Increasing efficiency for estimating treatment-biomarker interactions with historical data

Philip S. Boonstra[1], Jeremy M. G. Taylor[1], and Bhramar Mukherjee[1]

Abstract

Detecting a treatment-biomarker interaction, which is a task better suited for large sample sizes, in a phase II trial, which has a small sample size, is challenging. In this paper, we investigate how two plausibly-available sources of historical data may contain partial information to help estimate the treatment-biomarker interaction parameter in a randomized phase II study. The parameter is not identified in either historical dataset alone; nonetheless, both can provide some information about the parameter and, consequently, increase the precision of its estimate. To illustrate the potential for gains in efficiency and implications for the design of the study, we consider Gaussian outcomes and biomarker data and calculate the asymptotic variance using the expected Fisher information matrix. We quantify the gain in efficiency both through a numerical study and, in a simplified setting, insights derived from an algebraic development of the problem. We find that a non-negligible gain in precision is possible, even if the historical and prospective data do not arise from identical underlying models.

Keywords: auxiliary data, phase II clinical trials, precision, randomized clinical trial, variance

[1]Department of Biostatistics, University of Michigan, Ann Arbor 48109

Corresponding Author: Philip S. Boonstra, 1415 Washington Hts., Ann Arbor, MI, 48109, USA,

Email: philb@umich.edu

1 Introduction

Many new targeted therapies being evaluated for treatment of cancer are not expected to work equally well for all subjects in a population. For example, colorectal tumors exhibiting high-frequency microsatellite instability may be more responsive to treatments targeting the PI3K-AKT-mTOR pathway.[1] Or treatments may be prospectively tailored toward patients with specific molecular attributes, for example EGFR activation.[2] Trials of such targeted therapies frequently include measurements of biomarker(s), where, for example, if the biomarker is high the belief is that the treatment will be more effective. In statistical terms this prior belief can be represented by a treatment-biomarker interaction in a model for the outcome.[3,4,5] As treatment biomarker combinations are evaluated in early phase studies, it is important to gather data to assess whether this prior belief regarding the existence of interactions could be true. Such information is crucial to help inform the design of the next study to further evaluate the treatment.

In this paper we consider the situation of a modestly-sized, prospective randomized phase II trial, which is not intended to be definitive but rather informative about whether to proceed to a more definitive phase III trial. The possible future phase III trial might be restricted to subjects with high values of the biomarker, or it may be an all-comers trial. The size, expense and likelihood of reaching a sound conclusion from the phase III trial may depend on the selected design. Thus learning as much information as possible from the phase II trial about how the treatment effect varies by the level of the biomarker is of great importance. For trial design, and statistical models in general, it is well-known that larger sample sizes are needed to test and estimate interactions than are needed to test and estimate main effects.[6] Because the sample size of the phase II trial is already modest, any means of increasing the precision of the estimate of the interaction would be very useful. Motivated by this, we posit that historical data can increase the precision of estimates of treatment-biomarker interactions. Although such data by itself does not typically provide direct information

about the interaction parameter, it is plausible that, when combined with the data from the prospective phase II trial, it could lead to improved efficiency in the estimation of the treatment-biomarker interaction.

The motivation for this research comes from a phase II study currently ongoing at the University of Michigan. In the study, which has several unique features that we will not consider, castration-resistant prostate cancer patients are evaluated for ETS transcription factors *ERG* and *ETV1* in their biopsied, metastatic lesion. Overexpression of these biomarkers represents an ETS gene fusion in the prostate tumor, which is driven by an androgen-sensitive promoter.[7] Additionally, ETS-mediated oncogenic features such as metastasis and tumor growth depend on PARP1, thus the PARP1-inhibitor ABT-888 can specifically target ETS-positive prostate cancers.[8] The trial design is a stratified randomized phase II with a planned sample size of 148 evaluable patients and two treatment arms: abiraterone with prednisone (A+P) versus abiraterone with prednisone and ABT-888 (A+P+B). The trial will evaluate the role of ETS gene fusion as a predictive biomarker comparing A+P to A+P+B in this patient population. Two hypothetical sources of additional data that would help inform about the interaction are (i) data from subjects who all received A+P and have measurements of the ETS gene fusion or (ii) trial data from subjects to whom either A+P or A+P+B were given but for which ETS gene fusion status is not known. This second study may either be observational or a randomized trial.

In this paper, we consider a general statistical framework for a problem of this type and undertake an evaluation of the potential for such auxiliary data to help in the analysis of the data from the phase II study. In this general framework, Thall and Simon[9] investigate how to use historical control data to improve estimates of the *main effect* of treatment. More recently, Neuenschwander et al.[10] discuss how to incorporate historical estimates of treatment effects and characterize their contribution through the calculation of an effective historical sample size. However, Cuffe[11] shows that incorporating historical control data for

estimating treatment effects can decrease power if the historical and prospective data are too incompatible, which may not be known until after the prospective trial is complete. In the context of augmenting human data with that from animal studies, DuMouchel and Harris[12] suggest to ascertain multiple sources, so as to minimize sensitivity to any single historical dataset. We will focus on the efficiency of estimation for the *interaction* term, which will be evaluated by consideration of the expected Fisher information matrix. For our purposes, to illustrate the potential efficiency gain, we restrict our attention to the case where both the response variable and the biomarker are continuous, and the data can be described by a linear model with Gaussian errors. The research has implications for both the analysis and the design of the planned phase II trial. The analysis question is how much gain in efficiency is possible from the auxiliary data. There are two possible design issues, one being how much auxiliary data should be acquired or gathered and the other whether it is useful to have an unequal randomization between treatment arms in the prospective trial. In the remainder of the paper, we introduce notation and the models for the problem (Section 2), interpret the results of our numeric and analytic study of the problem (Sections 3 and 4), and close with a brief discussion (Section 5). The algebraic details of our work are primarily in the Web Appendix.

