Research Article

Statistics in Medicine

Received 13 September 2008, Accepted 4 March 2010 Published online 23 April 2010 in Wiley Interscience

(www.interscience.wiley.com) DOI: 10.1002/sim.3931

Predicting treatment effects using biomarker data in a meta-analysis of clinical trials

Y. Li*† and J. M. G. Taylor

A biomarker (*S*) measured after randomization in a clinical trial can often provide information about the true endpoint (*T*) and hence the effect of treatment (*Z*). It can usually be measured earlier and more easily than *T* and as such may be useful to shorten the trial length. A potential use of *S* is to completely replace *T* as a surrogate endpoint to evaluate whether the treatment is effective. Another potential use of *S* is to serve as an auxiliary variable to help provide information and improve the inference on the treatment effect prediction when *T* is not completely observed. The objective of this report is to focus on its role as an auxiliary variable and to identify situations when *S* can be useful to increase efficiency in predicting the treatment effect in a new trial in a multiple-trial setting. Both *S* and *T* are continuous. We find that higher efficiency gain is associated with higher trial-level correlation but not individual-level correlation when only *S*, but not *T* is measured in a new trial; but, the amount of information recovery from *S* is usually negligible. However, when *T* is partially observed in the new trial and the individual-level correlation is relatively high, there is substantial efficiency gain by using *S*. For design purposes, our results suggest that it is often important to collect markers that have high adjusted individual-level correlation with *T* and at least a small amount of data on *T*. The results are illustrated using simulations and an example from a glaucoma clinical trial.

Keywords: auxiliary variables; biomarker; clinical trials; meta analysis; mixed model; surrogate

1. Introduction

A biomarker (S) in a clinical trial is a type of variables intended to provide information about the true endpoint (T) and the effect of treatment (Z). It is often an intermediate physical or laboratory indicator in a disease progression process, and can be measured earlier and is often easier to collect than T. Examples of biomarkers include CD4 counts in AIDS, blood pressure and serum cholesterol level in cardiovascular disease, and prostate-specific antigen in prostate cancer studies. Early measurements are also used as biomarkers for the later measurements, such as the earlier vision test result as a biomarker for the later result in a study on patients with age-related muscular degeneration [1]. Different investigators use different terminologies for the roles of the biomarkers [2]. In this paper, we call S a surrogate endpoint when the potential use of S is to completely replace T to evaluate whether the treatment is effective [3]. Alternatively, when S is used to help provide information or enhance the efficiency of the estimator of the treatment effect on T when T is not completely observed, we call S as auxiliary variables [4]. When the true endpoints are rare, later-occurring or costly to obtain, the proper use of good biomarkers can substantially reduce the trial size and duration, hence lower the expense and lead to earlier decision making.

Previous research on the biomarker has often focused on the potential role of S as a surrogate endpoint for T. In a landmark article, Prentice [3] proposed a formal definition for perfect surrogacy and provided validation criteria for a single-trial setting. The criteria require that changes in S fully capture the effect of treatment on T. This paper inspired much research in the field, but the criteria are considered too restrictive for practical use. To relax the criteria, a surrogacy measure based on the proportion of the treatment effect explained (PTE) by S was proposed [5] and further studied and extended by several other authors (e.g. [6–8]). Freedman [5] also suggested that the PTE confidence interval's lower bound be >0.75 for a marker to be acceptable as a surrogate endpoint. However, this requires the treatment effect on T

Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109-2029, U.S.A.
**Correspondence to: Y. Li, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109-2029, U.S.A.*
†E-mail: yunlisph@umich.edu

Contract/grant sponsor: National Institutes of Health; contract/grant number: CA129102

to be very strong, which is rarely observed in practice [7, 9]. The PTE estimator is also highly variable and can be out of the [0,1] range [6, 10]; hence, its practical use is limited.

From a biological aspect, there are often multiple causal pathways leading to disease and complex mechanisms by which the treatment functions; hence, a biomarker may or may not mediate the effect of the treatment on T and the surrogacy measures are often not directly transferable from one study to another. Another problem is that S may not capture the harmful side effect of the treatment. These associated uncertainties in the use of S in replacing T to test a new treatment can lead to incorrect, even harmful conclusions [10, 11]. As a result, very few biomarkers have been accepted as valid surrogate endpoints for T and their potential use as substitutes has been less than promising.

With new biomarkers being discovered and developed at a phenomenal rate, the clinical research community continue to be extremely interested in biomarkers in clinical trials. In this paper, we focus on the use of S as an auxiliary variable in helping predict the treatment effect on T. As we shall see, this role of a biomarker proves to be more promising. One of the most common scenarios for S to be useful as an auxiliary outcome is when one has more information on S than that on T for a study population. This occurs often in practice, since patients are usually recruited into a trial sequentially in calendar time and S is observed more often and earlier than T, particularly on those enrolled early. Previous surrogacy measures are often proposed based on summary statistics in order to identify a replacement for T, and they are not usually suggested explicitly for the purpose of prediction. In the presence of individual-level data, a biomarker may actually be effective as an auxiliary outcome in enhancing inference, but not be identified as such using existing surrogacy measures. A strong association between S and T does not suffice for S to be a substitute for T; as Baker *et al.* [12] stated, 'a correlate does not make a surrogate'. However, when individual data on T exist, a strong association can inform and increase the efficiency of treatment effect prediction, as we shall demonstrate.

A number of authors have explored the role of biomarkers as auxiliary variables. However, the opinions on their values have been mixed, as noted by Cook and Lawless [13]. In much of the previous work, the information recovered from S appears to be very small [4, 14, 15] unless in rare situations when S and T are very highly correlated; however, when there is more structural relationship between S and T, it is more likely to achieve significant efficiency gain by using S [13]. Most of the work mentioned above has focused on the situation when T is the time to an event. When S and T are continuous data, Venkatraman and Begg [16] proposed fully nonparametric tests that incorporate the information from S and found that the amount of efficiency gain through S for these tests is small except in rare occasions when the correlation between S and T is extremely high. A homogeneous sample such as the single trial setting has often been considered in the previous work. When we can identify a group of trials which have similar treatment groups and patient populations, it is natural to use a meta-analytic approach to predict the treatment effect in a new trial. This approach could allow one to account for the heterogeneity among different trials and borrow information from previous trials to improve the efficiency.

In this paper, we will focus on examining the extent of information gain from S in a multiple trial setting. We will examine the situation when T is either completely missing or partially missing in a new trial when we have information on S, T, and Z in the previous trials. The objective is to predict the treatment effect on T in the new trial when S and T are continuous and Z is binary. We examine the factors, particularly, the correlation between S and T and the fraction of missing T that impact the extent of increase in the precision of the treatment effect estimate resulting from utilizing S to identify the situations when S can be beneficial. The results are intended to be of practical value and directly applicable to clinical trials.

In Section 2, we introduce a commonly used bivariate mixed model. In Section 3, we summarize several related methods used to predict the effect of Z on T in a new trial when T is either completely missing or partially missing in the new trial. The methods include those proposed by Buyse *et al.* [9], Gail *et al.* [17], and Henderson [18]. In Section 4, we examine the extent of information recovery from S and its relation to the correlation between S and T. In Section 5, we evaluate the methods and efficiency gain through simulations. In Section 6, we give a data example. In Section 7, we present conclusions.

2. The model

Suppose we have n randomized trials, $i=1,\ldots,n$, where the nth trial is labeled as new and there are m_i patients in the ith trial. Let $Z=0,1$ denote the placebo and treatment groups, respectively, and (S_{ij}, T_{ij}, Z_{ij}) represent S, T, and Z for individual j in trial i. We are interested in predicting the actual treatment effect on T in the new trial (δ_{Tn}) based on previous $(n-1)$ existing trials and whatever data are available in the nth trial. A commonly used bivariate mixed model used to describe the joint distribution of S_{ij} and T_{ij} [9] is

$$\begin{aligned} S_{ij} &= \alpha_0+\alpha_1 Z_{ij}+a_{0i}+a_{1i}Z_{ij}+\varepsilon_{Sij} \\ T_{ij} &= \gamma_0+\gamma_1 Z_{ij}+r_{0i}+r_{1i}Z_{ij}+\varepsilon_{Tij}, \end{aligned} \tag{1}$$

Statistics in Medicine

where

$$\begin{pmatrix}\varepsilon_{Sij}\\ \varepsilon_{Tij}\end{pmatrix}\sim \text{MVN}\left(\begin{pmatrix}0\\0\end{pmatrix},\quad \sigma=\begin{pmatrix}\sigma_{ss} & \sigma_{st}\\ & \sigma_{tt}\end{pmatrix}\right), \tag{2}$$

and

$$\begin{pmatrix}a_{0i}\\ r_{0i}\\ a_{1i}\\ r_{1i}\end{pmatrix}\sim \text{MVN}\left(\begin{pmatrix}0\\0\\0\\0\end{pmatrix},\quad D=\begin{pmatrix}d_{ss} & d_{st} & d_{sa} & d_{sr}\\ & d_{tt} & d_{ta} & d_{tr}\\ & & d_{aa} & d_{ar}\\ & & & d_{rr}\end{pmatrix}\right). \tag{3}$$

The treatment effect in the nth trial is $\delta_{Tn}=\gamma_1+r_{1n}$. Let $Y_i^{\mathrm{T}}=(S_{ij},T_{ij})$, $\varepsilon_i^{\mathrm{T}}=(\varepsilon_{Sij},\varepsilon_{Tij})$, $\beta^{\mathrm{T}}=(\alpha_0,\gamma_0,\alpha_1,\gamma_1)$ and $\eta_i^{\mathrm{T}}=(a_{0i},r_{0i},a_{1i},r_{1i})$. The model (1) can be written in a general mixed model notation as $Y_i=X_i\beta+U_i\eta_i+\varepsilon_i$, where β denotes the fixed effects, η_i denotes the random effects, X_i and U_i are the corresponding design matrices. The vector Y_i follows a bivariate normal distribution with mean $X_i\beta$ and variance $V_i=U_iDU_i^{\mathrm{t}}+\Sigma_i$ where Σ_i is a $2m_i\times 2m_i$ matrix with m_i blocks of σ on the main diagonal and zeros elsewhere.

