiency virus (HIV).[4–6] There are, however, two major problems with this approach. First, the 'proportion' of treatment effect explained has been shown to be an erroneous concept when there are competing mechanisms of action, including some that adversely affect the clinical outcome, in that the proportion can then take values outside the range of zero to one.[7,8] Second, the proportion of the treatment effect explained is very imprecisely estimated so that one can draw only very weak conclusions unless the magnitude of the treatment difference on the clinical outcome is more than about four times its standard error.[3] This is problematic in that most trials are underpowered to give such evidence, and, even if they had sufficient power, they would likely be stopped before they accumulated such strong statistical evidence. Related to this, it is difficult to investigate surrogacy with this methodology in clinical trials in which the results might act as a counterexample, for example, in a trial with a precisely estimated treatment difference on clinical outcome close to zero but in which there is a large difference for the response variable.

Based on these issues, it is of interest to develop alternative methods for investigating the value of a response variable as a surrogate marker. In this paper, we propose one such approch that uses a meta-analysis of results from clinical trials to investigate the association between treatment differences on a response variable shown to be predictive of disease progression, and treatment differences on the clinical outcome.[9] More specifically, we aim to model this association and then to determine the model's reliability for predicting the treatment difference on clinical outcome given an observed difference on the response variable. Previously, A'Hern *et al.*[10] considered the idea of using a meta-analysis to assess the possibility of such an association in breast cancer clinical trials in which the response variable was the response of the tumour and the clinical outcome was survival, but they did not discuss the statistical problem in detail. In particular, although they allowed for the precision in estimating the treatment difference on clinical outcome, they did not take into account the level of precision in estimating the treatment difference on the response variable. This is an issue we address in this paper. Fleming[11] also pursued the idea of

a meta-analysis to look at the surrogate marker issue in both AIDS and cancer clinical trials. His approach, however, involved a qualitative review of trial results rather than a quantitative analysis as in this paper.

In Section 2, we propose a model for such a meta-analysis and describe how we might interpret parameters in the model in the context of investigating surrogacy. In Section 3, we present a Bayesian approach for estimating parameters in the model, for hypothesis testing and prediction. This develops the work of DuMouchel[1,2] that provided a Bayesian solution to the more traditional use of meta-analysis for pooling results of clinical trials to investigate the effect of a particular treatment. Then, in Section 4, we apply the methods to data from some HIV clinical trials to assess the value of changes in CD4 cell count as a surrogate marker. Finally, in Section 5, we discuss several issues and extensions relating to the model and its application.

2. MODEL

Consider a meta-analysis of N clinical trials each of which, for simplicity, involves the comparison of two treatments. The two treatments might be the same two treatments across all trials if we wish to investigate the surrogacy of a response variable for that specific pair of treatments. Alternatively, the two treatments may differ between trials if we wish to investigate surrogacy more generally, for example, for a specific class of treatments. The latter is likely of greater interest as the results obtained might be more generalizable to future clinical trials and treatments.

In the ith trial, denote the true treatment difference on the clinical outcome as θ_i and the true treatment difference on the response variable as γ_i. We assume that we have estimates $\hat{\theta}_i$ and $\hat{\gamma}_i$, respectively, available from each trial and that the size of each study is sufficiently large, that $(\hat{\theta}_i, \hat{\gamma}_i)$ is normally distributed about (θ_i, γ_i). For example, if the clinical outcome is a binary variable, then $\hat{\theta}_i$ might be the log odds ratio for the comparison of outcome rates for the two treatments, and if the response variable is continuous, then $\hat{\gamma}_i$ might be the difference in the mean marker values for subjects on one versus the other treatment. Thus, within the ith clinical trial, we assume that the following model applies:

$$\begin{pmatrix} \hat{\theta}_i \\ \hat{\gamma}_i \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \theta_i \\ \gamma_i \end{pmatrix}, \begin{pmatrix} \sigma_i^2 & \rho_i \sigma_i \delta_i \\ \rho_i \sigma_i \delta_i & \delta_i^2 \end{pmatrix} \right) \tag{1}$$

where σ_i^2 and δ_i^2 are variances that reflect sampling variation and ρ_i is the correlation between the estimated treatment differences conditional upon the true differences.

Considering the γ_i as fixed effects, we assume a simple linear model to describe the relationship between θ_i and γ_i across the N clinical trials (though the approach generalizes readily to more complex models):

$$\theta_i | \gamma_i \sim \mathrm{N}(\alpha + \beta\gamma_i, \tau^2). \tag{2}$$

Then, by combining the within-trial distribution given by (1) with the between-trial distribution given by (2), we obtain a bivariate normal distribution:

$$\begin{pmatrix} \hat{\theta}_i \\ \hat{\gamma}_i \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \alpha + \beta\gamma_i \\ \gamma_i \end{pmatrix}, \begin{pmatrix} \sigma_i^2 + \tau^2 & \rho_i \sigma_i \delta_i \\ \rho_i \sigma_i \delta_i & \delta_i^2 \end{pmatrix} \right).$$

In this model, β measures the association between the treatment differences on the response variable and the clinical outcome so that $\beta = 0$ corresponds to the case where the response variable is not, in fact, a surrogate marker, since knowing γ_i then does not help in predicting θ_i. In addition, if $\beta \neq 0$, having $\tau^2 = 0$ would imply that we could predict θ_i perfectly given γ_i. In fitting this model, as it is of interest to show that $\beta \neq 0$, then a necessary condition is that there is heterogeneity in the θ_i's. We can assess this by fitting a Bayesian equivalent of the usual random effects meta-analysis model[12,13] whereby $\hat{\theta}_i \sim \mathrm{N}(\theta_i, \sigma_i^2)$ and $\theta_i \sim \mathrm{N}(\alpha_\theta, \tau_\theta^2)$ and showing that $\tau_\theta^2 \neq 0$.

It is also useful to consider the role of the parameter α in this model. We might expect that α is zero for a response variable that is a good surrogate marker so that having no treatment difference on the marker (that is, $\gamma_i = 0$), implies no treatment difference on the clinical outcome (that is, $\theta_i = 0$). This is then consistent with Prentice's definition for a surrogate marker[2] given in Section 1. In addition, for the comparison of any pair of treatments A and B, it would not then matter whether we expressed the difference as A − B or B − A. Having α non-zero is more difficult to justify as it implies that there is, on average, a treatment difference on the clinical outcome unexplained by a treatment difference on the response variable. This implies that there is some aspect of treatment mediated through a mechanism other than that associated with the response variable. Thus, investigation of the possibility that α is non-zero is a useful step in assessing surrogacy as well as the investigation of β and τ^2. If the meta-analysis involves comparisons from a class of drugs, this investigation requires some rationale for how one orders treatments in each pairwise comparison; as most major trials involve a control or standard treatment, the natural order is to express differences as the outcome for the new treatment minus that for the control treatment.

It might also be possible to consider the γ_i as random effects, sampled from some distribution that we need to define, rather than as fixed effects. This increases the precision of the estimates for β and τ^2 if we specify the distribution correctly. The choice of the distribution, however, might be problematic. For instance, the ordering of the treatment within each comparison is important even in the constrained model with $\alpha = 0$; for example, if we chose treatment pairs such that all the $\hat{\gamma}_i$ were positive, then this would affect the distribution on the γ_i. In addition, if the meta-analysis involved comparisons of, say, two test treatments, each compared to the same standard treatment, and if each of the test treatments produced a different effect on the response variable, then we would anticipate a bimodal distribution for the γ_i's. This would complicate the model, making it more sensitive to model misspecification, particularly if the number of treatment comparisons is small.

3. MODEL-FITTING, TESTING, AND PREDICTION

Taking a Bayesian approach, we place 'non-informative' prior distributions on the fixed effects, γ_i and the regression coefficients:

$$\alpha \sim \mathrm{N}(0, A_\alpha)$$

$$\beta \sim \mathrm{N}(0, A_\beta)$$

$$\gamma_i \sim \mathrm{N}(0, A_{\gamma_i})$$

with each of A_α, A_β and A_{γ_i} large (see example for specific values).