2 Notation and Models

Let Y be the outcome measure. A binary variable $T \in \{-1/2, 1/2\}$ indicates the new drug or treatment, with $T = -1/2$ for those who receive the standard treatment. The other covariates in the model are $\boldsymbol{W} = \{W_1, W_2, \ldots, W_q\}^\top \in \mathbb{R}^q$, assumed to be a length-q vector of commonly measured prognostic variables, and $V \in \mathbb{R}$, which is the biomarker. The biomarker is chosen so that patients with higher values of V are expected to be more responsive to the new treatment. In the simplest case of a continuous Y, the outcome model

may be given by:

$$Y|\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{X}^\top \boldsymbol{\beta}, \sigma^2), \tag{1}$$

where the length-$(q+4)$ vector $\boldsymbol{X} = \{1, T, \boldsymbol{W}^\top, V, T \times V\}^\top$, $\boldsymbol{\beta} = \{\beta_0, \beta_T, \boldsymbol{\beta}_{\boldsymbol{W}}^\top, \beta_V, \theta\}^\top$ and $\boldsymbol{\beta}_{\boldsymbol{W}} = \{\beta_{W1}, \beta_{W2}, \ldots, \beta_{Wq}\}^\top$. Our primary goal in this paper is to determine if V is predictive, meaning that a patient's measure of V determines in part the extent the treatment T affects Y. Algebraically, this is true if $\theta \neq 0$. In contrast, our focus is not in determining whether V is prognostic, which means that typical values of the outcome Y depend on V for patients treated under the standard of care, i.e. when $T = -1/2$. The biomarker is prognostic if $\beta_V - \theta/2 \neq 0$. Clearly, V may be both prognostic and predictive, but inference on θ is of primary interest.

Group 1: Prospective randomized clinical trial (RCT). We call the primary, small prospective phase II study group 1 and assume it has n_1 observations. The treatment T may take on both values in this group. Because the data in group 1 come from a RCT, T is assumed to be independent of both V and $\boldsymbol{W}$.

Group 2: Historical data under standard of care. In group 2, $T = -1/2$ for all observations, corresponding to a set of patients treated under the standard of care, and the biomarker V is measured. We will call these data, containing n_2 observations, group 2. By itself, group 2 only provides information about the quantity $\beta_V - \theta/2$, i.e. whether or not the biomarker is prognostic, and θ is not identified because T does not vary in this group. However, as we will show, even by helping to inform the quantity $\beta_V - \theta/2$, group 2 may still provide a gain in efficiency for estimating θ.

Group 3: Historical data that is missing biomarker information. In these data (group 3), T varies, as with group 1, but the biomarker V is unmeasured. These may be patient data from earlier-phase studies of the treatment of interest, either observational or a randomized study, before the biomarker V was identified or in which it was not measured. It is reasonable to

assume that the distribution of T is conditionally independent of (the unobserved) V given $\boldsymbol{W}$. This is trivially satisfied if group 3 arises from an RCT. However, even if group 3 comes from an observational study, so that the distribution of T depends on the prognostic variables $\boldsymbol{W}$, conditional independence may be assumed because V was not measured or not known. The assumption is critical because it differentiates this scenario from a more problematic retrospective validation of purely observational data, in which T and V may be highly correlated.[13] In practice, conditional independence will not hold exactly; however, if $\boldsymbol{W}$ is sufficiently rich, assuming so will be a good approximation. Depending on the underlying model for V or $V|\boldsymbol{W}$, by itself, group 3 will typically only provide information on expressions containing both θ and the remaining elements of $\boldsymbol{\beta}$, and θ alone is not identified. However, as with group 2, group 3 may contribute a gain in efficiency for estimating θ when combined with group 1.

We will quantify this efficiency gain in estimating θ from the inclusion of groups 2 or 3 through a numerical evaluation of Fisher information matrices. The assumed biomarker and treatment models are, respectively,

$$V|\boldsymbol{W} \sim \mathrm{N}(\gamma_0 + \boldsymbol{\gamma}^\top \boldsymbol{W}, \tau^2), \tag{2}$$

$$T|\boldsymbol{W} + 0.5 \sim \mathrm{Bin}(\mathrm{expit}\{\alpha_0 + \boldsymbol{\alpha}^\top \boldsymbol{W}\}), \tag{3}$$

where $\mathrm{expit}\{x\} = [1 + \exp(-x)]^{-1}$. Model (3) does not apply to group 2, because T does not vary in that group. We also assume that $\boldsymbol{\alpha} = \mathbf{0}_q$ in group 1, because it is a randomized trial. If group 3 corresponds to data from an observational study, then $\boldsymbol{\alpha} \neq \mathbf{0}_q$. When V is missing, as in group 3, the marginal outcome model can be determined by mixing over the original outcome model (1) and the biomarker model (2), leading to

$$Y|T, \boldsymbol{W} \sim$$
$$N\left([\beta_0 + \gamma_0\beta_V] + T[\beta_T + \gamma_0\theta] + \boldsymbol{W}^\top[\boldsymbol{\beta_W} + \beta_V\boldsymbol{\gamma}] + T\boldsymbol{W}^\top[\theta\boldsymbol{\gamma}], \tau^2[\beta_V + T\theta]^2 + \sigma^2\right). \tag{4}$$