3. Methods for predicting the treatment effect δ_{Tn} in the new trial

In this section, we introduce several related methods used to predict the effect of Z on T in a new trial when T is either completely missing or partially missing in the new trial.

3.1. Buyse et al. method

Buyse *et al.* (BMBRG) [9] assumed the same model and suggested a method to estimate δ_{Tn} when T is completely unobserved in the nth trial. First, they fit a bivariate mixed model to the data from trial 1 through $(n-1)$ to obtain the estimates of D, α_0, γ_0, α_1, and γ_1, denoted by $\hat{D}$, $\hat{\alpha}_0$, $\hat{\gamma}_0$, $\hat{\alpha}_1$, and $\hat{\gamma}_1$, respectively. Second, they fit a linear regression $S_{nj}=\mu_{0Sn}+\delta_{Sn}Z_{nj}+\varepsilon_{Snj}$ in the nth trial. One then obtains that $\hat{a}_{0n}=\hat{\mu}_{0Sn}-\hat{\alpha}_0$ and $\hat{a}_{1n}=\hat{\delta}_{Sn}-\hat{\alpha}_1$ where $\hat{\mu}_{0Sn}$ and $\hat{\delta}_{Sn}$ are estimates of μ_{0Sn} and δ_{Sn} based on data from the nth trial. Given that β, D, σ, a_{0n}, and a_{1n} are known, BMBRG showed that δ_{Tn} follows a normal distribution with conditional mean

$$E(\delta_{Tn})=\gamma_1+(d_{sr}\quad d_{ar})\begin{pmatrix}d_{ss} & d_{sa}\\ d_{sa} & d_{aa}\end{pmatrix}^{-1}\begin{pmatrix}a_{0n}\\ a_{1n}\end{pmatrix}, \tag{4}$$

and conditional variance

$$\text{var}(\delta_{Tn})=d_{rr}-(d_{sr}\quad d_{ar})\begin{pmatrix}d_{ss} & d_{sa}\\ d_{sa} & d_{aa}\end{pmatrix}^{-1}\begin{pmatrix}d_{sr}\\ d_{ar}\end{pmatrix}. \tag{5}$$

While various methods can be used to obtain the estimate for δ_{Tn}, denoted by $\hat{\delta}_{Tn}$, in our simulations, we replace β, D, σ, a_{0n}, and a_{1n} with their estimates in equations (4) and (5) as often done in practice. Specifically, we obtain β, D, and σ using a restricted maximum likelihood method from PROC MIXED in SAS. We estimate μ_{0Sn} and δ_{Sn} using PROC GLM in SAS and then obtain the estimates for a_{0n} and a_{1n}. However, this often leads to underestimation of $\text{var}(\hat{\delta}_{Tn})$.

3.2. Gail et al. method

Gail *et al.* (GPHC) [17] proposed to estimate δ_{Tn} without involving models for the joint distribution of (S_{ij},T_{ij}) at the individual level. The method applies to the situation when T is completely unobserved in the nth trial. Let $\mu_{Ti}^{\mathrm{T}}=(\mu_{0Ti},\mu_{1Ti})$ represent the marginal means of T in the $Z=0$ and 1 groups in the ith trial and similarly for $\mu_{Si}^{\mathrm{T}}=(\mu_{0Si},\mu_{1Si})$. GPHC assume that $(\mu_{0Ti},\mu_{0Si},\mu_{1Ti},\mu_{1Si})^{\mathrm{T}}$ follows a multivariate normal distribution with covariance φ, where φ is a 4×4 matrix representing the between-trial variance; hence, its estimate $(\hat{\mu}_{0Ti},\hat{\mu}_{0Si},\hat{\mu}_{1Ti},\hat{\mu}_{1Si})^{\mathrm{T}}$ follows a multivariate normal distribution with the covariance $\varphi+\omega_i$, where ω_i is a 4×4 matrix with two block diagonal matrices denoting the within-trial variance for each treatment group. The elements of μ_{Ti}, μ_{Si}, and φ are connected with the parameters in the model (1) in the following way: $\mu_{0Ti}=\gamma_0+r_{0i}$, $\mu_{1Ti}=\gamma_0+r_{0i}+\gamma_1+r_{1i}$, $\mu_{0Si}=\alpha_0+a_{0i}$,

$\mu_{1Si}=\alpha_0+a_{0i}+\alpha_1+a_{1i}$, $\varphi_{11}=d_{tt}+d_{bb}+2d_{tb}$, $\varphi_{12}=d_{ts}+d_{ab}+d_{ta}+d_{sb}$, $\varphi_{13}=d_{tt}+d_{tb}$, $\varphi_{14}=d_{ts}+d_{sb}$, $\varphi_{22}=d_{ss}+d_{aa}+2d_{sa}$, $\varphi_{23}=d_{st}+d_{ta}$, $\varphi_{24}=d_{ss}+d_{sa}$, $\varphi_{33}=d_{tt}$, $\varphi_{34}=d_{st}$, and $\varphi_{44}=d_{ss}$.

GPHC show that μ_{Tn} given $\hat{\mu}_{Sn}$ (and β, φ, and ω) follows a normal distribution with mean

$$E(\mu_{Tn})=\begin{pmatrix}\gamma_0\\ \gamma_0+\gamma_1\end{pmatrix}+\begin{pmatrix}\varphi_{12} & \varphi_{14}\\ \varphi_{23} & \varphi_{34}\end{pmatrix}\begin{pmatrix}\varphi_{22}+\omega_{22n} & \varphi_{24}\\ \varphi_{24} & \varphi_{44}+\omega_{44n}\end{pmatrix}^{-1}\begin{pmatrix}a_{0n}\\ a_{0n}+a_{1n}\end{pmatrix},$$

and variance

$$\mathrm{var}(\mu_{Tn})=\begin{pmatrix}\varphi_{11} & \varphi_{13}\\ \varphi_{13} & \varphi_{33}\end{pmatrix}-\begin{pmatrix}\varphi_{12} & \varphi_{14}\\ \varphi_{23} & \varphi_{34}\end{pmatrix}\begin{pmatrix}\varphi_{22}+\omega_{22n} & \varphi_{24}\\ \varphi_{24} & \varphi_{44}+\omega_{44n}\end{pmatrix}^{-1}\begin{pmatrix}\varphi_{12} & \varphi_{14}\\ \varphi_{23} & \varphi_{34}\end{pmatrix}^{\mathrm{T}},$$

where ω_{22n} denotes the variance of $\hat{\mu}_{0Sn}$ and ω_{44n} for $\hat{\mu}_{1Sn}$.

The treatment effect on T in the new trial, δ_{Tn}, has mean

$$E(\delta_{Tn})=(-1\ \ 1)E(\mu_{Tn}), \tag{6}$$

and variance

$$\mathrm{var}(\delta_{Tn})=(-1\ \ 1)\mathrm{var}(\mu_{Tn})\begin{pmatrix}-1\\ 1\end{pmatrix}. \tag{7}$$

If we drop the terms w_{22n} and w_{44n} from the above expressions, we obtain the identical expressions as those of the BMBRG mean and variance. The GPHC formula takes into account the uncertainty associated with estimating a_{0n} and a_{1n} while BMBRG does not. Similar to BMBRG, GPHC also assume that β, D, and σ are known in deriving equations (6) and (7). Since the uncertainties of β, D, and σ are not accounted for here, $\mathrm{var}(\hat{\delta}_{Tn})$ is often underestimated. Gail *et al.* (2000) noted that this method is analogous to the generalized estimating equations (GEE) [19]. We note that the GEE approach can handle the situation when T is partially observed in the new trial, thus the GPHC method could be generalized and would be worthy of further investigation.