For the between-studies variance, τ^2, it is less straightforward to define a 'non-informative' prior and so we consider three possibilities used previously:

1. Prior I (DuMouchel prior[12]): $\pi(\tau^2) = \frac{\sigma_c}{(\sigma_c + \tau)^2}\frac{1}{2\tau}$ where σ_c^2 is the harmonic mean of the within-study variances of the treatment difference on the clinical outcome, σ_i^2.
2. Prior II (shrinkage prior[14]): $\pi(\tau^2) = \frac{\sigma_c^2}{(\sigma_c^2 + \tau^2)^2}$, with σ_c^2 as above.
3. Prior III (flat prior[15]): $\pi(\tau^2) = \mathrm{d}\tau^2$

Priors I and II allow for the possibility that $\tau^2 = 0$. In terms of the relative behaviour of the three priors, prior I tends to produce a posterior distribution for τ^2 which is closer to zero than that for prior III, with prior II tending to produce an intermediate distribution. It is useful to assess the sensitivity of the results to the choice of prior; we illustrate this in the example.

For the covariance matrix in (1), we follow DuMouchel[12] by proceeding as if σ_i^2 and δ_i^2 are known and we replace them with their estimates in drawing inferences; thus, our approach is an empirical Bayes approach. The correlation ρ_i presents additional complexity, particularly as clinical trial publications never report $\hat{\rho}_i$. If the patient-level data are available, we can estimate ρ_i using the bootstrap technique[16] thus avoiding the need to model explicitly the joint relationship between the response variable and clinical outcome (a model that may be difficult to specify). We describe the use of the bootstrap in more detail in our example in Section 4. If the patient-level data are unavailable for some or all clinical trials, then the situation is more complex. In the Appendix, we show using a simulation study that, in practice, the value of ρ_i is likely small in magnitude and is unlikely to vary much between studies. Thus one approach is to assume that the ρ_i's for the trials with no data available are equal to some common value ρ. We could set the value of ρ to the average of the ρ_i's from trials for which patient-level data are available, or some arbitrary value reasonably close to zero in magnitude (for example, up to about 0·2 in magnitude; see the Appendix for more details). We could then evaluate the sensitivity of the conclusions by assessing the impact of changes in the choice of ρ on the resulting model estimates.

To fit the model, we can use Markov chain Monte Carlo (MCMC) techniques,[17] specifically the Gibbs sampler. The Gibbs sampler involves sequentially sampling from the distribution of each parameter, conditional on the current values of the other parameters and the data (known as the full conditional distribution). The full conditional distributions for all parameters in the model are Normal distributions, except for the full conditional distribution for τ^2, distributed as inverse gamma for prior III and as a non-standard distribution for priors I and II. To sample from the full conditional distribution for τ^2 for models using priors I and II, we use the Metropolis algorithm.[18] The Metropolis algorithm is a simple way to sample from any distribution known only up to its normalizing constant.

To test for an association between treatment differences on the response variable and treatment differences on the clinical outcome, we can compute Bayes factors[19] for $H_0: \beta = 0$; as the model with $\beta = 0$ is nested within the full model, we can do this using the Savage Dickey density ratio.[20] To compute the Bayes factor, however, we need to have a proper prior for β. We use a normal prior on the regression coefficients with mean zero. For the variance, we use the number of treatment comparisons included in the meta-analysis multiplied by the covariance matrix of the parameter estimates from the weighted least squares regression (assuming γ_i known). This is a simple, intuitive prior with covariance matrix equal to about one treatment comparison, similar to the unit information priors discussed in Kass and Wasserman.[21] We can use a similar

approach for testing $H_0: \alpha = 0$. For testing $\tau^2 = 0$, we can also employ Bayes factors for the two proper priors, I and II.

Having fit the model, it is of interest to predict the value of a future clinical outcome θ^* given the observed value of the response variable, $\hat{\gamma}^*$. To do this, note that conditional on regression coefficients, τ^2, and the data:

$$f(\theta^* | \hat{\gamma}^*, \delta^{*2}, \alpha, \beta, \tau^2, \hat{\theta}, \hat{\gamma}) \sim \mathrm{N}((\alpha + \beta\hat{\gamma}^*), \tau^2 + \beta^2\delta^{*2}).$$

The predictive distribution of interest, however, is for $\theta^* | \hat{\gamma}^*, \delta^{*2}, \hat{\theta}, \hat{\gamma}$; that is, we want the distribution of the true treatment difference on the clinical outcome given the observed treatment difference on the response variable and its variance. We can compute this distribution by using the output from the MCMC run:[22]

$$f(\theta^* | \hat{\gamma}^*, \delta^{*2}, \hat{\theta}, \hat{\gamma}) = \sum_{m=1}^{M} f(\theta^* | \hat{\gamma}^*, \delta^{*2}, \hat{\theta}, \hat{\gamma}, \alpha_{(m)}, \beta_{(m)}, \tau^2_{(m)}) \tag{3}$$

where m indexes the sample (of size M) from the joint posterior distribution of (α, β, τ^2) (these are merely the values obtained from running the Gibbs sampler). In these expressions, we replace δ^{*2} by $\hat{\delta}^{*2}$. We wrote FORTRAN programs to fit the model and compute predictive distributions, but we could also fit these models using the upcoming edition of the software for Bayesian analysis, BUGS, which allows for multivariate normal distributions.[23]

4. EXAMPLE: CD4 CELL COUNT AS A POTENTIAL SURROGATE MARKER

As an example of the application of these methods, we explore the association between treatment differences on the development of AIDS or death (a composite endpoint as the clinical outcome) and treatment differences in CD4 cell count (the response variable) in HIV clinical trials. The use of CD4 cell count as a potential surrogate marker is supported by the fact that infection with the HIV results in a gradual decline in immune function of which a fundamental component is the decline in the number of CD4 cells in the blood. As the CD4 cell count declines, there is increased risk of various opportunistic infections and malignancies which constitute the diagnosis of AIDS, and of death; thus, CD4 cell count is prognostic of progression to AIDS or death and so merits consideration as a potential surrogate marker.

The example uses data from the 15 phase II/III randomized clinical trials of the HIV Disease Section of the Adult AIDS Clinical Trials Group of the National Institutes of Health, which had data available as of May 1996, which had at least six months of follow-up on some patients and in which at least one patient developed AIDS or died. We excluded two other trials carried out by the Group as they compared different strategies of alternating drug therapy rather than continuous therapy with the same drug(s). All but two of the trials assessed the effects of a particular class of anti-HIV drugs known as nucleoside analogues. This class includes zidovudine (ZDV, also known as AZT), zalcitabine (ddC), and didanosine (ddI). The other two trials included newer types of drug, nevirapine (NVP) and saquinavir (SQV). A complicating feature of this meta-analysis arises from the fact that seven of the 15 trials had three treatment arms and one had four treatment arms. The trials with three (four) arms give rise to three (six) pairwise comparisons. However, given two (three), the remaining comparison(s) is determined so it is not possible to include all in an analysis. We have chosen to include in our analysis comparisons that involve the standard treatment as one of the treatments. It makes little difference, however, to the results of the analysis as to which comparisons one drops.

Table I. Treatment differences for the log hazard ratio for the development of AIDS or death over 2 years ($\hat{\theta}_i$) and the difference in mean change in CD4 cell count between baseline and 6 months ($\hat{\gamma}_i$) for studies of the AIDS Clinical Trial Group ($\hat{\sigma}_i$, $\hat{\delta}_i$, $\hat{\rho}_i$ are estimates of the standard error of $\hat{\theta}_i$, the standard error of $\hat{\gamma}_i$, and the correlation between $\hat{\theta}_i$ and $\hat{\gamma}_i$)