The variance of $Y|T, \boldsymbol{W}$ depends on the value of the treatment T. Under these models, then, four functions of the parameters contain θ, through which group 3 contributes its relevant information: $\beta_T + \gamma_0\theta$, $\theta\boldsymbol{\gamma}$, $\tau^2[\beta_V - \theta/2]^2 + \sigma^2$ (the error variance when $T = -1/2$), $\tau^2[\beta_V + \theta/2]^2 + \sigma^2$ (the error variance when $T = 1/2$).

REMARK 1: Although models (1) – (3) are written so that all parameters are shared between the three groups, we will weaken this assumption in our numerical study. The intercepts β_0 and γ_0 may vary between groups, which is important because it allows for the populations to differ in ways that may not be explained by $\boldsymbol{W}$.

REMARK 2: It is not necessary to write an explicit model for $\boldsymbol{W}$, because it is fully observed in all groups. For ease of notation, we denote the first two moments of $\boldsymbol{W}$ as $\mathrm{E}[\boldsymbol{W}]$ and $\mathrm{E}[\boldsymbol{W}\boldsymbol{W}^\top]$ but do not require that these be equal between groups.

3 Numerical study of efficiency gains

In this section, we assess, through a numerical study, the benefit to the analysis of group 1, assumed to be RCT data, from adding either the group 2 or group 3 data. To do this, we estimate the asymptotic variance of the interaction parameter θ by numerically calculating and inverting the expected Fisher information matrices of the parameters from models (1) and (2) and recording the resulting entry corresponding to θ. The expected Fisher information matrices for one observation from groups 1, 2, or 3 are respectively denoted as $\boldsymbol{I}_1$, $\boldsymbol{I}_2$ or

$\boldsymbol{I}_3$ and are derived in Appendices A1–A3:

$$\boldsymbol{I}_1 = \frac{1}{\sigma^2}\begin{pmatrix} 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ T & 1/4 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \boldsymbol{W} & T\boldsymbol{W} & \boldsymbol{W}\boldsymbol{W}^\top & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ V & TV & V\boldsymbol{W}^\top & V^2 & \cdot & \cdot & \cdot & \cdot & \cdot \\ TV & V/4 & TV\boldsymbol{W}^\top & TV^2 & V^2/4 & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \mathbf{0}_q^\top & 0 & 0 & 1/(2\sigma^2) & \cdot & \cdot & \cdot \\ 0 & 0 & \mathbf{0}_q^\top & 0 & 0 & 0 & \sigma^2/\tau^2 & \cdot & \cdot \\ \mathbf{0}_q & \mathbf{0}_q & \mathbf{0}_q\mathbf{0}_q^\top & \mathbf{0}_q & \mathbf{0}_q & \mathbf{0}_q & (\sigma^2/\tau^2)\boldsymbol{W} & (\sigma^2/\tau^2)\boldsymbol{W}\boldsymbol{W}^\top & \cdot \\ 0 & 0 & \mathbf{0}_q^\top & 0 & 0 & 0 & 0 & \mathbf{0}_q^\top & \sigma^2/(2\tau^4) \end{pmatrix}, \tag{5}$$

$$\boldsymbol{I}_2 = \frac{1}{\sigma^2}\begin{pmatrix} 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ -1/2 & 1/4 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \boldsymbol{W} & -\boldsymbol{W}/2 & \boldsymbol{W}\boldsymbol{W}^\top & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ V & -V/2 & V\boldsymbol{W}^\top & V^2 & \cdot & \cdot & \cdot & \cdot & \cdot \\ -V/2 & V/4 & -V\boldsymbol{W}^\top/2 & -V^2/2 & V^2/4 & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \mathbf{0}_q^\top & 0 & 0 & 1/(2\sigma^2) & \cdot & \cdot & \cdot \\ 0 & 0 & \mathbf{0}_q^\top & 0 & 0 & 0 & \sigma^2/\tau^2 & \cdot & \cdot \\ \mathbf{0}_q & \mathbf{0}_q & \mathbf{0}_q\mathbf{0}_q^\top & \mathbf{0}_q & \mathbf{0}_q & \mathbf{0}_q & (\sigma^2/\tau^2)\boldsymbol{W} & (\sigma^2/\tau^2)\boldsymbol{W}\boldsymbol{W}^\top & \cdot \\ 0 & 0 & \mathbf{0}_q^\top & 0 & 0 & 0 & 0 & \mathbf{0}_q^\top & \sigma^2/(2\tau^4) \end{pmatrix} \tag{6}$$