To estimate δ_{Tn} and $\mathrm{var}(\hat{\delta}_{Tn})$, in our simulations, we first obtain $\hat{\varphi}+\hat{\omega}_i$ by calculating the covariances of the treatment- and trial-specific means where $\hat{\varphi}$ and $\hat{\omega}_i$ denote the estimates of φ and ω_i, respectively. We then calculate the treatment-specific covariances of S and T within each trial and then average them over different trials to obtain $\hat{\omega}_i$. From these, we calculate $\hat{\varphi}$. We calculate the overall treatment-specific means as $\hat{\gamma}_0$ and $\hat{\gamma}_1$ (i.e. the estimates of γ_0 and γ_1) and the variances for each treatment group in the new trial for $\hat{\omega}_{22n}$ and $\hat{\omega}_{44n}$ (i.e. the estimates of ω_{22n} and ω_{44n}). We estimate μ_{0Sn} and μ_{1Sn} and then calculate a_{0n} and a_{1n}. Then we plug in these estimates into (6) and (7) to obtain the mean and variance for $\hat{\delta}_{Tn}$.

3.3. *Henderson method (HD)*

While both BMBRG and GPHC methods only apply to the situation when T is completely missing in the new trial, the HD method applies to the situations when T is either completely missing, partially missing, or completely observed in the new trial. Using the general mixed model notation, we can obtain the estimates of β and η_n (denoted by $\hat{\beta}$ and $\hat{\eta}_n$) by solving the mixed model equation which is described by Henderson [18] (details in Appendix A) and their sum follows a normal distribution with mean

$$E(\hat{\beta}+\hat{\eta}_n)=\beta+DU_n^{\mathrm{T}}V_n^{-1}(Y_n-X_n\beta). \tag{8}$$

and variance

$$\begin{aligned}\mathrm{var}(\hat{\beta}+\hat{\eta}_n-\beta-\eta_n)=&\left(\sum_{i=1}^{n}X_i^{\mathrm{T}}V_i^{-1}X_i\right)^{-1}+D-DU_n^{\mathrm{T}}V_n^{-1}U_nD\\ &+DU_n^{\mathrm{T}}V_n^{-1}X_n\left(\sum_{i=1}^{n}X_i^{\mathrm{T}}V_i^{-1}X_i\right)^{-1}X_n^{\mathrm{T}}V_n^{-1}U_nD-2DU_n^{\mathrm{T}}V_n^{-1}X_n\left(\sum_{i=1}^{n}X_i^{\mathrm{T}}V_i^{-1}X_i\right)^{-1}.\end{aligned}$$

The treatment effect for the nth trial has mean

$$E(\hat{\delta}_{Tn})=(0\ \ 0\ \ 0\ \ 1)(\beta+\eta_n) \tag{9}$$

and variance

$$\text{var}(\hat{\delta}_{Tn}) = (0\ \ 0\ \ 0\ \ 1)\text{var}(\hat{\beta} + \hat{\eta}_n - \beta - \eta_n)(0\ \ 0\ \ 0\ \ 1)^{\mathrm{T}}. \tag{10}$$

Note that $\hat{\eta}_n$ is the best linear unbiased predictor (BLUP) and can be derived as an empirical Bayes estimator [20, 21]. When T is completely missing in the nth trial, the expression of $E(\hat{\delta}_{Tn})$ in (9) is exactly the same as the GPHC estimate in (6). Different from GPHC and BMBRG, the variance formula in (10) accounts for the uncertainty associated with estimating β, but it treats D and σ as known quantities. In the implementations, we obtain these estimates using PROC MIXED in SAS.

3.4. *Empirical Bayes estimate and conditional posterior variance* (*EB-CPV*)

Let r be the number of patients in the new trial on whom we have information on both S and T. The empirical Bayesian estimate of δ_{Tn} can be obtained as the posterior mode estimate when we assume flat priors for the fixed effects and multivariate normal priors for the random effects [21]. Its expression is identical to the HD estimate in equation (9) [21]. When β, D, and σ are known, the conditional posterior variance (CPV) of δ_{Tn} can approximate the variance of $\hat{\delta}_{Tn}$ [22]. We obtain the CPV of δ_{Tn} as (details in Appendix C)

$$\text{var}(\delta_{Tn}) = (0\ \ 1)(\Psi_d^{-1} + \Phi_e^{-1})^{-1}(0\ \ 1)^{\mathrm{T}}, \tag{11}$$

where Ψ_d is a function only of the between-trial covariances given by $\Psi_{11} - \Psi_{12}\Psi_{22}^{-1}\Psi_{21}$, and Φ_e is a function only of the within-trial covariances given by $\Phi_e = \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{12} & \phi_{22} \end{pmatrix}$. The elements of Ψ_d and Φ_e are listed below:

$$\Psi_{11} = \begin{pmatrix} d_{tt} & d_{tr} \\ d_{tr} & d_{rr} \end{pmatrix}, \quad \Psi_{12} = \begin{pmatrix} d_{st} & d_{ta} \\ d_{sr} & d_{ar} \end{pmatrix}, \quad \Psi_{21} = \begin{pmatrix} d_{st} & d_{sr} \\ d_{sr} & d_{ar} \end{pmatrix},$$

$$\Psi_{22} = \begin{pmatrix} d_{ss} & d_{sa} \\ d_{sa} & d_{aa} \end{pmatrix}, \quad \phi_{11} = \frac{(\sigma_{tt} - \sigma_{st}^2\sigma_{ss}^{-1})\sum_{j=1}^{r} Z_{nj}^2}{r\sum_{j=1}^{r} Z_{nj}^2 - (\sum_{j=1}^{r} Z_{nj})^2},$$

$$\phi_{12} = \frac{(\sigma_{tt} - \sigma_{st}^2\sigma_{ss}^{-1})\sum_{j=1}^{r} Z_{nj}}{r\sum_{j=1}^{r} Z_{nj}^2 - (\sum_{j=1}^{r} Z_{nj})^2}, \quad \phi_{22} = \frac{r(\sigma_{tt} - \sigma_{st}^2\sigma_{ss}^{-1})}{r\sum_{j=1}^{r} Z_{nj}^2 - (\sum_{j=1}^{r} Z_{nj})^2}.$$

When T is completely missing in the nth trial, i.e. $r = 0$, the CPV simplifies to

$$\text{var}(\delta_{Tn}) = (0\ \ 1)\Psi_d(0\ \ 1)^{\mathrm{T}}, \tag{12}$$

an expression equivalent to the BMBRG variance formula in (5). The CPV formula can be viewed as the generalization of the BMBRG variance formula. Note that the CPV underestimates the prediction variance because they treat β, D, σ, a_{0n}, and a_{1n} as known quantities. Morris [23] and Ghosh and Rao [22] showed that a better estimator of the prediction variance can be obtained by adding to the CPV a second term that takes into account the uncertainty about all parameters.

3.5. *Bayesian estimation* (*denoted by Bayes*)

An alternative method to obtain the distributions of the parameters of interest is a fully Bayesian estimation method which is also applicable when T is either partially missing or completely missing. We assume that flat priors for the fixed effects, i.e. $p(\alpha_0) \propto 1$, $p(\gamma_0) \propto 1$, $p(\alpha_1) \propto 1$, and $p(\gamma_1) \propto 1$, and vague priors for the rest of parameters, specifically, $\sigma^{-1} \sim W(a, E)$ and $D^{-1} \sim W(c, F)$, where W refers to the Wishart distribution. We use $a = 3$, $c = 5$, $E = (a+1)^{-1}I_2$ and $F = (c+1)^{-1}I_4$. A data augmentation method is used to implement the procedure (details in Appendix B). The Bayesian estimation method naturally takes into consideration the uncertainty associated with estimating every parameter [24], but it can be sensitive to the prior specifications. While it is computationally intensive to conduct extensive simulations to evaluate the properties of this method, it is very feasible to analyze data using this method.

4. Efficiency gain and correlation

In this section, we study the precision of the predicted treatment effects ($\hat{\delta}_{Tn}$) and the factors that impact the precision, particularly, the correlation between S and T and the fraction of missingness.

4.1. Correlation

In a multiple-trial setting, with a bivariate mixed model assumption, the treatment adjusted individual-level or within-trial correlation between S and T is $R^2_{\text{indiv}}=\sigma^2_{st}/\sigma_{ss}\sigma_{tt}$. The trial-level correlation between S and T is defined by Buyse *et al.* [9] as

$$R^2_{\text{trial}}=\frac{(d_{sr}\ \ d_{ar})\begin{pmatrix}d_{ss} & d_{sa}\\ d_{sa} & d_{aa}\end{pmatrix}^{-1}\begin{pmatrix}d_{sr}\\ d_{ar}\end{pmatrix}}{d_{rr}}.$$

The between-trial correlation R^2_{trial} assesses how well the treatment effect on T in the new trial can be predicted by that on S. While R^2_{trial} is identified as the key factor that impacts the degree of efficiency gain from S in the research by Buyse *et al.* [9] and Gail *et al.* [17], as we shall see in the following, R^2_{indiv} plays an even more important role than R^2_{trial} in obtaining substantial efficiency gain from S with respect to the estimated treatment effect on T when T is partially observed.