Study	Reference	Test treatment*	Standard treatment*	$\hat{\theta}_i(\hat{\sigma}_i)$	$\hat{\gamma}_i(\hat{\delta}_i)$	$\hat{\rho}_i$
002	31	ZDV[600]	ZDV[1500]	0·048 (0·092)	−9·2 (9·0)	−0·14
016	32	ZDV[1200]	placebo	−1·035 (0·370)	56·0 (11·8)	−0·02
019a	33	ZDV[1500]	placebo	−0·235 (0·282)	28·8 (11·0)	−0·13
		ZDV[500]	placebo	−0·594 (0·307)	46·1 (10·7)	−0·15
019b	34	ZDV[1500]	placebo	−1·313 (0·651)	67·1 (16·8)	0·01
		ZDV[500]	placebo	−0·359 (0·465)	37·2 (16·3)	−0·00
036	35	ZDV[1500]	placebo	−0·598 (0·707)	32·2 (18·0)	−0·06
112	†	ddC[‡]	ZDV[‡]	−0·447 (0·732)	−4·7 (6·1)	0·17
114	†	ddC[2·25]	ZDV[600]	0·267 (0·121)	−9·1 (5·6)	−0·22
116a	36	ddI[750]	ZDV[600]	0·096 (0·156)	11·8 (8·4)	−0·15
		ddI[500]	ZDV[600]	−0·022 (0·161)	12·8 (8·6)	−0·19
116b	37	ddI[750]	ZDV[600]	0·180 (0·130)	15·9 (5·3)	−0·07
		ddI[500]	ZDV[600]	−0·355 (0·137)	22·2 (5·4)	−0·11
118	38	ddI[200]	ddI[750]	0·112 (0·121)	−8·9 (5·8)	−0·06
		ddI[500]	ddI[750]	0·166 (0·120)	−5·5 (5·8)	−0·05
119	39	ddC[2·25]	ZDV[600]	−0·035 (0·340)	12·8 (9·5)	−0·08
155	40	ZDV/ddC[600/2·25]	ZDV[600]	−0·102 (0·121)	27·5 (4·2)	−0·09
		ddC[2·25]	ZDV[600]	0·083 (0·129)	17·1 (4·5)	−0·10
175	41	ZDV/ddC[600/2·25]	ZDV[600]	−0·348 (0·202)	36·1 (6·5)	−0·13
		ZDV/ddI[600/400]	ZDV[600]	−0·467 (0·207)	71·2 (6·4)	−0·17
		ddI[400]	ZDV[600]	−0·487 (0·207)	40·9 (6·4)	−0·19
229	42	ZDV/SQV[600/1800]	ZDV/ddC[600/2·25]	0·148 (0·518)	7·3 (10·2)	−0·13
		ZDV/ddC/SQV[600/2·25/1800]	ZDV/ddC[600/2·25]	−0·841 (0·680)	15·9 (10·2)	−0·16
241	43	ZDV/ddI/NVP[600/400/400]	ZDV/ddI[600/400]	0·211 (0·258)	25·8 (7·3)	−0·17

* Figures in brackets are the total daily dose (mg), ZDV = zidovudine, ddI = didanosine, ddC = zalcitabine, NVP = nevirapine, SQV = saquinavir
† Not published
‡ The ddC dose was weight dependent and the ZDV dose depended on the dose of ZDV being received prior to study entry

We estimate the difference in the effects of a pair of treatments on the response variable, γ_i, using the mean changes in CD4 count from the pre-treatment baseline assessment to the assessment at about six months. We chose six months as this is the duration of many phase II HIV clinical trials, and, as noted earlier, important uses of surrogate markers include the selection of drugs for further study and preliminary approval from regulatory bodies of a drug based on marker effects over such a period. We estimated the difference in the effects on the clinical outcome, θ_i, using the log hazard ratio obtained from a proportional hazards model for the time to development of AIDS or death, whichever came first, within 24 months of starting treatment. We chose 24 months as these studies typically had little follow-up beyond this, and, in clinical practice, many patients will not remain on the same treatment for periods much longer than this. In calculating the log hazard ratio and the difference in mean changes in CD4 cell counts, we always compared the new treatment to the standard treatment.

Table I lists the 15 clinical trials, the drugs (treatments) involved and the estimates of treatment difference for both the response variable and the clinical outcome. Note that two trials (studies

002 and 118) involved the comparison of multiple doses of the same drug (ZDV in study 002, ddI in study 118). In both trials, the objective was to determine whether lowering the dose might give reduced toxicities without materially affecting any benefit in delay of disease progression. Thus we consider the high dose as the standard treatment.

In the analysis, we need to make allowance for the correlation between the estimates for the treatment differences obtained from trials with three or four treatment arms. We can do this by expanding the first stage of the model (for three treatment arms) to:

$$\begin{pmatrix} \hat{\theta}_{i1} \\ \hat{\theta}_{i2} \\ \hat{\gamma}_{i1} \\ \hat{\gamma}_{i2} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \theta_{i1} \\ \theta_{i2} \\ \gamma_{i1} \\ \gamma_{i2} \end{pmatrix}, \begin{pmatrix} \sigma_{i1}^2 & \rho_{i\theta}\sigma_{i1}\sigma_{i2} & \rho_{i11}\sigma_{i1}\delta_{i1} & \rho_{i12}\sigma_{i1}\delta_{i2} \\ \rho_{i\theta}\sigma_{i1}\sigma_{i2} & \sigma_{i2}^2 & \rho_{i21}\sigma_{i2}\delta_{i1} & \rho_{i22}\sigma_{i2}\delta_{i2} \\ \rho_{i22}\delta_{i1}\sigma_{i1} & \rho_{i21}\delta_{i1}\sigma_{i2} & \delta_{i1}^2 & \rho_{i\gamma}\delta_{i1}\delta_{i2} \\ \rho_{i12}\delta_{i2}\sigma_{i1} & \rho_{i11}\delta_{i2}\sigma_{i2} & \rho_{i\gamma}\delta_{i1}\delta_{i2} & \delta_{i2}^2 \end{pmatrix} \right).$$

We readily obtain the correlation between treatment comparisons within a study, $\rho_{i\theta}$ and $\rho_{i\gamma}$, from proportional hazards models or linear regression models, respectively; as for the variances $\sigma_{i1}^2, \sigma_{i2}^2, \delta_{i1}^2, \delta_{i2}^2$, we assumed these known in the analysis. A similar extension applies for the four treatment arm study.

To estimate the correlation between the estimators of the treatment difference on the log hazard ratio and on change in CD4 count, we employed the non-parametric bootstrap.[16] Within each study, we took a random sample (with replacement) of patients within each treatment group in that study with the size of the sample being equal to the size of the treatment group within that study. We did this 1000 times (which gives adequate precision in estimating the correlation[16]), each time recomputing the log hazard ratio, $\hat{\theta}_i$, and the difference in mean change in CD4 cell count, $\hat{\gamma}_i$, for a pair of treatments, i. We then used the correlation of $\hat{\gamma}_i$ and $\hat{\theta}_i$ over the 1000 replicates as the estimate for ρ_i. One could also use this bootstrap technique to give estimates of the variances, σ_i^2 and δ_i^2; doing this gave almost identical estimates to those obtained directly from the proportional hazards model.

Figure 1 shows a plot of the association between the log hazard ratio of developing AIDS or dying versus the difference in the mean change in CD4 cell count for the placebo-controlled comparisons (labelled 'p') and the comparisons with active controls (labelled 'a'). Also shown are the 95 per cent confidence regions for the difference in clinical outcome and the difference in the response variable. The large size of the confidence regions for many of the comparisons shows the lack of precision in estimating these differences. The plot also suggests that a linear association between differences on the clinical outcome and differences on the response variable is reasonable.

Table II shows the results from four models fitted to evaluate the association suggested in Figure 1. Medians of the posterior distributions are presented together with the 95 per cent credible intervals (the limits of the interval are the 2·5th and 97·5th percentiles of the posterior distribution). For the regression parameters, α and β, the posterior means were similar to the medians reflecting the approximate symmetry of the posterior distributions for these parameters. In model A, we fit a linear association between θ_i and γ_i, as in equation (2). Model B is similar to model A except that we constrained α to be zero. Models C and D parallel models A and B, respectively, except that they allow α and β to differ between the placebo- and active-controlled treatment comparisons (as labelled by the subscripts p and a) thus allowing investigation of whether the type of control affects the association found. For each model, we show results for the three different priors for τ^2, described in Section 2, so that we can evaluate the impact of the choice of prior. For the priors on the regression coefficients, α and β, and on the true treatment