$$\boldsymbol{I}_3 = \frac{1}{\delta^2}\begin{pmatrix} 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ T & 1/4 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \boldsymbol{W} & T\boldsymbol{W} & \boldsymbol{W}\boldsymbol{W}^\top & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \tilde{V} & T\tilde{V} & \tilde{V}\boldsymbol{W}^\top & \tilde{V}^2{+}2\tau^4D^2/\delta^2 & \cdot & \cdot & \cdot & \cdot & \cdot \\ T\tilde{V} & \tilde{V}/4 & T\tilde{V}\boldsymbol{W}^\top & T\tilde{V}^2{+}2T\tau^4D^2/\delta^2 & \tilde{V}^2/4{+}\tau^4D^2/(2\delta^2) & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \tau^2D/\delta^2 & T\tau^2D/\delta^2 & 1/(2\delta^2) & \cdot & \cdot & \cdot \\ D & TD & D\boldsymbol{W}^\top & \tilde{V}D & T\tilde{V}D & 0 & D^2 & \cdot & \cdot \\ D\boldsymbol{W} & TD\boldsymbol{W} & D\boldsymbol{W}\boldsymbol{W}^\top & \tilde{V}D\boldsymbol{W} & T\tilde{V}D\boldsymbol{W} & \mathbf{0}_q & D^2\boldsymbol{W} & D^2\boldsymbol{W}\boldsymbol{W}^\top & \cdot \\ 0 & 0 & 0 & \tau^2D^3/\delta^2 & T\tau^2D^3/\delta^2 & D^2/(2\delta^2) & 0 & 0 & D^4/(2\delta^2) \end{pmatrix} \tag{7}$$

where, in $\boldsymbol{I}_3$, $D = \beta_V + T\theta$, $\tilde{V} = \mathrm{E}[V|\boldsymbol{W}] = \gamma_0 + \boldsymbol{W}^\top\boldsymbol{\gamma}$, and $\delta = (\tau^2D^2 + \sigma^2)^{1/2}$. Reading down the diagonal of each matrix, the elements correspond to $\beta_0, \beta_T, \boldsymbol{\beta}_{\boldsymbol{W}}, \beta_V, \theta, \sigma^2, \gamma_0, \boldsymbol{\gamma}$, and τ^2, and so the fifth row/column, corresponding to θ, is of primary interest. Compared to $\boldsymbol{I}_1$ or $\boldsymbol{I}_2$, $\boldsymbol{I}_3$ contains additional non-zero off-diagonal elements, which may depend on the values of β_V or θ, because it is required to marginalize over the missing biomarker measurements V, which are fully observed in groups 1 and 2.

For each group, given a set of parameter values, we independently sample 100,000 draws of $\boldsymbol{W} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{W}})$, $T|\boldsymbol{W}$ (for groups 1 and 3), and $V|\boldsymbol{W}$ (for groups 1 and 2), and calculate the average values of $\boldsymbol{I}_1$, $\boldsymbol{I}_2$, and $\boldsymbol{I}_3$. Given these estimates of typical information matrices, define the following measures of relative efficiency for estimating θ, given some non-negative

number k:

$$\begin{aligned}
\mathrm{RE}_2(k) &= \{(\boldsymbol{I}_1 + k\boldsymbol{I}_2)^{-1}\}_{[5,5]}/\{\boldsymbol{I}_1^{-1}\}_{[5,5]}, \\
\mathrm{RE}_3(k) &= \{(\boldsymbol{I}_1 + k\boldsymbol{I}_3)^{-1}\}_{[5,5]}/\{\boldsymbol{I}_1^{-1}\}_{[5,5]}, \\
\mathrm{RE}_{23}(k) &= \{(\boldsymbol{I}_1 + 0.5k[\boldsymbol{I}_2 + \boldsymbol{I}_3])^{-1}\}_{[5,5]}/\{\boldsymbol{I}_1^{-1}\}_{[5,5]}, \\
\mathrm{RE}_{\max}(k) &= \{(\boldsymbol{I}_1 + k\boldsymbol{I}_1)^{-1}\}_{[5,5]}/\{\boldsymbol{I}_1^{-1}\}_{[5,5]} = 1/(1+k).
\end{aligned}$$

For an arbitrary sample size n_1, $\mathrm{RE}_2(k)$, $\mathrm{RE}_3(k)$, $\mathrm{RE}_{23}(k)$, and $\mathrm{RE}_{\max}(k)$ are, respectively, the variance of θ from analyzing (i) group 1 containing n_1 observations and group 2 containing kn_1 observations, (ii) group 1 containing n_1 observations and group 3 containing kn_1 observations, (iii) group 1 containing n_1 observations, group 2 containing $0.5kn_1$ observations, and group 3 containing $0.5kn_1$ observations, and (iv) group 1 containing $n_1(1+k)$ observations, relative to the variance of θ from analyzing group 1 alone, containing n_1 observations. The quantity k is the sample size ratio of the added data versus the original group 1 data, and the value of n_1 is arbitrary. The final metric, $\mathrm{RE}_{\max}(k)$, may be considered the ideal efficiency gain, in which the additional data are most informative for estimating θ because group 1 increases in size. As mentioned, group 1 is supposed to correspond to RCT data, meaning that $\boldsymbol{\alpha} = \mathbf{0}_q$ throughout. We vary $\boldsymbol{\alpha}$ in group 3 to represent RCT data ($= \mathbf{0}_q$) or observational data ($\neq \mathbf{0}_q$).

We now discuss the parameter values considered, which are summarized in Table 1.