4.2. Prediction precision and correlation

We examine the impacts of R^2_{indiv} and R^2_{trial} on the prediction precision using the CPV formula in equation (11). We note that when there is an equal number of patients per treatment group in the new trial, the elements of Φ_e in CPV simplify to

$$\phi_{11}=\frac{2\sigma_{tt}(1-R^2_{\text{indiv}})}{r},\quad \phi_{12}=\frac{2\sigma_{tt}(1-R^2_{\text{indiv}})}{r},\quad \phi_{22}=\frac{4\sigma_{tt}(1-R^2_{\text{indiv}})}{r}.$$

When T is completely missing in the new trial, the CPV simplifies to $\text{var}(\delta_{Tn})=d_{rr}(1-R^2_{\text{trial}})$; hence, the factors that determine the precision of the predictor of the treatment effect on T are R^2_{trial} and d_{rr} which are between-trial level. When T is partially observed, the additional important factors are within-trial level including R^2_{indiv}, σ_{tt}, and r. Since the within-trial covariances in Φ_e are usually significantly smaller than the between-trial covariances in Ψ_d, we find that Φ_e usually dominates and Ψ_d has a negligible impact on the CPV. Although the CPV usually underestimates the prediction variance, our simulation studies show that it usually accounts for the majority of the total variance, and a comparison between (11) and (12) should suffice to provide algebraic intuition about the prediction variance.

5. Simulations

5.1. The setup

We conduct simulation studies to evaluate the bias, efficiency, and coverage rates (CR) of the confidence intervals for the predicted treatment effect in a new trial using the above methods. For comparison purposes, we also estimate δ_{Tn} based on observed T using the simple estimate without any distributional assumption (denoted by SIMPLE). That is, $\hat{\delta}_{Tn}=\sum_k T_{nk1}/m_{n1}-\sum_l T_{nl0}/m_{n0}$, where T_{nk1} represents T on patient k in the $Z=1$ group in the nth trial and similarly for T_{nl0}, m_{n1} represents the number of patients in the $Z=1$ group in the nth trial and similarly for m_{n0}.

We generate 500 data sets based on the bivariate mixed model in (1). We assume equal number of patients per trial and let $m_i=m$. The parameter specifications are $\beta^{\text{T}}=(1,2,1,1)$, $d_{ss}=0.5$, $d_{tt}=0.2$, $d_{aa}=3.5$, $d_{rr}=1.6$, $\sigma_{ss}=1$, and $\sigma_{tt}=0.3$. To examine the impact of the trial-level correlation, we vary the correlation matrices for the random effects

$$\begin{pmatrix}1 & 0.57 & 0.37 & 0.22\\ 0.57 & 1 & 0.24 & 0.21\\ 0.37 & 0.24 & 1 & 0.3\\ 0.22 & 0.21 & 0.3 & 1\end{pmatrix},\ \begin{pmatrix}1 & 0.57 & 0.37 & 0.22\\ 0.57 & 1 & 0.24 & 0.21\\ 0.37 & 0.24 & 1 & 0.7\\ 0.22 & 0.21 & 0.7 & 1\end{pmatrix}\ \text{and}\ \begin{pmatrix}1 & 0.57 & 0.37 & 0.22\\ 0.57 & 1 & 0.24 & 0.21\\ 0.37 & 0.24 & 1 & 0.9\\ 0.22 & 0.21 & 0.9 & 1\end{pmatrix},$$

which correspond to $R^2_{\text{trial}}=0.1, 0.5$, and 0.8, respectively. To examine the impact of the individual-level correlation, we vary R^2_{indiv} from 0.1, 0.5, to 0.9. We vary n, m, and the percentage of missingness in the new trial (denoted by p). For each different data sets, we have a different underlying true treatment effect δ_{Tn} because δ_{Tn} is not fixed and follows a known distribution. Its average across 500 data sets is denoted by $\bar{\delta}_{Tn}$. For each data set and each method used, we obtain $\hat{\delta}_{Tn}$, its standard error (SE), its CI as $\hat{\delta}_{Tn}\pm 1.96\times(\text{SE})$ and an indicator variable for whether the 95 per cent CI

Table I. Simulation results based on 500 data sets.

n	m	Per cent missing	Methods	Bias	RMSE	SE	CR
10	100	0	SIMPLE	−0.005	0.111	0.109	95.0
			HD	−0.005	0.112	0.106	94.8
			Bayes	−0.006	0.112	0.109	94.8
			EB-CPV	−0.005*	0.112*	0.093†	90.8
		50	SIMPLE	−0.002	0.149	0.139	95.4
			HD	0.000	0.143	0.139	94.0
			Bayes	−0.004	0.145	0.145	94.8
			EB-CPV	0.000*	0.143*	0.132†	91.2
		100	BMBRG	−0.013	1.193	0.736	76.6
			GPHC	−0.010	1.125	0.755	80.2
			HD	−0.011	1.125	0.795	82.0
			Bayes	−0.037	1.139	1.105	94.8
40	100	0	SIMPLE	−0.005	0.111	0.109	95.0
			HD	−0.005	0.110	0.108	95.0
			EB-CPV	−0.005*	0.110*	0.093†	91.4
		50	SIMPLE	−0.002	0.149	0.155	95.4
			HD	0.003	0.140	0.142	95.8
			EB-CPV	0.003*	0.140*	0.130†	93.2
		100	BMBRG	−0.008	0.965	0.887	92.8
			GPHC	−0.008	0.965	0.877	92.6
			HD	−0.008	0.965	0.869	92.4
40	300	0	SIMPLE	0.002	0.063	0.063	94.2
			HD	0.002	0.063	0.063	94.4
			EB-CPV	0.002*	0.063*	0.066†	90.9
		50	SIMPLE	−0.002	0.089	0.090	94.4
			HD	0.000	0.085	0.083	93.6
			EB-CPV	0.000*	0.085*	0.077†	91.4
		100	BMBRG	−0.002	0.920	0.868	93.8
			GPHC	−0.002	0.919	0.871	93.8
			HD	−0.002	0.919	0.882	94.0
55	100	0	SIMPLE	−0.005	0.111	0.109	95.0
			HD	−0.005	0.110	0.108	95.0
			EB-CPV	−0.005*	0.110*	0.094†	90.8
		50	SIMPLE	−0.002	0.149	0.155	95.4
			HD	0.002	0.140	0.142	95.8
			EB-CPV	0.002*	0.140*	0.131†	93.4
		100	BMBRG	−0.033	0.950	0.883	93.8
			GPHC	−0.033	0.948	0.898	94.0
			HD	−0.033	0.948	0.898	94.2

$\beta^{\mathrm{T}}=(1,2,1,1)$, $d_{ss}=0.5$, $d_{tt}=0.2$, $d_{aa}=3.5$, $d_{rr}=1.6$, $\sigma_{ss}=1$, $\sigma_{tt}=0.3$, $R^2_{\mathrm{indiv}}=0.5$ and $R^2_{\mathrm{trial}}=0.5$. SIMPLE: simple estimate. HD: Henderson method. Bayes: Bayesian estimation. CPV: conditional posterior variance. BMBRG: method by Buyse *et al.* [9]. GPHC: method by Gail *et al.* [17]. EB-CPV: EB estimate with CPV: variance. *: obtained using HD. †: obtained using CPV.

contains δ_{Tn} or not. Let $\bar{\hat{\delta}}_{Tn}$ denote the average of $\hat{\delta}_{Tn}$ across 500 data sets. We examine the method's performance by its average bias (Bias = $\bar{\hat{\delta}}_{Tn}-\bar{\delta}_{Tn}$), the average SE, the root mean squared error (RMSE = $\sqrt{\sum(\hat{\delta}_{Tn}-\delta_{Tn})^2/500}$), and the CR over all simulated data sets. As we will see all estimates are unbiased, the relative efficiency (RE) of two estimators can be approximated by the inverse of the ratio of the two corresponding RMSE^{2}s.

5.2. *Method evaluation*

In Table I, we present Bias, RMSE, SE, and CR of $\hat{\delta}_{Tn}$ using the respective methods including SIMPLE, HD, BMBRG, GPHC, Bayes, and EB-CPV from simulations with various combinations of n, m, and the percentage of missingness. We let $R^2_{\mathrm{indiv}}=0.5$ and $R^2_{\mathrm{trial}}=0.5$. When T is completely or 50 per cent observed in the new trial, SIMPLE and HD generate estimates which are unbiased, have similar RMSE and confidence intervals with nominal-level or close-to-nominal-level CR; on the other hand, CPV consistently gives underestimated prediction variances (i.e. SE < RMSE).

When T is completely missing in the new trial, BMBRG, GPHC, and HD all underestimate the variances of $\hat{\delta}_{Tn}$. When the number of the trials is relatively large ($n=40, 55$), the extent of underestimation is minor; however, with a small number of trials ($n=10$), the extent can be more severe and the CRs can be less than 85 per cent. Although HD is expected to have better CR than GPHC and GPHC is expected to be better than BMBRG because they account for more uncertainty of the parameters, the advantages of HD and GPHC over BMBRG are small and all methods give similar CRs. The Bayes method we used gives more precise estimates of the variances and the CRs are around the 95 per cent nominal level.

The SIMPLE and HD methods give estimates with similar precision which shows that the efficiency gain from the bivariate normal assumption is small. When T is partially or completely observed, the increase in m can improve the precision of the estimates while a larger n does not necessarily improve much precision. When T is completely missing, there is a minor gain in the precision when n and m increase.