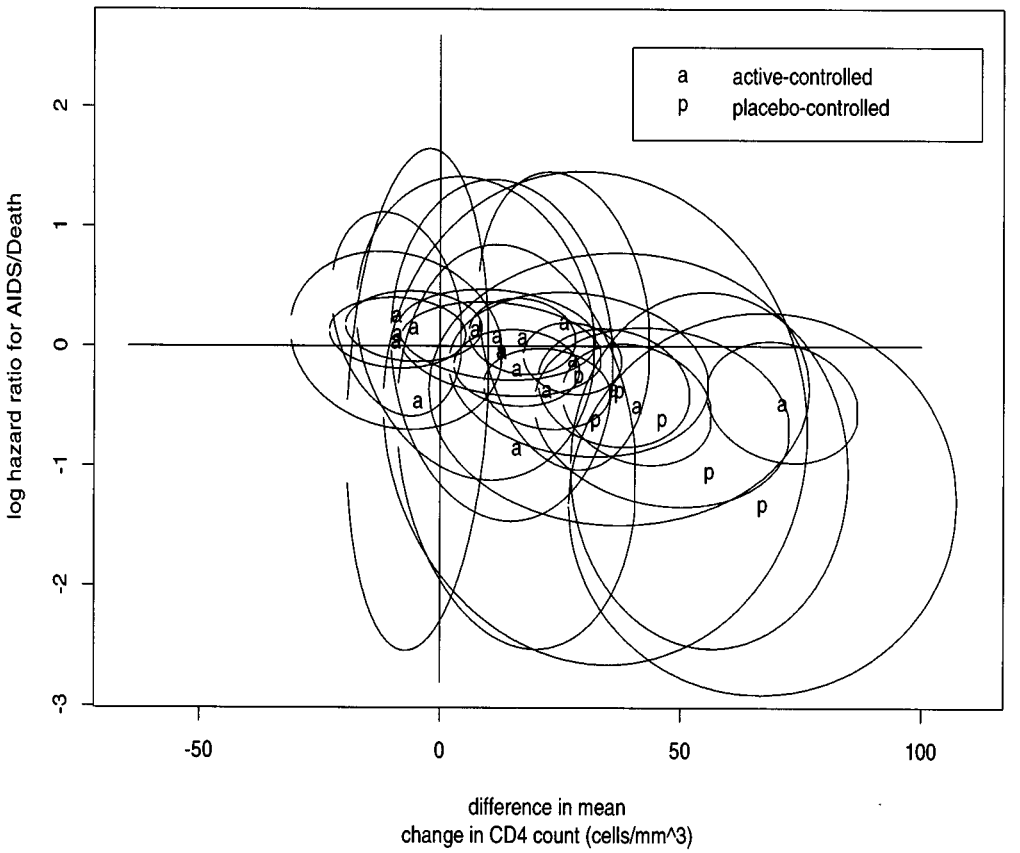

Figure 1. Association between the log hazard ratio for the development of AIDS or dying during 24 months and the difference in mean change in CD4 cell count over 6 months for all trials

differences on the mean change in CD4 cell count, the γ_i's, we chose the variances, A_α, A_β, and $A_{\gamma i}$, respectively, to be $1{\cdot}0e + 08$, which are essentially flat over the region of plausible values.

Before contrasting the different models, it is notable that for all four models there is very strong evidence of an association as shown by the negative values for β with credible intervals that exclude zero. Kass and Raftery suggest the following benchmarks for evaluating Bayes factors: 1–3, not worth more than a bare mention; 3–10, substantial evidence; 10–100, strong evidence; >100, decisive evidence.[19] For models A, B and D, the evidence against $\beta = 0$ (or $\beta_p = \beta_a = 0$) was decisive with Bayes factors exceeding 500 while the evidence was strong for model C, with Bayes factors exceeding 50. In addition, across the four models, the medians and credible intervals for τ^2 are very similar for any particular prior so that conclusions concerning τ^2 seem relatively insensitive to any differences that might exist in the association between the placebo- and active-controlled studies or to the possibility of non-zero intercepts.

Inspection of model A and of model C shows credible intervals for the intercepts that include zero. The Bayes factors for comparing model B to model A were slightly larger than 1 in favour of model B, that is, $\alpha = 0$, but did not provide conclusive evidence about this choice. The Bayes

Table II. Estimation of regression parameters, α and β and the variance, τ^2 for four models to describe the association between the log hazard ratio for AIDS or death and the difference in mean change in CD4 cell count, for various priors for τ^2 (medians and 95 per cent credible intervals reported)

Model	Parameters	Prior I*	Prior II*	Prior III*
A	α	0·072 [−0·030, 0·181]	0·072 [0·038, 0·189]	0·071 [−0·045, 0·194]
	β	−0·010 [−0·014, −0·006]	−0·010 [−0·015, −0·006]	−0·010 [−0·15, −0·005]
	τ^2	0·0009 [0·0000, 0·0161]	0·0040 [0·0002, 0·0268]	0·0063 [0·0002, 0·0432]
B	α	0	0	0
	β	−0·008 [−0·012, −0·005]	−0·009 [−0·012, −0·005]	−0·009 [−0·013, −0·005]
	τ^2	0·0012 [0·0000, 0·0172]	0·0047 [0·0001, 0·0266]	0·0070 [0·0002, 0·0428]
C	α_a	0·070 [−0·031, 0·177]	0·070 [−0·044, 0·186]	0·069 [−0·54, 0·189]
	β_a	−0·008 [−0·013, −0·004]	−0·008 [−0·013, −0·003]	−0·008 [−0·013, −0·003]
	α_p	0·226 [−0·747, 1·76]	0·213 [−0·748, 1·76]	0·202 [−0·763, 1·75]
	β_p	−0·019 [−0·052, 0·002]	−0·019 [−0·052, 0·002]	−0·019 [−0·052, 0·003]
	τ^2	0·0011 [0·0000, 0·0168]	0·0042 [0·0002, 0·0263]	0·0063 [0·0003, 0·0431]
D	α_a	0	0	0
	β_a	−0·007 [−0·011, −0·003]	−0·007 [−0·011, −0·003]	−0·007 [−0·011, −0·002]
	α_p	0	0	0
	β_p	−0·015 [−0·025, −0·007]	−0·015 [−0·025, −0·007]	−0·015 [−0·025, −0·007]
	τ^2	0·0011 [0·0000, 0·0167]	0·0041 [0·0002, 0·0255]	0·0061 [0·0003, 0·0394]

* Prior I: $\pi(\tau^2) = \dfrac{\sigma_c}{(\sigma_c + \tau)^2}\dfrac{1}{2\tau}$; prior II: $\pi(\tau^2) = \dfrac{\sigma_c^2}{(\sigma_c^2 + \tau^2)^2}$; prior III: $\pi(\tau^2) = \mathrm{d}\tau^2$

factors for comparing model D to model C were greater than 17, providing strong evidence in favour of zero intercepts. Consideration of the overlap in the credible intervals for β_p and β_a in models C and D, and of the credible intervals for α_p and α_a in model D, suggests little evidence for differential associations for the placebo-controlled treatment differences from the active-controlled treatment differences. Bayes factors for jointly testing $\alpha_p = \alpha_a$ and $\beta_p = \beta_a$ were all larger than 12, again providing strong evidence in favour of a common intercept and slope. Hence comparison of the four models, by using Bayes factors and examining credible intervals, does not suggest any strong evidence against the use of the simplest model, that is, model B, to describe the association seen in Figure 1.

Focusing on model B, consider the posterior distribution for τ^2. Of key interest is the level of evidence for $\tau^2 = 0$, that is, whether, given the true difference in mean change in CD4 cell count, we might establish the true hazard ratio for developing AIDS or dying. It is only possible to compute Bayes factors to assess this when using proper priors for τ^2, that is, priors I and II. The Bayes factors for testing $H_0: \tau^2 = 0$ was greater than 3 for both these priors, indicating positive, but not strong, evidence for τ^2 being zero. We obtained similar results when we assessed this hypothesis using one of the other models. We return to the relevance of the differences between the median values and credible intervals for τ^2 for the three priors below when we consider the use of the model to predict treatment differences in clinical outcome.