$Y|\boldsymbol{X}$ [Model (1)] Because (1) is a Gaussian-linear model, the information matrices from groups 1 and 2 do not change with any elements of $\boldsymbol{\beta}$, and the error variance σ^2 acts only as a multiplicative scalar. However, the information matrix from group 3, which requires marginalization over V, does depend on β_V, θ, and σ^2 (Appendix A3). We consider three sets of values for β_V and θ, $\{\beta_V, \theta\} \in \big\{\{0,1\}, \{1/\sqrt{5}, 2/\sqrt{5}\}, \{2/\sqrt{5}, 1/\sqrt{5}\}\big\}$. Graphical depictions of these three choices are given in Figure 1. For a fixed fraction of explained

variance $R^2_\sigma \in \{0.1, 0.3, 0.5\}$, we determine σ^2 by solving $R^2_\sigma = \boldsymbol{\beta}^\top \text{Var}[\boldsymbol{X}]\boldsymbol{\beta}/(\boldsymbol{\beta}^\top \text{Var}[\boldsymbol{X}]\boldsymbol{\beta} + \sigma^2)$. Appendix A4 has details on this calculation. The parameters β_T and $\boldsymbol{\beta_W}$ are required only for calculating σ^2; we use $\beta_T = 0$ and $\boldsymbol{\beta_W} = (q^{-1/2})\mathbf{1}_q$ (so that $\boldsymbol{\beta}^\top_{\boldsymbol{W}}\boldsymbol{\beta_W} = 1$). Finally, we allow the intercept β_0 to differ between groups, meaning that each group's information matrix has a distinct row and column for β_0, and no group contributes information regarding another group's intercept. However, because we are focusing only on efficiency, the information contribution remains the same for an arbitrary set of three intercepts and no actual values need to be chosen.

$V|\boldsymbol{W}$ [Model (2)] For the biomarker model, we use $\boldsymbol{\gamma} = (q^{-1/2})\mathbf{1}_q$, so that $\boldsymbol{\gamma}^\top\boldsymbol{\gamma} = 1$. We chose γ_0 so that, for a given $\text{E}[\boldsymbol{W}]$ (see following paragraph) and $\boldsymbol{\gamma}$, we have $\text{E}[V] = \gamma_0 + \boldsymbol{\gamma}^\top \text{E}[\boldsymbol{W}] = 0$. As already noted, we assume that γ_0, like β_0, varies between groups, so that each group's information matrix contains a distinct row and column for γ_0. To determine τ^2, we solved $R^2_\tau = \boldsymbol{\gamma}^\top \text{Var}[\boldsymbol{W}]\boldsymbol{\gamma}/(\boldsymbol{\gamma}^\top \text{Var}[\boldsymbol{W}]\boldsymbol{\gamma} + \tau^2)$ using $R^2_\tau \in \{0.1, 0.3, 0.5\}$.

$T|\boldsymbol{W}$ [Model (3)] and model for $\boldsymbol{W}$ For group 1, we assume that $\boldsymbol{\alpha} = \mathbf{0}_q$ and $\alpha_0 = 0$, corresponding to a RCT with equal randomization probabilities. For group 2, no model for T is required because T is fixed at $-1/2$. For group 3, we considered $\boldsymbol{\alpha} \in \{\mathbf{0}_q, (2q^{-1/2})\mathbf{1}_q\}$ but still chose α_0 so that $E[T] = \text{E}[\text{expit}\{\alpha_0 + \boldsymbol{\alpha}^\top \boldsymbol{W}\}] - 0.5 = 0$. We drew $\boldsymbol{W}$ from a multivariate normal distribution with $\text{E}[\boldsymbol{W}] \equiv \boldsymbol{\mu} = \mathbf{0}_q$ and $\text{Var}[\boldsymbol{W}] \equiv \boldsymbol{\Sigma_W} = \text{diag}\{1, 1, \ldots, 1\}$.

Graphical depictions of how each RE measure changes with k for $\{\beta_V, \theta\} = \{1/\sqrt{5}, 2/\sqrt{5}\}$ are given in Figure 2 ($\boldsymbol{\alpha} = \mathbf{0}_q$ in group 3) and Figure 3 ($\boldsymbol{\alpha} = (2q^{-1/2})\mathbf{1}_q$ in group 3). Analogous figures for the other values of $\{\beta_V, \theta\}$ are in Figures A1–A4 in the Appendix, for which much of the following interpretation of Figures 2 and 3 remains the same. Within each figure, we show results for all values of R^2_σ and R^2_τ. In the top-right corner of each panel is the partial-R^2 of θ, which is a measure in $(0, 1)$ of how much of R^2_σ is due to θ. As the partial-R^2 increases, θ explains a greater proportion of the variance in model (1).

The typical initial drop in $\text{RE}_2(k)$ is steep, suggesting that even a small amount of data

from group 2 is helpful. However, there are diminishing gains from increasingly large sample sizes in group 2, and the curves for group 2 eventually plateau at some positive asymptote. Under this model configuration, the asymptote is $\lim_{k\to\infty} \mathrm{RE}_2(k) = 0.5$. Section 4 discusses this in more detail. The efficiency gains from a small amount of group 3 are less noticeable. Empirically, there is also an asymptote for possible efficiency gains from group 3 (not shown). However, the height of the asymptote changes and may be smaller than 0.5, as in the rightmost column of Figures 2 and 3. When $\boldsymbol{\alpha} = \mathbf{0}_q$ (Figure 2), $\mathrm{RE}_3(k)$ is smaller than when $\boldsymbol{\alpha} \neq \mathbf{0}_q$ (Figure 3), meaning, when the group 3 data are from an observational study, they are generally less informative for estimating θ than when they come from an RCT.