5.3. R^2_{indiv}, R^2_{trial}, *percentage of missingness and information recovery from* S

Figure 1(a) shows the RE of $\hat{\delta}_{Tn}$ when T is completely missing in the new trial compared with the estimate before any deletion in T occurs using the HD method. RE is defined as the inverse of the ratio of the two variances. We vary R^2_{indiv} and R^2_{trial} and let $n=40$ and $m=100$. We find that while the increases in R^2_{indiv} have negligible impact on the precision, the increase in R^2_{trial} can improve the precision more than any other factor. These findings agree with the algebraic intuition from the CPV variance formula in (12). Relative to the estimate based on completely observed data, the RE varies from 0.7, 1.2 to 3.4 per cent as we increase R^2_{trial} from 0.1, 0.5 to 0.8. As a result, when we completely rely on S and summary statistics from previous trials to predict $\hat{\delta}_{Tn}$, the extent of information recovery is often limited and the precision of $\hat{\delta}_{Tn}$ is usually insufficient to be clinically useful.

Figure 1(b) presents the RE of $\hat{\delta}_{Tn}$ when T is 50 per cent missing compared with the estimate before any deletion of T using the HD method. We find that high R^2_{indiv} can lead to a large gain of efficiency from the use of S. When R^2_{indiv} is large (e.g. 0.7 or 0.9), most of the information on δ_{Tn} is recovered from S and the precision of the estimate is close to that when T is completely observed. On the other hand, the magnitude of R^2_{trial} does not have much impact on the amount of efficiency gain from S. The observations here are in agreement with the CPV variance formula in (11).

Figure 1(c) shows the RE of $\hat{\delta}_{Tn}$ when T is partially or completely missing compared with the estimate before any deletion of T. Naturally, the higher the proportion of available T, the smaller the RMSE, and thus the greater the precision for the treatment effect prediction. Interestingly, we find that there is a substantial efficiency gain from the information on S with even a small fraction of observed T, particularly when R^2_{indiv} is high. For example, when 30 per cent T are observed, the lost information due to missingness is almost completely recovered from S when $R^2_{\text{indiv}}=0.9$.

6. Data analysis: a glaucoma study

The evaluation of the extent of information recovery from S in predicting the treatment effect on T in a new trial is illustrated using the Collaborative Initial Glaucoma Treatment Study (CIGTS) [25]. Glaucoma is a group of diseases that cause vision loss and is a leading cause of blindness. High pressure in the eyes, i.e. intraocular pressure (IOP), is a major risk factor of glaucoma. The CIGTS is a randomized multi-center clinical trial to compare the effects of two types of treatments, surgery and medicine, on reducing IOP among glaucoma patients. Patients are enrolled between 1993 and 1997. A total of 607 patients are included in the study and among them, 307 are randomly assigned into the medicine group. IOP (recorded in mmHg) has been measured at different time points following the treatment. For the purpose of this paper, we take the IOP measurement at month 96 as T and that at month 12 as S. We assume that the IOP measurements are normally distributed. To evaluate the situation of a meta-analysis where data are from different trials; we treat the different centers in the CIGTS study as independent trials testing a similar group of treatments. A preliminary analysis of these data shows that the estimate of the between-trial variances, $\hat{D}$, is nonpositive definite. Mimicing the approach of Gail *et al.* [17], we rescale up the data size by simulating S_{ij} and T_{ij} from bivariate normal distributions for each trial and treatment group with the trial-specific and treatment-specific means and variance–covariances from the real data. Nonetheless, our results are generalizable. The CIGTS study includes 14 centers from which we delete five centers (i.e. 5, 7, 12, 13, 14) either because they had too few observations or because of nonpositive definite covariance matrices within center. We also deleted two outliers that are greater than 35 mmHg. For the centers included ($n=9$), we increase the sample sizes to 335, 176, 385, 264, 539, 368, 286, 528, and 319. The trial-specific and treatment-specific means and correlations for S and T are listed in Table II.

The HD method is used to fit the rescaled data for which $\hat{D}$ is positive definite and the estimates of R^2_{trial} and R^2_{indiv}, denoted by $\hat{R}^2_{\text{trial}}$ and $\hat{R}^2_{\text{indiv}}$, are obtained as 0.25 and 0.15, respectively. We randomly select Center 8 as the new trial

Statistics in Medicine

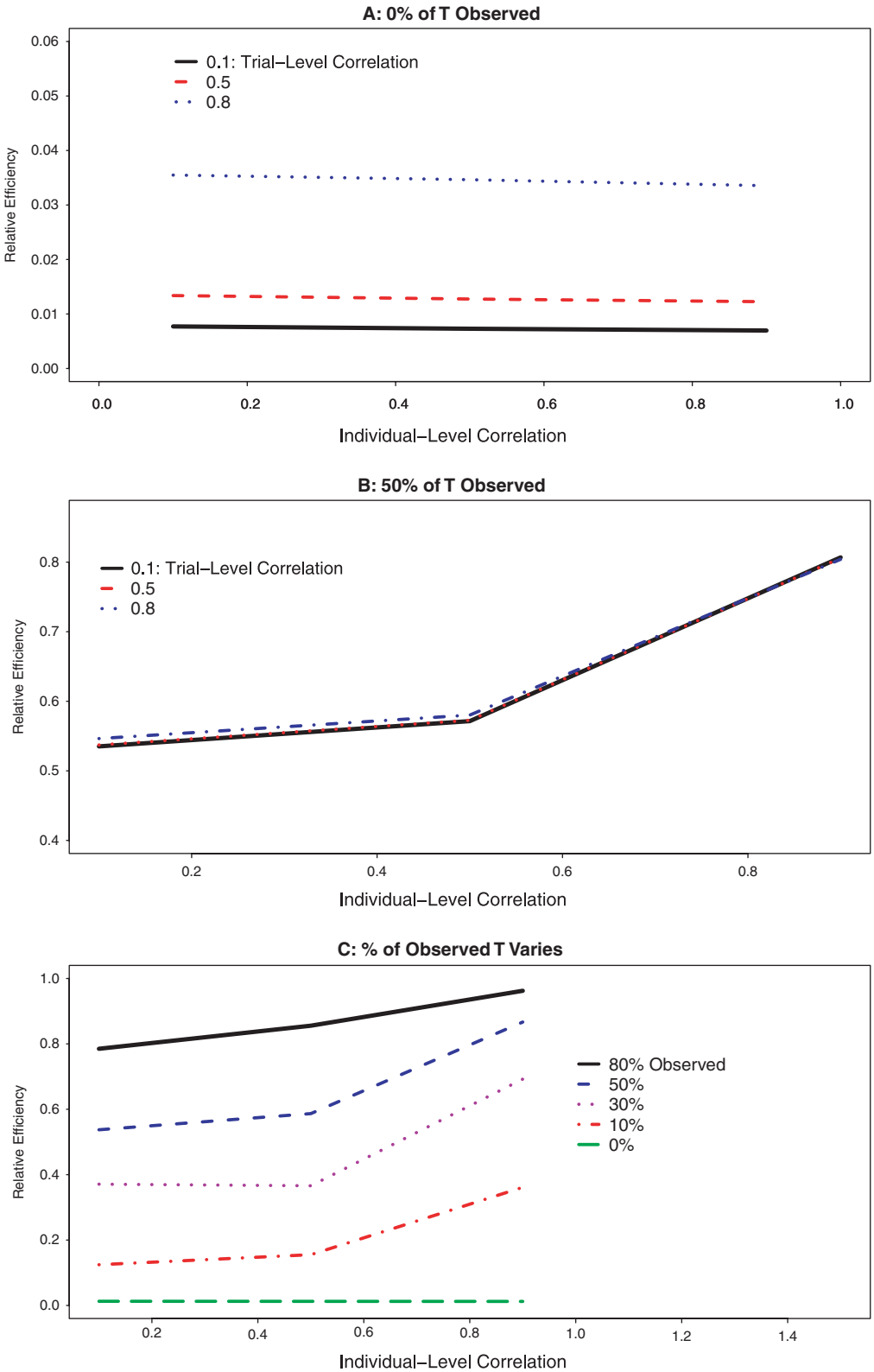

Figure 1. Simulation results based on 500 data sets. Relative efficiency of the new treatment effect estimate using S when T is not completely observed to that when T is completely observed. A: 0 per cent of T observed in the new trial; B: 50 per cent of T observed in the new trial; C: Percentage of observed T varies in the new trial.

Table II. Description of pseudodata in glaucoma study: treatment-specific means and individual-level correlations for each center.