Now, we consider how one might check the predictive value of the model. To do this, we fit the model with one of the comparisons omitted and then use the model to predict the log hazard of developing AIDS or dying for the omitted comparison given the observed difference in mean

Table III. Results of the analysis to assess the predictive fit of model B with prior II

Trial	Comparison excluded				θ* for comparison excluded			
	Test treatment	Standard treatment	β	τ^2	Observed ($\hat{\theta}$*)	Predicted median	Predicted 95% interval	Z-score*
002	ZDV[600]	ZDV[1500]	−0·009	0·0054	0·05	0·08	[−0·22, 0·39]	0·19
016	ZDV[1200]	placebo	−0·008	0·0043	−1·04	−0·44	[−1·23, 0·34]	1·47
019a	ZDV[1500]	placebo	−0·008	0·0047	−0·24	−0·24	[−0·88, 0·37]	−0·02
	ZDV[500]	placebo	−0·008	0·0046	−0·59	−0·37	[−1·05, 0·29]	0·66
019b	ZDV[1500]	placebo	−0·008	0·0044	−1·31	−0·55	[−1·81, 0·79]	1·15
	ZDV[500]	placebo	−0·009	0·0045	−0·36	−0·31	[−1·30, 0·66]	0·09
036	ZDV[1500]	placebo	−0·008	0·0045	−0·60	−0·24	[−1·74, 1·12]	0·47
112	ddC[†]	ZDV[†]	−0·009	0·0046	−0·45	0·06	[−1·38, 1·49]	0·69
114	ddC[2·25]	ZDV[600]	−0·008	0·0037	0·27	0·08	[−0·22, 0·38]	−1·23
116a	ddI[750]	ZDV[600]	−0·009	0·0047	0·10	−0·11	[−0·49, 0·26]	−1·06
	ddI[500]	ZDV[600]	−0·008	0·0050	−0·02	−0·10	[−0·50, 0·28]	−0·42
116b	ddI[750]	ZDV[600]	−0·009	0·0054	−0·18	−0·14	[−0·47, 0·19]	0·24
	ddI[500]	ZDV[600]	−0·008	0·0043	−0·36	−0·18	[−0·54, 0·15]	0·98
118	ddI[200]	ddI[750]	−0·009	0·0050	0·11	0·08	[−0·23, 0·40]	−0·20
	ddI[500]	ddI[750]	−0·009	0·0053	0·17	0·05	[−0·25, 0·36]	−0·73
119	ddC[2·25]	ZDV[600]	−0·009	0·0046	−0·04	−0·10	[−0·83, 0·57]	−0·21
155	ZDV/ddC[600/2·25]	ZDV[600]	−0·009	0·0057	−0·10	−0·24	[−0·58, 0·07]	−0·87
	ddC[2·25]	ZDV[600]	−0·009	0·0048	0·08	−0·14	[−0·48, 0·19]	−1·37
175	ZDV/ddC[600/2·25]	ZDV[600]	−0·009	0·0052	−0·35	−0·30	[−0·77, 0·18]	0·20
	ZDV/ddI[600/400]	ZDV[600]	−0·010	0·0046	−0·47	−0·70	[−1·28, −0·16]	−0·86
	ddI[400]	ZDV[600]	−0·008	0·0045	−0·49	−0·34	[−0·83, 0·12]	0·59
229	ZDV/SQV[600/1800]	ZDV/ddC[600/2·25]	−0·009	0·0045	0·15	−0·06	[−1·06, 0·96]	−0·40
	ZDV/ddC/SQV[600/2·25/1800]	ZDV/ddC[600/2·25]	−0·008	0·0047	−0·84	−0·16	[−1·50, 1·21]	0·99
241	ZDV/ddI/NVP[600/400/400]	ZDV/ddI[600/400]	−0·009	0·0043	0·21	−0·24	[−0·79, 0·31]	−1·59

* Z-score $= \dfrac{\hat{\theta}* - \hat{\theta}_{\text{pred}}}{\sqrt{\{\text{var}(\hat{\theta}_{\text{pred}})\}}}$ where $\hat{\theta}_{\text{pred}}$ and $\text{var}(\hat{\theta}_{\text{pred}})$ are the predicted mean and variance, respectively

† The ddC dose was weight dependent and the ZDV dose depended on the dose of ZDV being received prior to study entry

Table IV. Prediction of the true treatment difference on the log hazard ratio for AIDS or death (θ^*) given various values and standard errors of the estimated difference in mean change in CD4 cell count ($\hat{\gamma}^*$) for several priors for τ^2 for model B (medians and 95 per cent prediction intervals are reported)

$\hat{\gamma}^*$	Prior*	$\delta^* = 0$	$\delta^* = 5$	$\delta^* = 10$	$\delta^* = 15$
0	I	0·00 (–0·12, 0·12)	0·00 (–0·14, 0·14)	0·00 (–0·21, 0·21)	0·00 (–0·29, 0·29)
	II	0·00 (–0·18, 0·18)	0·00 (–0·19, 0·19)	0·00 (–0·24, 0·24)	0·00 (–0·31, 0·31)
	III	0·00 (–0·22, 0·22)	0·00 (–0·23, 0·23)	0·00 (–0·28, 0·28)	0·00 (–0·34, 0·34)
10	I	–0·09 (–0·22, 0·05)	–0·09 (–0·24, 0·07)	–0·09 (–0·31, 0·12)	–0·09 (–0·39, 0·20)
	II	–0·09 (–0·27, 0·09)	–0·09 (–0·28, 0·11)	–0·09 (–0·34, 0·15)	–0·09 (–0·41, 0·23)
	III	–0·09 (–0·31, 0·14)	–0·09 (–0·33, 0·15)	–0·09 (–0·37, 0·18)	–0·09 (–0·44, 0·25)
20	I	–0·17 (–0·31, –0·03)	–0·17 (–0·34, –0·01)	–0·17 (–0·40, 0·04)	–0·17 (–0·48, 0·11)
	II	–0·17 (–0·36, 0·01)	–0·17 (–0·38, 0·03)	–0·17 (–0·44, 0·07)	–0·17 (–0·51, 0·14)
	III	–0·17 (–0·41, 0·05)	–0·17 (–0·42, 0·07)	–0·17 (–0·47, 0·10)	–0·17 (–0·53, 0·16)
30	I	–0·25 (–0·42, –0·10)	–0·25 (–0·44, –0·08)	–0·25 (–0·51, –0·04)	–0·25 (–0·58, 0·03)
	II	–0·26 (–0·46, –0·06)	–0·26 (–0·48, –0·05)	–0·26 (–0·54, –0·01)	–0·26 (–0·61, 0·05)
	III	–0·26 (–0·50, –0·02)	–0·26 (–0·52, –0·01)	–0·26 (–0·57, 0·02)	–0·26 (–0·64, 0·07)
40	I	–0·34 (–0·52, –0·16)	–0·34 (–0·55, –0·15)	–0·34 (–0·62, –0·11)	–0·34 (–0·69, –0·05)
	II	–0·34 (–0·57, –0·13)	–0·34 (–0·59, –0·12)	–0·34 (–0·64, –0·08)	–0·34 (–0·72, –0·03)
	III	–0·34 (–0·61, –0·09)	–0·34 (–0·63, –0·08)	–0·34 (–0·68, –0·05)	–0·34 (–0·75, –0·01)
50	I	–0·42 (–0·64, –0·21)	–0·42 (–0·67, –0·21)	–0·42 (–0·72, –0·18)	–0·42 (–0·80, –0·13)
	II	–0·43 (–0·68, –0·19)	–0·43 (–0·71, –0·18)	–0·43 (–0·75, –0·15)	–0·43 (–0·82, –0·10)
	III	–0·43 (–0·72, –0·15)	–0·43 (–0·74, –0·14)	–0·43 (–0·78, –0·13)	–0·43 (–0·85, –0·08)
60	I	–0·51 (–0·76, –0·27)	–0·51 (–0·78, –0·26)	–0·51 (–0·84, –0·24)	–0·51 (–0·91, –0·20)
	II	–0·51 (–0·80, –0·25)	–0·51 (–0·82, –0·24)	–0·51 (–0·87, –0·22)	–0·51 (–0·93, –0·18)
	III	–0·52 (–0·84, –0·21)	–0·52 (–0·86, –0·21)	–0·52 (–0·90, –0·18)	–0·52 (–0·96, –0·15)

* Prior I: $\pi(\tau^2) = \dfrac{\sigma_c}{(\sigma_c + \tau)^2}\dfrac{1}{2\tau}$; prior II: $\pi(\tau^2) = \dfrac{\sigma_c^2}{(\sigma_c^2 + \tau^2)^2}$; prior III: $\pi(\tau^2) = \mathrm{d}\tau^2$

change in CD4 cell count for that comparison. We can then compare this predicted value with that observed by noting that the standard deviation of the predicted value for treatment comparison i is $\sqrt{\{\sigma_i^2 + \mathrm{var}(\theta_i|\hat{\gamma}_i, \delta_i^2, \hat{\theta}_{(-i)}, \hat{\gamma}_{(-i)}\}}$, where $(\hat{\theta}_{(-i)}, \hat{\gamma}_{(-i)})$ denotes the data without treatment comparison i (see equation (3) for computation of the specified distribution). We repeat this with omission of each comparison in turn. Table III shows the results of such an analysis using model B and prior II. None of the comparisons seems very influential as the posterior medians for both τ^2 and β do not show changes that would substantially alter any conclusions that one might draw from the model. In addition, all 24 comparisons had the predicted treatment difference on the clinical outcome within two standard deviations of the observed value and 95 per cent credible intervals that included the observed value. These results indicate a good predictive fit. Similar analyses using either prior I or prior III gave the same conclusions.