The figures also contain $\mathrm{RE}_{23}(k)$, which is the relative efficiency from adding $0.5kn_1$ observations from each of groups 2 and 3. In all cases, at around $k = 1$, $\mathrm{RE}_{23}(k)$ is smaller than either $\mathrm{RE}_2(k)$ or $\mathrm{RE}_3(k)$, suggesting an apparent synergy from the joint addition of data from both groups 2 and 3. Also, in the cases we considered, the asymptote is empirically always nearly zero, meaning that, with enough data, adding equal numbers of observations from both groups can match the ideal scenario of having an equivalently-sized group 1. In other words, adding some data from both groups would seem most preferred.

We reiterate that these results are for the case when β_0 and γ_0 are assumed to differ between groups. These results are also invariant to the *true* value(s) of β_0, because it does not appear in the information matrix, and numerical studies (not shown) suggest that these are invariant to the true values of $\boldsymbol{\mu}$ and γ_0, even when they vary between groups. Intuitively, then, these parameters are not primarily responsible for the efficiency gain in estimating θ.

4 Combining groups 1 and 2

In this section, we analytically assess the potential contribution from group 2 toward inference on θ, with the goal of optimally randomizing the group 1 patients. Here, we assume it is

known that $\gamma_0 = 0$ and $\mathrm{E}[\boldsymbol{W}] = \mathbf{0}_q$ for both groups. The matrix $(1/n_1)(\boldsymbol{I}_1 + k\boldsymbol{I}_2)^{-1}$, with $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$ as in (5) and (6), is the large-sample covariance matrix of the parameters, where $k = n_2/n_1$ is the sample size ratio as in Section 3. Some algebra (Appendix A5) gives that the large-sample variance of $\hat{\theta}$ is

$$\begin{aligned} v_{\hat{\theta}}(\mu_T, n_1, k) &\equiv (1/n_1)\{(\boldsymbol{I}_1 + k\boldsymbol{I}_2)^{-1}\}_{[5,5]} \\ &= \left(\frac{1}{n_1}\right)\left(\frac{\sigma^2}{\boldsymbol{\gamma}^\top \boldsymbol{\Sigma}_{\boldsymbol{W}} \boldsymbol{\gamma} + \tau^2}\right)\left(\frac{1+k}{(1+2k)/4 + \mu_T(k - \mu_T)}\right). \end{aligned} \tag{8}$$

From this, only the last expression contains k, and so the variance reduction from adding group 2 does not depend on the values of any model parameters except μ_T. We can use (8) to calculate the maximum possible efficiency gain from adding group 2:

$$\lim_{k\to\infty} \frac{v_{\hat{\theta}}(\mu_T, n_1, k)}{v_{\hat{\theta}}(\mu_T, n_1, 0)} = \lim_{k\to\infty} \frac{(1+k)(1/4 - \mu_T^2)}{(1+2k)/4 + \mu_T(k - \mu_T)} = 1/2 - \mu_T.$$

When $\mu_T = 0$, the addition of group 2 can reduce the variance of $\hat{\theta}$ by no more than 1/2, compared to using group 1 alone; this is the asymptote for $\mathrm{RE}_2(k)$ in Figures 2 and 3. A larger variance reduction may be possible as μ_T increases toward 1/2.

For designing a trial that will generate data corresponding to group 1, we can also use (8) to choose how to optimally randomize the n_1 individuals, given n_2, so that we have the largest possible efficiency for estimating θ. That is, choose μ_T to minimize $v_{\hat{\theta}}(\mu_T, n_1, k)$:

$$\frac{\partial \log v_{\hat{\theta}}(\mu_T, n_1, k)}{\partial \mu_T} = \frac{-(k - 2\mu_T)}{(1+2k)/4 + \mu_T(k - \mu_T)} \stackrel{\text{set}}{=} 0 \qquad \Rightarrow \qquad \mu_T^{\text{opt}} = k/2 = n_2/(2n_1). \tag{9}$$

A consequence of this is that, if $k \geq 1$, equivalently $n_2 \geq n_1$, then the asymptotic variance of $\hat{\theta}$ is minimized by assigning all patients in group 1 to the treatment arm, i.e. $\mu_T = 1/2$. In other words, when group 2 is large and precision of $\hat{\theta}$ is of primary importance, the prospective phase II trial should be a single-arm trial. Although the result assumes that the generating model for the outcome is identical between groups, which is why randomization within

group 1 is often recommended,[14] this may be relaxed somewhat. Namely, the expression for $v_{\hat{\theta}}(\mu_T, n_1, n_2)$ in (8) is free of the intercept and treatment effect, respectively β_0 and β_T. Also, $\hat{\beta}_0$ and $\hat{\beta}_T$ are independent of $\hat{\theta}$. Therefore, it is possible to parametrize (1) so that β_0 and/or β_T differ between groups 1 and 2 (subject to constraints on identifiability) while still being able to show that that μ_T^{opt} in (9) remains the same.