		Medicine	Surgery	Individual-level correlation	
Center	Sample size	(Means of S, T)	(Means of S, T)	Medicine	Surgery
1	670	(17.63, 16.52)	(13.76, 14.59)	0.367	0.608
2	352	(17.22, 16.42)	(14.63, 12.98)	−0.455	0.467
3	770	(19.27, 17.58)	(15.81, 16.17)	0.589	0.548
4	528	(17.17, 15.51)	(10.93, 12.88)	0.176	0.540
5	1078	(18.52, 18.67)	(14.99, 15.32)	0.435	0.407
6	736	(18.62, 18.89)	(15.13, 17.11)	−0.16	−0.0056
7	572	(18.35, 15.34)	(14.59, 14.53)	0.177	0.396
8	1056	(18.59, 16.16)	(13.60, 13.72)	0.31	0.95
9	638	(17.56, 16.82)	(14.19, 14.61)	0.042	0.756

Table III. Estimate treatment effect on IOP at the 96th month utilizing information from early IOP measures at the 12th month in the glaucoma study.

p	Estimate	Standard error	p-Value
Center = 8			
SIMPLE*	−2.45	0.29	<0.0001
No missing†	−2.33	0.22	<0.0001
100 per cent missing†	−1.58	0.79	0.063
90 per cent missing†	−1.50	0.47	0.0059
80 per cent missing†	−2.37	0.39	<0.0001
50 per cent missing†	−2.61	0.29	<0.0001
20 per cent missing†	−2.19	0.23	<0.0001
Center = 9			
SIMPLE*	−2.21	0.30	<0.0001
No missing†	−2.32	0.27	<0.0001
100 per cent missing†	−2.68	0.82	0.0053
90 per cent missing†	−2.19	0.61	0.0023
80 per cent missing†	−2.30	0.49	<0.0002
50 per cent missing†	−2.04	0.36	<0.0001
20 per cent missing†	−2.15	0.30	<0.0001

*: Based on complete data before any deletion. †: HD method was used.

and delete some proportion of T in Center 8 to examine the extent of efficiency gain through the use of S. The missing mechanism is missing completely at random [26]. The results are listed in Table III. Without missing T, $\hat{\delta}_{Tn}$ is −2.45 with the SE of 0.29. When T is completely missing, $\hat{\delta}_{Tn}$ is −1.58 with the SE of 0.79. When 20 or 50 per cent of T are missing, the precision of $\hat{\delta}_{Tn}$ using S is comparable to that based on completely observed T. Even with 80 per cent missing, the SE is substantially smaller than that when 100 per cent of T is missing. For further illustration, we treat Center 9 as a new trial and obtain similar results. With this rescaled data, we have artificially increased the sample size by approximately five-fold for each trial, hence, the power to detect the treatment effect is much larger than the original data. When δ_{Tn} is predicted solely based on S, the RE is only about 10 per cent compared with that when T is not missing, and $\hat{\delta}_{Tn}$ reaches the significance level of 0.05 for Center 9 and is not quite significant for center 8. In this particular study we completely observe S by the end of year 1998 but only start to observe T in year 2001, thus by solely relying on S to predict δ_{Tn}, we can significantly shorten the trial length, but the result is only of borderline significance. In practice many trials do not have such a strong effect so when δ_{Tn} is predicted solely based on S, the substantial loss in precision often results in failure to detect any real treatment effect difference. In the CIGTS, by October 2002, about 20 per cent of T would have been observed and the treatment effect is clearly significant, illustrating the benefit of significant increase in the precision of $\hat{\delta}_{Tn}$ by utilizing a small fraction of T. There is a also a considerable time saving compared with collecting T on all subjects, which would have required follow-up to 2005.

7. Discussion

In this report, we examine the role of biomarkers as auxiliary variables in predicting the treatment effect and identify situations when biomarkers can be beneficial in a multiple-trial setting. While previous literature on the use of biomarkers as substitutes for the true endpoints has been mostly negative and the proposed surrogate measures are often not useful in practice, we show that it is possible for S to be useful as auxiliary variables in helping provide information and enhancing the inference on T. Although a high correlation between S and T does not qualify S as a good surrogate [12], we show that the correlation is a critical measure in determining the extent of information recovery from S.

In a multiple-trial setting, when T is completely unobserved, R^2_{indiv} has little impact on the amount of information recovered from S; on the other hand, the higher the R^2_{trial}, the higher the efficiency gain from S. However, even with a relatively high R^2_{trial}, the predicted treatment effect based on data from other trials and biomarkers in the new trial solely is usually too imprecise to be clinically useful. On the other hand, when the predicted treatment effect on T solely based on S would be sufficient to detect the difference in the treatment effect, the benefit of reducing the trial length can be enormous. Examples include the situation when the statistical power to detect treatment effect is very large or when R^2_{trial} is close to 1 such as the ovarian cancer example in [9]. However, these cases are usually rare in practice. On the contrary, when T is partially observed in the new trial, we find that a high R^2_{indiv} is a very important determinant in increasing the precision of the predicted treatment effect from S but the impact of R^2_{trial} is negligible. With even a small fraction of T and a high R^2_{indiv}, the information on the treatment effect is mostly recovered and the prediction precision is close to that when T is completely observed. It appears that some data on T are essential to provide the basis for individual-level predictions of T from S and take advantage of the distributional assumption between S and T, and hence to give a much more efficient treatment estimate.

We compare the BMBRG, GPHC, and HD methods when T is completely missing. Each method gave unbiased estimates; but the variances were underestimated, particularly when the number of the trials was small. Either a bootstrap [17] or fully Bayesian or measurement-error approach [27] could remedy this problem. When T is partially observed, we use two methods: HD and EB-CPV. We find that the underestimation of the variance from the HD method becomes negligible but CPV consistently underestimates the variance. We note that we only consider the case of missing T being missing completely at random and that all methods are applicable when the missing mechanism is missing at random [26].

In conclusion, biomarkers would seem to have a useful role as auxiliary variables. Future research should focus on their roles as auxiliary variables and identify scenarios when biomarkers can increase the precision of the treatment effect. For design purposes, our results suggest that it is often important to collect at least some data on the true endpoint and more information on biomarkers which have high adjusted individual-level correlations with the true endpoint. With appropriate utilization of high-quality biomarkers in estimating the treatment effect when the true endpoint is not completely observed, one can reach a desired level of precision earlier, hence shortening the study period and reducing the cost. In our study, we consider continuous S and T. For future research, it would also be interesting to investigate the factors that impact the efficiency gain and the extent of it when S and T are other types of data such as binary, categorical, and time to an event.

Appendix A: Henderson method [18]

Let $Y=X\beta+U\eta+\varepsilon$, where the vectors Y, η, and ε and the matrix X are obtained from stacking the vectors Y_i, η_i, and ε_i and the matrices X_i, respectively, underneath each other, and where U is the block-diagonal matrix with blocks U_i on the main diagonal and zeros elsewhere. Let $\mathscr{D}$ and Σ be block-diagonal with blocks D and Σ_i on the main diagonal and zeros elsewhere. We have the following relationships: $E(\eta)=0$, $E(\varepsilon)=0$, $\text{var}(\eta)=\mathscr{D}$, $\text{var}(\varepsilon)=\Sigma$, $\text{cov}(\eta,\varepsilon)=0$ and we let $\mathscr{V}=U\mathscr{D}U^{\mathrm{T}}+\Sigma$. The estimates of $\mathscr{D}$, Σ and $\mathscr{V}$ are denoted by $\hat{\mathscr{D}}$, $\hat{\Sigma}$, and $\hat{\mathscr{V}}$. Henderson [18] proposed a method to obtain estimates of β and η by solving the mixed model equation as follows:

$$\begin{bmatrix} X^{\mathrm{T}}\hat{\Sigma}^{-1}X & X^{\mathrm{T}}\hat{\Sigma}^{-1}U \\ U^{\mathrm{T}}\hat{\Sigma}^{-1}X & U^{\mathrm{T}}\hat{\Sigma}^{-1}U+\hat{\mathscr{D}}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\eta} \end{bmatrix} = \begin{bmatrix} X^{\mathrm{T}}\hat{\Sigma}^{-1}Y \\ U^{\mathrm{T}}\hat{\Sigma}^{-1}Y \end{bmatrix}.$$

The solution can be written as:

$$\hat{\beta}=(X^{\mathrm{T}}\hat{\mathscr{V}}^{-1}X)^{-1}X^{\mathrm{T}}\hat{\mathscr{V}}^{-1}Y,$$

$$\hat{\eta}=\hat{\mathscr{D}}U^{\mathrm{T}}\hat{\mathscr{V}}^{-1}(Y-X\hat{\beta}).$$

The covariance matrix of $(\hat{\beta}-\beta, \hat{\eta}-\eta)$ is

$$C=\begin{bmatrix} X^{\mathrm{T}}\Sigma^{-1}X & X^{\mathrm{T}}\Sigma^{-1}U \\ U^{\mathrm{T}}\Sigma^{-1}X & U^{\mathrm{T}}\Sigma^{-1}U+\mathscr{D}^{-1} \end{bmatrix}^{-1}.$$

McLean and Sanders [28] and McLean *et al*. (1991) [29] show that C can also be written as

$$C=\begin{bmatrix} C_{11} & C_{21} \\ C_{21} & C_{22} \end{bmatrix}^{-1},$$

where

$$\begin{aligned} C_{11} &= (X^{\mathrm{T}}\mathscr{V}^{-1}X)^{-1}, \\ C_{21} &= -\mathscr{D}U^{\mathrm{T}}\mathscr{V}^{-1}XC_{11}, \\ C_{22} &= (U^{\mathrm{T}}\Sigma^{-1}U+\mathscr{D}^{-1})^{-1}-C_{21}X^{\mathrm{T}}\mathscr{V}^{-1}U\mathscr{D} \\ &= \mathscr{D}-\mathscr{D}U^{\mathrm{T}}\mathscr{V}^{-1}U\mathscr{D}-C_{21}X^{\mathrm{T}}\mathscr{V}^{-1}U\mathscr{D}. \end{aligned}$$