One of the most important uses of the meta-analysis is prediction of the treatment difference on the clinical outcome given the observed treatment difference on the response variable. Table IV shows the relevant predictions for various values of the estimated treatment difference on mean change in CD4 cell count, $\hat{\gamma}^*$, and for various values of its standard deviation, δ^*. For any given values for $\hat{\gamma}^*$ and δ^*, the choice of prior distribution for τ^2 has little impact on the median predicted value, though the intervals tend to widen a little in moving from prior I to II to III.

Consider first the column for which $\delta^* = 0$; this corresponds to the idealistic situation in which the true difference on the mean change in CD4 cell count is known. In this situation, if τ^2 was equal to zero, it is possible to predict θ^* exactly and the prediction interval has zero width. Thus the intervals shown in the column for which $\delta^* = 0$ in Table IV reflect the uncertainty in the predictions that arise from the possibility that τ^2 is non-zero. These intervals exclude $\theta^* = 0$ for values of $\hat{\gamma}^*$ greater than 20 or 30 cells/mm^3, and indicate that we require this magnitude of true difference in change in CD4 cell count before we have reasonable certainty of a corresponding clinical benefit, though such a benefit might be minimal.

In practice, of course, the true difference in change in CD4 cell count is unknown. Moving across any particular row in Table IV shows the effect of decreasing precision in estimation of the treatment difference on the change in CD4 cell count on the ability to predict, reliably, the associated difference in clinical outcome. For standard deviations, δ^*, of 10 or 15 cells/mm^3 (which might be typical in practice as indicated by the confidence regions in Figure 1), we require estimated differences in the mean change of the order of 30 or 40 cells/mm^3 before we have reasonable certainty of a corresponding clinical benefit. As shown in Table I, such differences were uncommon in the studies of the AIDS Clinical Trials Group. Requiring such evidence, however, might form a basis for preliminary approval of a drug pending the results of clinical outcome studies, particularly if the drug concerned is similar in nature to those included in the meta-analysis. In contrast, for selecting drugs for taking forward from phase II to phase III clinical trials, we might find a smaller difference in mean change in CD4 cell count acceptable depending on the magnitude of clinical benefit that we anticipate from the model.

5. DISCUSSION

It is interesting to contrast our results from the meta-analysis to the results found in analyses of data from individual clinical trials. The published results have concerned placebo-controlled trials[4–6,24] and have presented mixed results on the quality of CD4 cell count as a surrogate marker for the development of AIDS or death, though none of them suggests that it is a particularly strong surrogate marker, probably explaining no more than about 30 per cent of the effect of ZDV on clinical outcome. Although further research is needed to relate the results of analyses within individual studies to the results of a meta-analysis, there are some reasons why the results may differ.

First, it may be possible for τ^2 to be zero or very close to zero, so that a treatment difference on a response variable is a good predictor of a treatment difference on clinical outcome, even if there are alternative mechanisms of treatment action not mediated through that response variable. This is provided that the percentage of the effects of treatments on clinical outcome explained by the response variable is very similar for all treatments. Lack of similarity is more worrisome and would be captured in the meta-analysis by τ^2 being somewhat greater than zero.

The second concerns the characteristic of the trajectory of the response variable evaluated. In our meta-analysis, we focused on evaluation of a short-term change (over six months) in the CD4 cell count and assessed how treatment differences on that change are associated with longer-term differences on the clinical outcome (over two years). This is evaluation of the change in CD4 cell count as an 'intermediate' marker.[3,9] The focus of the analyses within individual studies, however, has generally been on evaluation of the extent to which the trajectory of the CD4 cell count throughout time explains the treatment difference on clinical outcome over time; this is to

evaluate CD4 cell count as a 'concurrent' marker.[9] It may be possible that the difference between the two questions asked might lead to different conclusions.

A third reason is the possibility of effect attenuation or regression dilution in estimation of the magnitude of surrogacy within an individual study due to within-subject variation and measurement error in the covariate of interest, an important issue with CD4 cell counts.[25,26] This is well known in epidemiologic and sociologic studies where one often finds that effects estimated across subjects within populations are smaller than those estimated across populations of subjects.[27,28] In addition, CD4 cell counts are often measured only every few months so that the within-study (time-dependent proportional hazards) models used require assumptions about how the count changes between measurements. Although some of the studies have attempted to circumvent these issues by modelling the 'underlying' CD4 cell count, they may be sensitive to the models required to achieve this, particularly as, in practice, CD4 cell counts are more difficult to obtain after a subject prematurely stops taking study medications.

As with any meta-analysis, it is important to avoid biases that can arise through selective inclusion of clinical trials in the analysis. Thus, one should include all trials with recorded data on the marker and clinical outcome over the durations of interest. This will require extensive searches particularly to identify unpublished studies as these perhaps more likely result in non-significant effects on the clinical outcome.[29] Note that if the numbers of clinical events within any trial are small, then we may find it necessary to extend our model to avoid making the assumption that $\hat{\theta}_i$ is normally distributed about θ_i. This might be possible using alternative Bayesian methods for meta-analysis.[30] Another issue with meta-analysis is the assumption that the within-study variances are known. As the summary statistics provide no information on these within-study variances, we assumed them known.[12] Of interest is an investigation of the sensitivity of the results to the fact that these variances are unknown, but estimated. For example, a multiple imputation type approach that perturbed the covariance matrices and that refit the model could help account for the uncertainty in these parameters. How to perturb these covariance matrices in an unresolved issue.

Before closing, it is useful to note that the investigation of surrogacy within both individual studies and in a meta-analysis may be fraught with difficulty if the response variable is not a perfect surrogate marker. We have already noted the importance of establishing that any response variable is prognostic of disease progression before we undertake an analysis so that we can avoid other variables that might be associated with treatment differences on clinical outcome but not actually be mediators of treatment effects. For example, we may find increased drug toxicities with increased clinical benefits though the toxicities are not predictive of disease progression. Such variables might not perform well when we consider other drugs in future studies. Beyond this, if the response variable mediates only partially the effect on clinical outcome, then it is possible that we may detect that association whether our analysis is model-based within an individual study or involves a meta-analysis. It is then possible, however, that we find the association negated in other trials, particularly if they involve other drugs, if a drug has an adverse effect through other mechanisms. In this respect, one may find the sensitivity analysis that we have described particularly useful in establishing just how well a model performs with respect to individual drug comparisons. Hence, we should view this as an important component of model validation before we make any claims of surrogacy.

In conclusion, we have proposed a methodology that uses meta-analysis to investigate the value of potential surrogate markers. This approach is relatively simple to apply and should provide an analysis that complements investigations within individual studies. In particular, its

strength lies in those situations where individual studies are underpowered to provide strong conclusions in their own right.

APPENDIX: INVESTIGATING PLAUSIBLE VALUES FOR ρ_i

Estimation of ρ_i requires the specification of a joint model for the relationship between the treatment differences on the marker and on the clinical outcome within each trial, or application of a bootstrap technique as we have done in our example. Thus this estimation requires the availability of the individual patient data from each clinical trial; this may not always be obtainable. In this Appendix, we present results from a short simulation study to show that ρ_i is likely small in magnitude and that the assumption of a common value for all the ρ_i's is reasonable when patient-level data are unavailable. One can use these results to identify plausible values for ρ_i which one can use when investigating the association of interest, though we recommend consideration of several values and the evaluation of the sensitivity of conclusions to the choice of value.

In each simulation, we conducted 1000 clinical trials that involved 100 subjects assigned to each of two treatments. For each subject j in each clinical trial, we sampled values of the response variable, X_j, from a normal distribution with mean γz_j and variance 1, where z_j is a binary treatment indicator. Thus, γ measures the treatment difference per standard deviation of values of the response variable in the population sampled. We then simulated clinical outcome data for each subject based on the following model, $\text{logit}(p_j) = \alpha' + \beta' z_j + \delta X_j$, where p_j is the probability of a clinical outcome for subject j. In the model, we set β' to be zero to indicate that the response variable is, in fact, a perfect marker, according to Prentice's criteria,[2] as the response variable then explains all of the treatment difference. With these simulated data, in each trial we fitted the models $\text{logit}(p_j) = \alpha + \theta z_j$ and $E(X_j) = \eta + \gamma z_j$ to provide estimates of the treatment differences on the clinical outcome, $\hat{\theta}$, and the response variable, $\hat{\gamma}$. We then estimated the correlation between $\hat{\theta}$ and $\hat{\gamma}$ across the 1000 simulated trials.