An analogous result of finding an optimal μ_T when combining groups 1 and 3 would also be desirable, but the integration over the missing measurements of V adds complexity to the matrix $\boldsymbol{I}_1 + k\boldsymbol{I}_3$, and a closed-form solution to its inverse, while theoretically possible, does not lend itself to the same intuition as above.

5 Discussion

Phase II trials are, by definition, of a smaller scope. However, it is not uncommon for a primary or secondary aim of such a trial to be identifying a particular biomarker-treatment interaction, for which large sample sizes are necessary for efficient estimation. Given this, it is important to use as much available historical data as possible. Through some asymptotic calculations for a specific model, we have demonstrated the potential gain in efficiency in estimating a biomarker-treatment interaction by using auxiliary data. The exact numerical results are dependent on the specifics of the model, such as continuous outcome and biomarker variables and Gaussian distributions. While we have not investigated it, we hypothesize that the findings would be qualitatively similar for other data types and distributions. We will investigate this in our future work.

From a design perspective, in Section 4, we established conditions under which, if sufficiently many observations from group 2 are available, assigning more, or even all, group 1 patients to the treatment may increase the resulting precision of $\hat{\theta}$ relative to equal-armed randomization. A critical assumption here was that the generating models underlying each group

are similar, although they need not be identical: the intercept and main effect for treatment may differ in the outcome model between groups. However, there are other reasons why it is important to have equal randomization between treatment arms, including differences in the models beyond those considered above[15] or study aims that are of a higher priority than identifying biomarker-treatment interactions. Or, to balance the desire for efficiency gains with possible model discordance, a biased randomization design could assign most, but not all, patients to the experimental treatment arm, so that contemporaneous information about the control arm is available.

The most important result from this study is in Section 3 and suggests that more auxiliary data, from either group 2 or group 3, is always better. In every parametrization we considered, including three different types of interaction that could be encountered in practice (Figures 2 and 3 and A1–A4 in the Appendix), there was an initial steep drop from adding group 2 data, with considerable gains in efficiency possible up to $k = 1$, that is, up to a doubling of the sample size of the original group 1 data. The efficiency gains become marginal as more historical data from a single group are used. If observations from both groups were available, then a synergistic effect was evident, and greater efficiency gains were observed. This synergy is a result of each group contributing different information about the outcome and biomarker models. A practical recommendation when studying treatment-biomarker interactions in phase II trials is therefore to critically assess the types of historical data available to see how they might be incorporated into the analysis of the phase II trial.

While we allow the intercepts in the outcome and biomarker models, respectively, equations (1) and (2), and the distribution of $\boldsymbol{W}$ to vary between groups, we do require that the parameters in the outcome model related to the biomarker, namely β_V and θ, be the same between groups. A violation to this assumption of the models between groups being "similar enough" may change the results of our analysis, in particular the initial steep drop in variance, but the impact from groups 2 and 3 could be down-weighted, for example, through a Bayesian

analysis with a commensurate or power prior.[16,17] That is, apart from using the likelihood, we have not prescribed a particular analysis method, and, by incorporating historical data into a prior, the Bayesian approach may be ideal for managing its contribution to the analysis.

Beyond issues related to differences in models, although a biomarker-treatment interaction comprises a fundamental aspect in our statistical formulation, detection of a linear statistical interaction is neither necessary nor sufficient for determining whether a biomarker is informative for patient-specific treatment assignments. With respect to necessity, there are other ways to characterize a predictive biomarker, an example being identifying subgroups defined by regions of the covariate space in which the treatment is more likely to be effective.[18,19,20] In contrast to a single biomarker V, methods for subgroup identification are better-suited for more complex scenarios in which the treatment effect is expected to vary with respect to several biomarkers and/or covariates. By focusing on the single biomarker situation, which is not uncommon, we are able to better quantify the information gain from the historical data. With respect to sufficiency, two biomarkers may have the same interaction coefficient but differ in their clinical utility;[21,22] in the framework of the models considered in this paper, clinical utility increases as θ increases in magnitude relative to β_V. These references, however, do not consider the addition of auxiliary data, and this article is a "proof-of-concept" paper supporting the contention that if auxiliary data are available, they may aid in the design and subsequent analysis of the proposed randomized phase II study.

Acknowledgments

This work was supported by the National Institutes of Health [CA129102]. Code for the numerical study was written in `R`,[23] using the `lattice`[24] and `latticeExtra`[25] packages to create Figures 2–3, and is available at `http://www-personal.umich.edu/~philb`. The authors thank Stephanie Daignault for her assistance in describing the motivating trial.