In practice, the estimate, $\hat{C}$, is often obtained by substituting $\mathscr{D}$ and Σ in C with their estimates, as we have done in this paper. From the above, we can obtain the expression for the mean and variance for $\hat{\beta}+\hat{\eta}_n$ as follows:

$$\begin{aligned} E(\hat{\beta}+\hat{\eta}_n) &= \beta+DU_n^{\mathrm{T}}V_n^{-1}(Y_n-X_n\beta), \operatorname{var}(\hat{\beta}-\beta+\hat{\eta}_n-\eta_n)=(X^{\mathrm{T}}\mathscr{V}^{-1}X)^{-1}+D-DU_n^{\mathrm{T}}V_n^{-1}U_nD+DU_n^{\mathrm{T}}V_n^{-1}X_n \\ &\quad \times(X^{\mathrm{T}}\mathscr{V}^{-1}X)^{-1}X_n^{\mathrm{T}}V_n^{-1}U_nD-2DU_n^{\mathrm{T}}V_n^{-1}X_n(X^{\mathrm{T}}\mathscr{V}^{-1}X)^{-1} \\ &= \left(\sum_{i=1}^{n}X_i^{\mathrm{T}}V_i^{-1}X_i\right)^{-1}+D-DU_n^{\mathrm{T}}V_n^{-1}U_nD \\ &\quad +DU_n^{\mathrm{T}}V_n^{-1}X_n\left(\sum_{i=1}^{n}X_i^{\mathrm{T}}V_i^{-1}X_i\right)^{-1}X_n^{\mathrm{T}}V_n^{-1}U_nD-2DU_n^{\mathrm{T}}V_n^{-1}X_n\left(\sum_{i=1}^{n}X_i^{\mathrm{T}}V_i^{-1}X_i\right)^{-1}. \end{aligned}$$

Appendix B: Bayesian estimation

Iterate the following two steps until the parameters reach convergence:
Step 1: Impute missing T_{nj}'s from a normal distribution with mean and variance:

$$\begin{aligned} E(T_{nj}|S_{nj}, Z_{nj}) &= \gamma_0+r_{0n}-\sigma_{st}\sigma_{ss}^{-1}(\alpha_0+a_{0n})+(\gamma_1+r_{1n}-\sigma_{st}\sigma_{ss}^{-1}(\alpha_1+a_{1n}))Z_{nj}+\sigma_{st}\sigma_{ss}^{-1}S_{nj}, \\ \operatorname{var}(T_{nj}|S_{nj}, Z_{nj}) &= \sigma_{tt}-\sigma_{st}^2\sigma_{ss}^{-1}. \end{aligned}$$

Step 2: Apply the Gibbs sampling to the complete data to estimate the parameters

$$\begin{aligned} D^{-1}|\eta &\sim W\left(n+c, \left(\sum_{i=1}^{n}\eta_i\eta_i^{\mathrm{T}}+F^{-1}\right)^{-1}\right), \\ \sigma^{-1}|X, Y, Z, \beta, \eta &\sim W\left(\sum_{i=1}^{n}m_i+a, (VS+E^{-1})^{-1}\right), \\ \eta_i|X, Y, Z, \sigma, D &\sim \mathrm{MVN}\left(VE\times\left(\sum_{j=1}^{m_i}Z_{ij}^{\mathrm{T}}\sigma^{-1}(Y_{ij}-X_{ij}\beta)\right), VE\right), \\ \beta|X, Y, Z, \eta_i, \sigma &\sim \mathrm{MVN}\left(VB\times\left(\sum_{i=1}^{n}\sum_{j=1}^{m_i}X_{ij}\sigma^{-1}(Y_{ij}-U_{ij}\eta_i)\right), VB\right), \end{aligned}$$

where

$$VS=\sum_{i=1}^{n}\sum_{j=1}^{m_i}(Y_{ij}-X_{ij}\beta-U_{ij}\eta_i)(Y_{ij}-X_{ij}\beta-U_{ij}\eta_i)^{\mathrm{T}},$$

$$VB=\left(\sum_{i=1}^{n}\sum_{j=1}^{m_i}X_{ij}^{\mathrm{T}}\sigma^{-1}X_{ij}\right)^{-1},$$

$$VE=\left(\sum_{j=1}^{m_i}U_{ij}^{\mathrm{T}}\sigma^{-1}U_{ij}+D^{-1}\right)^{-1}.$$

From the distributions of β and η_i, we can obtain the distribution of δ_{Tn}.

Appendix C: conditional posterior variance of δ_{Tn}

Let $\alpha_0+a_{0i}=\mu_{0Si}$, $\gamma_0+r_{0i}=\mu_{0Ti}$, $\alpha_1+a_{1i}=\delta_{Si}$ and $\gamma_1+r_{1i}=\delta_{Ti}$. We can rewrite the model (1) as

$$S_{ij}=\mu_{0Si}+\delta_{Si}Z_{ij}+\varepsilon_{Sij},$$

$$T_{ij}=\mu_{0Ti}+\delta_{Ti}Z_{ij}+\varepsilon_{Tij}.$$

Assume that there are r observations with both S and T observed and m_n-r observations with just S observed in the nth trial. The likelihood can be written as

$$L(\phi|S,T,Z)=\left\{\prod_{i=1}^{n-1}\left[\prod_{j=1}^{m_i}N(Y_{ij}|\beta,\eta_i,\sigma,D,Z_{ij})\right]\right\}\left\{\prod_{j=1}^{r}N(Y_{nj}|\eta_n,\beta,Z_{nj},\sigma,D)\prod_{j=1}^{m_n-r}N(S_{nj}|\mu_{0sn},\delta_{sn},Z_{nj},\sigma_{ss})\right\}$$

$$=\prod_{i=1}^{n-1}\left[\prod_{j=1}^{m_i}\frac{1}{\sqrt{2\pi}|\sigma|^{1/2}}\exp\left\{-\frac{1}{2}\begin{pmatrix}S_{ij}-\mu_{0Si}-\delta_{Si}Z_{ij}\\T_{ij}-\mu_{0Ti}-\delta_{Ti}Z_{ij}\end{pmatrix}^{\mathrm{T}}\sigma^{-1}\begin{pmatrix}S_{ij}-\mu_{0Si}-\delta_{Si}Z_{ij}\\T_{ij}-\mu_{0Ti}-\delta_{Ti}Z_{ij}\end{pmatrix}\right\}\right]$$

$$\times\prod_{j=1}^{r}\frac{1}{\sqrt{2\pi}|\sigma|^{1/2}}\exp\left\{-\frac{1}{2}\begin{pmatrix}S_{nj}-\mu_{0Sn}-\delta_{Sn}Z_{nj}\\T_{nj}-\mu_{0Tn}-\delta_{Tn}Z_{nj}\end{pmatrix}^{\mathrm{T}}\sigma^{-1}\begin{pmatrix}S_{nj}-\mu_{0Sn}-\delta_{Sn}Z_{nj}\\T_{nj}-\mu_{0Tn}-\delta_{Tn}Z_{nj}\end{pmatrix}\right\}$$

$$\times\prod_{j=1}^{m_n-r}\frac{1}{\sqrt{2\pi}\sigma_{ss}^{1/2}}\exp\left\{-\frac{1}{2}(S_{nj}-\mu_{0Sn}-\delta_{Sn}Z_{nj})^2\right\}$$

$$\times\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}|D|^{1/2}}\exp\left\{-\frac{1}{2}\begin{pmatrix}\mu_{0Si}-\alpha_0\\\mu_{0Ti}-\gamma_0\\\delta_{Si}-\alpha_1\\\delta_{Ti}-\gamma_1\end{pmatrix}^{\mathrm{T}}D^{-1}\begin{pmatrix}\mu_{0Si}-\alpha_0\\\mu_{0Ti}-\gamma_0\\\delta_{Si}-\alpha_1\\\delta_{Ti}-\gamma_1\end{pmatrix}\right\},$$

which is proportional to the posterior density when we assume flat priors for the fixed effects and multivariate normal distributions for the random effects. The conditional posterior distributions of μ_{0Tn} and δ_{Tn} given the data and all other parameters are proportional to

$$\begin{pmatrix}\mu_{0Tn}\\\delta_{Tn}\end{pmatrix}\Bigg|\cdot\propto\prod_{j=1}^{r}\frac{1}{\sqrt{2\pi}|\sigma|^{1/2}}\exp\left\{-\frac{1}{2}ME^{\mathrm{T}}\times\sigma^{-1}\times ME\right\}\frac{1}{\sqrt{2\pi}|D|^{1/2}}\exp\left\{-\frac{1}{2}\begin{pmatrix}\mu_{0Sn}-\alpha_0\\\mu_{0Tn}-\gamma_0\\\delta_{Sn}-\alpha_1\\\delta_{Tn}-\gamma_1\end{pmatrix}^{\mathrm{T}}D^{-1}\begin{pmatrix}\mu_{0Sn}-\alpha_0\\\mu_{0Tn}-\gamma_0\\\delta_{Sn}-\alpha_1\\\delta_{Tn}-\gamma_1\end{pmatrix}\right\}$$