We repeated this simulation for various values of γ, the treatment difference on the response variable (per standard deviation of values of the response variable in the population sampled) between 0·2 and 2, the probability of a clinical outcome in the treatment group for which $z_j = 0$, given by $\frac{\exp(\alpha')}{1+\exp(\alpha')}$, between 0·1 and 0·5, and the odds ratio for clinical outcomes, $\exp(\delta)$, for a standard deviation difference in the response variable between 1·1 to 5·0. Table V shows the results from these simulation studies. In general, the correlation is small (between 0 and 0·2) except in the circumstances when the odds ratio, $\exp(\delta)$, is large (for example, 2 or higher). This agrees with our empirical findings in the example where the values of ρ_i shown in Table I are all small in magnitude. The correlation also tends to zero as the odds ratio, $\exp(\delta)$ tends to 1; this is expected as $\exp(\delta) = 1$ implies that the response variable is not predictive of clinical outcome. Note that if the response variable is not a perfect marker, then β' will be non-zero and so the correlation between $\hat{\gamma}$ and $\hat{\theta}$ is smaller than the values shown in Table V. Thus, we recommend looking at the sensitivity of the estimates to values of ρ_i from 0 up to about 0·2, though if the effect of treatment is to increase the level of the response variable while reducing the rate of clinical outcomes, then the correlations are the negative of the values shown and so one might consider the values from 0 down to about $-0{\cdot}2$. Furthermore, as there is little variation in the correlations in each column, even as the probability of a clinical outcome in the treatment group 0 varies, it appears reasonable to assume that ρ_i is constant across trials when estimating α, β and τ^2, especially for odds ratios less than two.

Table V. Correlation of $\hat{\gamma}$ and $\hat{\theta}$ from a simulation study for selected values of γ, the odds ratio for a clinical outcome associated with a standard deviation change in marker levels, exp δ, and the probability of clinical outcome in the control group, $\exp(\alpha')/[1 + \exp(\alpha')]$

	Probability: $\exp(\alpha')/[1 + \exp(\alpha')]$														
	0·1					0·2					0·5				
	exp δ					exp δ					exp δ				
γ	5	2	1·5	1·2	1·1	5	2	1·5	1·2	1·1	5	2	1·5	1·2	1·1
0·2	0·49	0·17	0·13	0·01	0·01	0·49	0·26	0·13	0·05	0·04	0·52	0·36	0·19	0·06	0·08
0·4	0·37	0·20	0·08	0·08	0·02	0·48	0·26	0·13	0·05	0·04	0·54	0·33	0·19	0·15	0·03
0·6	0·41	0·15	0·07	0·07	0·01	0·44	0·23	0·14	0·01	0·00	0·50	0·32	0·15	0·11	0·01
0·8	0·36	0·17	0·15	0·07	0·05	0·44	0·30	0·15	0·09	0·00	0·55	0·30	0·18	0·13	0·02
1	0·22	0·16	0·13	0·04	0·04	0·37	0·23	0·16	0·12	0·04	0·46	0·31	0·22	0·05	0·04
2	0·08	0·09	0·01	0·02	0·01	0·13	0·19	0·07	0·06	0·03	0·37	0·33	0·24	0·09	0·08

ACKNOWLEDGEMENTS

We are grateful to the AIDS Clinical Trials group and the protocol teams for permission to use the data in our example. This work was supported by a Howard Hughes Medical Institute Predoctoral Fellowship, grant AI 24643 from the National Institutes of Health, and by the Statistical and Data Analysis Center of the AIDS Clinical Trials group.

REFERENCES

1. Wittes, J., Lakatos, E. and Probstfield, J. 'Surrogate endpoints in clinical trials: cardiovascular diseases', *Statistics in Medicine*, **8**, 415–425 (1989).
2. Prentice, R. L. 'Surrogate markers in clinical trials: definition and operational criteria', *Statistics in Medicine*, **8**, 431–440 (1989).
3. Freedman, L. S., Graubard, B. L. and Schatzkin, A. 'Statistical validation of intermediate endpoints for chronic diseases', *Statistics in Medicine*, **11**, 167–178 (1993).
4. Choi, S., Lagakos, S. W., Schooley, T. T. and Volberding, P. A. '$CD4^+$ lymphocytes are an incomplete surrogate marker for clinical progression in persons with asymptomatic HIV infection taking zidovudine', *Annals of Internal Medicine*, **118**, 674–680 (1993).
5. De Gruttola, V., Wulfsohn, M., Fischl, M. A. and Tsiatis, A. 'Modeling the relationship between survival and CD4 lymphocytes in patients with AIDS and AIDS-related complex', *Journal of Acquired Immune Deficiency Syndromes*, **6**, 359–365 (1993).
6. Lin, D. Y., Fischl, M. A. and Schoenfeld, D. A. 'Evaluating the role of CD4-Lymphocyte counts as surrogate endpoints in human immunodeficiency virus clinical trials', *Statistics in Medicine*, **12**, 835–842 (1993).
7. De Gruttola, V. and Fleming, T. R. (Letter), *New England Journal of Medicine*, **334**, 1671–1672 (1996).
8. De Gruttola, V., Fleming, T., Lin, D. Y. and Coombs, R. 'Validating surrogate markers: Are we being naive?', Technical report, Department of Biostatistics, Harvard School of Public Health, Boston, MA, 1996.
9. Hughes, M. D., De Gruttola, V. and Welles, S. 'Evaluating surrogate markers', *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, **10**, (Suppl 2), S1–S8 (1995).
10. A'Hern, R. P., Ebbs, S. R. and Baum, M. B. 'Does chemotherapy improve survival in advanced breast cancer? A statistical overview', *British Journal of Cancer*, **57**, 615–618 (1988).
11. Fleming, T. R. 'Surrogate markers in AIDS and cancer trials', *Statistics in Medicine*, **13**, 1423–1435 (1994).
12. DuMouchel, W. 'Hierarchical Bayes linear models for meta-analysis' *National Institute of Statistical Sciences Technical Report*, **27** (1994).