$$\propto\exp\left\{-\frac{1}{2}\sum_{j=1}^{r}[T_{nj}-\mu_{0Tn}-\delta_{Tn}Z_{nj}-\sigma_{st}\sigma_{ss}^{-1}(S_{nj}-\mu_{0Sn}-\alpha_nZ_{nj})]^2\times q^{-1}\right\}$$

$$\times \exp\left\{-\frac{1}{2}MD^{\mathrm{T}} \times (\Psi_{11}-\Psi_{12}\Psi_{22}^{-1}\Psi_{21})^{-1} \times MD\right\}$$
$$= A \times B, \tag{C1}$$

where

$$\Psi_{11}=\begin{pmatrix} d_{tt} & d_{tr} \\ d_{tr} & d_{rr}\end{pmatrix}, \Psi_{12}=\begin{pmatrix} d_{st} & d_{ta} \\ d_{sr} & d_{ar}\end{pmatrix}, \Psi_{21}=\begin{pmatrix} d_{st} & d_{sr} \\ d_{ta} & d_{ar}\end{pmatrix}, \Psi_{22}=\begin{pmatrix} d_{ss} & d_{sa} \\ d_{sa} & d_{aa}\end{pmatrix},$$

$$ME=\begin{pmatrix} S_{nj}-\mu_{0Sn}-\delta_{Sn}Z_{nj} \\ T_{nj}-\mu_{0Tn}-\delta_{Tn}Z_{nj}\end{pmatrix}, MD=\begin{pmatrix} \mu_{0Tn}-\gamma_0 \\ \delta_{Tn}-\gamma_1\end{pmatrix}-\Psi_{12}\Psi_{22}^{-1}\begin{pmatrix} \mu_{0Sn}-\alpha_0 \\ \delta_{Sn}-\alpha_1\end{pmatrix}.$$

and $q=\sigma_{tt}-\sigma_{st}^2\sigma_{ss}^{-1}$.

The covariance contribution for μ_{0Tn} and δ_{Tn} from term B is $\Psi_d=\Psi_{11}-\Psi_{12}\Psi_{22}^{-1}\Psi_{21}$.

We define $Q_{nj}=T_{nj}-\sigma_{st}\sigma_{ss}^{-1}(S_{nj}-\mu_{0Sn}-\alpha_n Z_{nj})$. From (C1),

$$A=\exp\left\{-\frac{1}{2}\sum(Q_{nj}-\mu_{0Tn}-\delta_{Tn}Z_{nj})^2 q^{-1}\right\}$$
$$=\exp\left\{-\frac{1}{2}\left[\frac{\sum Z_{nj}^2}{q}\delta_{Tn}^2+\frac{r}{q}\mu_{0Tn}^2+\frac{\sum Q_{nj}^2}{q}-2\frac{\mu_{0Tn}\sum Q_{nj}}{q}-2\frac{\delta_{Tn}\sum Z_{nj}Q_{nj}}{q}+2\frac{\mu_{0Tn}\delta_{Tn}\sum Z_{nj}}{q}\right]\right\}.$$

A is proportional to a bivariate normal density. The covariance contribution from term A is defined as $\Phi_e=\begin{pmatrix}\phi_{11} & \phi_{12} \\ \phi_{12} & \phi_{22}\end{pmatrix}$, where

$$\phi_{11}=\frac{(\sigma_{tt}-\sigma_{st}^2\sigma_{ss}^{-1})\sum_{j=1}^r Z_{nj}^2}{r\sum_{j=1}^r Z_{nj}^2-(\sum_{j=1}^r Z_{nj})^2},$$

$$\phi_{12}=\frac{(\sigma_{tt}-\sigma_{st}^2\sigma_{ss}^{-1})\sum_{j=1}^r Z_{nj}}{r\sum_{j=1}^r Z_{nj}^2-(\sum_{j=1}^r Z_{nj})^2},$$

$$\phi_{22}=\frac{r(\sigma_{tt}-\sigma_{st}^2\sigma_{ss}^{-1})}{r\sum_{j=1}^r Z_{nj}^2-(\sum_{j=1}^r Z_{nj})^2}.$$

Combining the variance contributions from terms A and B, we can obtain the conditional posterior covariance for μ_{0Tn} and δ_{Tn} as: $(\Phi_e^{-1}+\Psi_d^{-1})^{-1}$. The corresponding conditional posterior variance for $\hat{\delta}_{Tn}-\delta_{Tn}$ is $(0 \quad 1)(\Phi_e^{-1}+\Psi_d^{-1})^{-1}(0 \quad 1)^{\mathrm{T}}$.

Acknowledgements

The authors thank Dr Brenda Gillespie for providing us with the CIGTS data. This research was supported by National Institutes of Health Grant CA129102.

References

1. Buyse M, Molenberghs G. Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics* 1998; **54**:1014–1029.
2. Baker SG, Kramer BS. Biomarker, Surrogate Endpoints, and Early Detection Imaging Tests: Reducing Confusion. Available from: http://www.icsa.org/bulletin/Bulletin-1-2004-Contents/A3-25-controverstial-issues-v4.doc.
3. Prentice RL. Surrogate endpoints in clinical trials, definition and operational criteria. *Statistics in Medicine* 1989; **8**:431–440.
4. Hsu C, Taylor JMG, Murray S, Commenges D. Survival analysis using auxiliary variables via nonparametric multiple imputation. *Statistics in Medicine* 2006; **25**:3503–3517.
5. Freedman LS, Graubard BI, Schatzkin A. Statistical validation of intermediate endpoints for chronic disease. *Statistics in Medicine* 1992; **11**:167–178.
6. Lin DY, Fleming TR, DeGruttola V. Estimating the proportion of treatment effect captured by a surrogate marker. *Statistics in Medicine* 1997; **16**:1515–1527.
7. Bycott PW, Taylor JMG. An evaluation of a measure of the proportion of the treatment effect explained by a surrogate marker. *Controlled Clinical Trials* 1998; **19**:555–568.

8. Wang Y, Taylor JMG. A measure of the proportion of treatment effect explained by a surrogate marker. *Biometrics* 2003; **58**:803–812.
9. Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics* 2000; **1**:49–67.
10. De Gruttola V, Fleming T, Lin DY, Coombs R. Perspective: validating surrogate markers—are we being nave? *The Journal of Infectious Diseases* 1997; **175**:237–246.
11. Fleming TR, DeMets DL. Surrogate endpoints in clinical trials: are we being misled? *Annals of Internal Medicine* 1996; **125**:605–613.
12. Baker SG, Kramer BS. A perfect correlate does not a surrogate make. *BMC Medical Research Methodology* 2003; **3**:16.
13. Cook RJ, Lawless JF. Some comments on efficiency gains from auxiliary information for right-censored data. *Journal of Statistical Planning and Inference* 2001; **96**:191–202.
14. Malani HM. A modification of the re-distribution to the right algorithm using disease markers. *Biometrika* 1995; **82**:515–526.
15. Murray S, Tsiatis AA. Nonparametric survival estimation using prognostic longitudinal covariates. *Biometrics* 1996; **52**:137–151.
16. Venkatraman ES, Begg CB. Properties of a nonparametric test for early comparison of treatments in clinical trials in the presence of surrogate endpoints. *Biometrics* 1999; **55**:1171–1176.
17. Gail M, Pfeiffer R, Houwelingen HCV, Carroll RJ. On meta-analytic assessment of surrogate outcomes. *Biostatistics* 2000; **1**:231–246.
18. Henderson CR. Best linear unbiased estimation and prediction under a selection model. *Biometrics* 1975; **31**:423–447.
19. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**:13–22.
20. Laird NM, Lang N, Stram D. Random-effects models for longitudinal data. *Biometrics* 1982; **38**:963–974.
21. Robinson GK. That BLUP is a good thing: the estimation of random effects. *Statistical Science* 1991; **6**:15–51.
22. Ghosh M, Rao NK. Small area estimation: an appraisal. *Statistical Science* 1994; **9**:55–76.
23. Morris C. Parametric empirical Bayes inference: theory and application (with Discussion). *Journal of American Statistical Association* 1983; **78**:47–65.
24. Louis TA, Zelterman D. Bayesian approaches to research synthesis. In *The Handbook of Research Synthesis*, Cooper H, Hedges LV (eds). Russell Sage Foundation: New York, 1994.
25. Musch DC, Lichter PR, Guire KE, Standardi CL, CIGTS Investigators. The Collaborative Initial Glaucoma Treatment Study (CIGTS): Study design, methods, and baseline characteristics of enrolled patients. *Ophthalmology* 1999; **106**:653–662.
26. Little RJA, Rubin DB. *Statistical Analysis with Missing Data* (2nd edn). Wiley: New York, 2002.
27. Burzykowski T, Molenberghs G, Buyse M. *The Evaluation of Surrogate Endpoints*, Chapter 18. Springer: New York, 2004.
28. McLean RA, Sanders WL. Approximating degrees of freedom for standard errors in mixed linear models. *Proceedings of the Statistical Computing Section*, American Statistical Association, New Orleans, 1988; 50–59.
29. McLean RA, Sanders WL, Stroup WW. A unified approach to mixed linear models. *The American Statistician* 1991; **45**:54–64.