13. DerSimonian, R. and Laird, N. 'Meta-analysis in clinical trials', *Controlled Clinical Trials*, **7**, 177–186 (1986).
14. Strawderman, W. E. 'Proper Bayes minimax estimators of the multivariate normal mean', *Annals of Mathematical Statistics*, **42**, 385–388 (1971).
15. Berger, J. O. *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, New York, 1995.
16. Efron, B. and Tibshirani, R. 'Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy', *Statistical Science*, **1**, 54–75 (1986).
17. Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. *Markov Chain Monte Carlo in Practice*, Chapman and Hall, London, 1996.
18. Smith, A. F. M. and Roberts, G. O. 'Bayesian computation via the Gibbs sampler and related Markov Chain Monte Carlo Methods', *Journal of the Royal Statistical Society B*, **55**, 3–23 (1993).
19. Kass, R. E. and Raftery, A. E. 'Bayes Factors', *Journal of the American Statistical Association*, **90**, 773–795 (1995).
20. Verdinelli, I. and Wasserman, L. 'Computing Bayes factors using the Savage Dickey density ratio', *Journal of the American Statistical Association*, **90**, 614–618 (1995).
21. Kass, R. E. and Wasserman, L. 'A reference Bayesian test for nested hypotheses and its relationship to the Schwarz Criterion', *Journal of the American Statistical Association*, **90**, 928–934 (1995).
22. Gelfand, A. E. and Smith, A. F. M. 'Sampling-based approaches to calculating marginal densities', *Journal of the American Statistical Association*, **85**, 398–409 (1990).
23. Thomas, A., Spiegelhalter, D. J. and Wilks, W. R. 'BUGS:A program to perform Bayesian inference using Gibbs sampling', *in* Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M. (eds.), *Bayesian Statistics 4*, Clarendon Press, Oxford, 1992, pp. 837–842.
24. Tsiatis, A. A., DeGruttola, V. and Wulfsohn, M. 'Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS', *Journal of American Statistical Association*, **90**, 27–37 (1995).
25. Hughes, M. D., Stein, D. S., Gundacker, H. M., Valentine, F. T., Phair, J. P. and Volberding, P. A. 'Within-subject variation in CD4 lymphocyte count in asymptomatic human immunodeficiency virus infection: implications for patient monitoring', *Journal of Infectious Diseases*, **169**, 28–36 (1994).
26. Hoover, D. R., Graham, N. M., Chen, B., Taylor, J. M., Phair, J., Zhou, S. Y. and Munoz, A. 'Effect of $CD4^+$ cell count measurement variability on staging HIV-1 infection', *Journal of Acquired Immune Deficiency Syndromes*, **5**, 794–802 (1992).
27. MacMahon, S., Peto, R., Cutler, J., Collins, R., Sorlie, P., Neaton, J., Abbott, R., Godwin, J., Dyer, A. and Stamler, J. 'Blood pressure, stroke, and coronary heart disease: part I, prolonged differences in blood pressure: prospective observational studies corrected for regression dilution bias', *Lancet*, **335**, 765–774 (1990).
28. Robinson, W. S. 'Ecological correlations and the behavior of individuals', *American Sociological Review*, **15**, 351–357 (1950).
29. Iyengar, S. and Greenhouse, J. B. 'Selection models and the file drawer problem', *Statistical Science*, **3**, 109–117 (1988).
30. Smith, T. C., Spiegelhalter, D. J. and Thomas, A. 'Bayesian approaches to a random effects meta-analysis: a comparative study', *Statistics in Medicine*, **14**, 2685–2699 (1995).
31. Fischl, M. A., Parker, C. B., Pettinelli, C., Wulfsohn, M., Hirsch, M. S., Collier, A. C., Antoniskis, D., Ho, M., Richman, D. D., Fuchs, E., Merigan, T. C., Reichman, R. C., Gold, J., Steigbigel, N., Leoung, G. S., Rasheed, S., Tsiatis, A. and the AIDS Clinical Trials Group. 'A randomized controlled trial of a reduced daily dose of zidovudine in patients with acquired immunodeficiency syndrome', *New England Journal of Medicine*, **323**, 1009–1014 (1990).
32. Fischl, M. A., Richman, D. D., Hansen, N., Collier, A. C., Carey, J. T., Para, M. F., Hardy, W. D., Dolin, R., Powderly, W. G., Allan, J. D., Wong, B., Merigan, T. C., McAuliffe, V. J., Hyslop, N. E., Rhame, F. S., Balfour, H. H., Spector, S. A., Volberding, P., Pettinelli, C., Anderson, J. and the AIDS Clinical Trials Group. 'The safety and efficacy of zidovudine in the treatment of subjects with mildly symptomatic human immunodeficiency virus type I (HIV) infection', *Annals of Internal Medicine*, **112**, 727–737 (1990).
33. Volberding, P. A., Lagakos, S. W., Grimes, J. M., Stein, D. S., Rooney, J., Meng, T. C., Fischl, M. A., Collier, A. C., Phair, J. P., Hirsch, M. S., Hardy, W. D., Balfour, H. H., Reichman, R. C. for the AIDS Clinical Trials Group. 'Immediate versus deferred zidovudine in asymptomatic HIV-infected subjects with CD4 cell counts of 500 per microliter or greater', *New England Journal of Medicine*, **333**, 401–413 (1995).

34. Volberding, P. A., Lagakos, S. W., Koch, M. A., Pettinelli, C., Myers, M. W., Booth, D. K., Balfour, H. H., Reichman, R. C., Bartlett, J. A., Hirsch, M. S., Murphy, R. L., Hardy, D., Soiero, R., Fischl, M. A., Bartlett, J. G., Merigan, T. C., Hyslop, N. E., Richman, D. D., Valentine, F. T., Corey, L. and the AIDS Clinical Trials Group of the National Institute of Allergy and Infectious Diseases. 'Zidovudine in asymptomatic human immunodeficiency virus infection', *New England Journal of Medicine*, **322**, 941–949 (1990).
35. Merigan, T. C., Amato, D. A., Balsley, J., Power, M., Price, W. A., Benoit, S., Perez-Michael, A., Brownstein, A., Kramer, A. S., Brettler, D., Aledort, L., Ragni, M. V., Andes, W. A., Gill, J. C., Goldsmith, J., Stabler, S., Sanders, N., Gjerset, G., Lusher, J. and the NHF-ACTG 036 Study Group. 'Placebo-controlled trial to evaluate zidovudine in treatment of human immunodeficiency virus infection in asymptomatic patients with hemophilia', *Blood*, **4**, 900–906 (1991).
36. Dolin, R., Amato, D. A., Fischl, M. A., Pettinelli, C., Beltangady, M., Liou, S. H., Brown, M. J., Cross, A. P., Hirsch, M. S., Hardy, W. D., Mildvan, D., Blair, D. C., Powderly, W. G., Para, M. F., Fife, K. H., Steigbigel, R. T., Smaldone, L. and the AIDS Clinical Trials Group. 'Zidovudine compared with didanosine in patients with advanced HIV type 1 infection and little or no previous experience with zidovudine', *Archives of Internal Medicine*, **155**, 961–974 (1995).
37. Kahn, J. O., Lagakos, S. W., Richman, D. D., Cross, A. C., Pettinelli, C., Liou, S. H., Brown, M., Volberding, P. A., Crumpacker, C. S., Beall, G., Sacks, H. S., Merigan, T. C., Beltangady, M., Smaldone, L., Dolin, R. and the NIAID AIDS Clinical Trials Group. 'A controlled trial comparing continued zidovudine with didanosine in human immunodeficiency virus infection', *New England Journal of Medicine*, **327**, 581–587 (1992).
38. Allan, J. D., De Gruttola, V., Cross, A., Seidlin, M., Bassett, R., Brown, M., McLaren, C., Smaldone, L. and Pettinelli, C. 'Executive summary of the final analysis of ACTG 118', National Institute of Infectious and Allergic Diseases, Bethesda, MD, 1993.
39. Fischl, M. A., Olson, R. M., Follansbee, S. E., Lalezari, J. P., Henry, D. H., Frame, P. T., Remick, S. C., Salgo, M. P., Lin, A. H., Nauss-Karol, C., Lieberman, J. and Soo, W. 'Zalcitabine compared with zidovudine in patients with advanced HIV-1 infection who received previous zidovudine therapy', *Annals of Internal Medicine*, **118**, 762–769 (1993).
40. Fischl, M. A., Stanley, K., Collier, A. C., Arduino, J. M., Stein, D. S., Feinberg, J. E., Allan, J. D., Goldsmith, J. C., Powderly, W. G. and the NIAID Clinical Trials Group. 'Combination and monotherapy with zidovudine and zalcitabine in patients with advance HIV disease', *Annals of Internal Medicine*, **122**, 24–32 (1995).
41. Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundacker, H., Schooley, R. T., Haubrich, R. H., Henry, W. K., Lederman, M. M., Phair, J. P., Niu, M., Hirsch, M. S., Merigan, T. C., for the AIDS Clinical Trials Group Study 175 Study Team. 'A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter', *New England Journal of Medicine*, **335**, 1081–1090 (1996).
42. Collier, A. C., Coombs, R. W., Schoenfeld, D. A., Bassett, R. L., Timpone, J., Baruch, A., Jones, M., Facey, K., Whitacre, C., McAuliffe, V. J., Friedman, H. M., Merigan, T. C., Reichman, R. C., Hooper, C., Corey, L. for the AIDS Clinical Trials Group. 'Treatment of human immunodeficiency virus infection with saquinavir, zidovudine and zalcitabine', *New England Journal of Medicine*, **334**, 1011–1017 (1996).
43. D'Aquila, R. T., Hughes, M. D., Johnson, V. A., Fischl, M. A., Sommadossi, J. P., Liou, S., Timpone, J., Myers, M., Basgoz, N., Niu, M., Hirsch, M. S. and the National Institute of Allergy and Infectious Diseases AIDS Clinical Trials Group Protocol 241 Investigators. 'Nevirapine, zidovudine, and didanosine compared with zidovudine and didanosine in patients with HIV-1 infection. A randomized, double-blind, placebo-controlled trial', *Annals of Internal Medicine*, **124**, 1019–1031 (1996).