References

1. Vilar E, Mukherjee B, Kuick R, et al. Gene expression patterns in mismatch repair-deficient colorectal cancers highlight the potential therapeutic role of inhibitors of the phosphatidylinositol 3-kinase-AKT-mammalian target of rapamycin pathway. *Clin Cancer Res*. 2009;15:2829–2839.
2. Ciardiello F, Tortora G. EGFR antagonists in cancer treatment. *N Engl J Med*. 2008;358:1160–1174.
3. Byar DP. Assessing apparent treatment-covariate interactions in randomized clinical trials. *Stat Med*. 1985;4:255–263.
4. Buyse M. Towards validation of statistically reliable biomarkers. *Eur J Cancer Suppl*. 2007;5:89–95.
5. Simon R. The use of genomics in clinical trial design. *Clin Cancer Res*. 2008;14:5984–5993.
6. Follmann DA. Subgroups and Interactions. In: Geller NL, editor. *Advances in clinical trial biostatistics*. vol. 13. Marcel Dekker, Inc.; 2003. p. 121–140.
7. Tomlins SA, Rhodes DR, Perner S, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*. 2005;310:644–648.
8. Brenner JC, Ateeq B, Li Y, et al. Mechanistic rationale for inhibition of poly(ADP-ribose) polymerase in ETS gene fusion-positive prostate cancer. *Cancer Cell*. 2011;19:664–678.
9. Thall PF, Simon R. Incorporating historical control data in planning phase II clinical trials. *Stat Med*. 1990;9:215–228.
10. Neuenschwander B, Capkun-Niggli G, Branson M, et al. Summarizing historical information on controls in clinical trials. *Clin Trials*. 2010;7:5–18.
11. Cuffe RL. The inclusion of historical control data may reduce the power of a confirmatory study. *Stat Med*. 2011;30:1329–1338.
12. DuMouchel WH, Harris JE. Bayes methods for combining the results of cancer studies

in humans and other species. *J Am Statist Assoc*. 1983;78:293–308.

13. Mandrekar SJ, Sargent DJ. Clinical trial designs for predictive biomarker validation: theoretical considerations and practical challenges. *J Clin Oncol*. 2009;27:4027–4034.
14. Ratain MJ, Sargent DJ. Optimising the design of phase II oncology trials: the importance of randomisation. *Eur J Cancer*. 2009;45:275–280.
15. Taylor JMG, Braun TM, Li Z. Comparing an experimental agent to a standard agent: Relative merits of a one-arm or randomized two-arm phase II design. *Clin Trials*. 2006;3:335–348.
16. Ibrahim JG, Chen MH. Power prior distributions for regression models. *Stat Sci*. 2000;15:46–60.
17. Hobbs BP, Carlin BP, Mandrekar SJ, et al. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*. 2011;67:1047–1056.
18. Senn SJ. *Statistical issues in drug development*. 2nd ed. Hoboken, NJ: John Wiley & Sons; 2008.
19. Ruberg SJ, Chen L, Wang Y. The mean does not mean as much anymore: Finding sub-groups for tailored therapeutics. *Clin Trials*. 2010;7:574–583.
20. Foster JC, Taylor JMG, Ruberg SJ. Subgroup identification from randomized clinical trial data. *Stat Med*. 2011;30:2867–2880.
21. Janes H, Pepe MS, Bossuyt PM, et al. Measuring the performance of markers for guiding treatment decisions. *Ann Intern Med*. 2011;154:253–259.
22. Huang Y, Gilbert PB, Janes H. Assessing treatment-selection markers using a potential outcomes framework. *Biometrics*. 2012;68:687–696.
23. R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria; 2013.
24. Sarkar D. *Lattice: Multivariate data visualization with R*. New York: Springer; 2008.
25. Sarkar D, Andrews F. *latticeExtra: Extra graphical utilities based on Lattice*; 2013. R

package version 0.6-26.

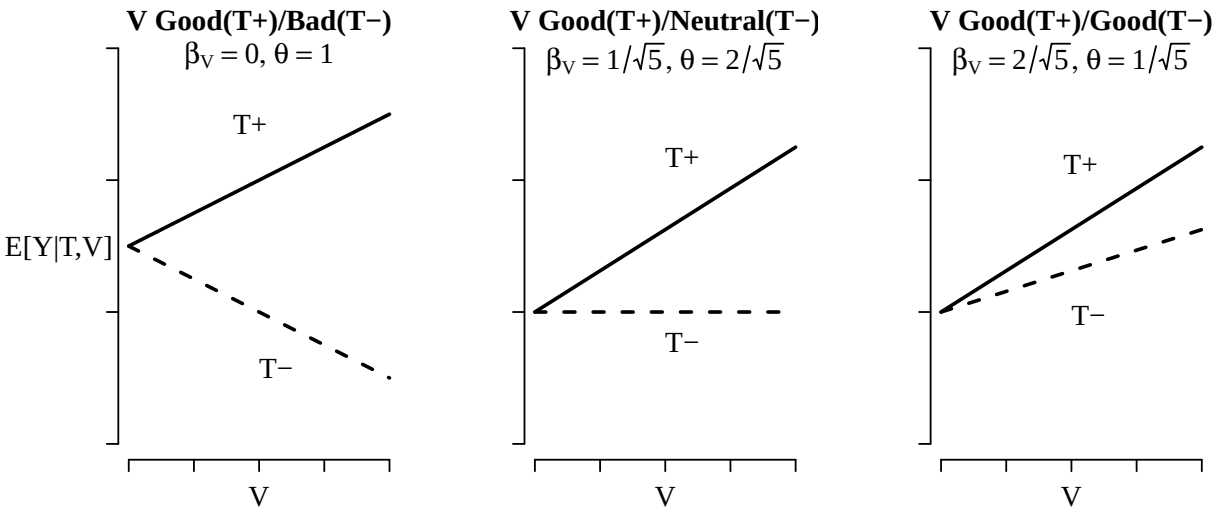

Figure 1: Graphical representation of how average values of Y change with the biomarker V and the two treatment groups under the three sets of values for $\{\beta_V, \theta\}$ considered in the numeric study of Section 3

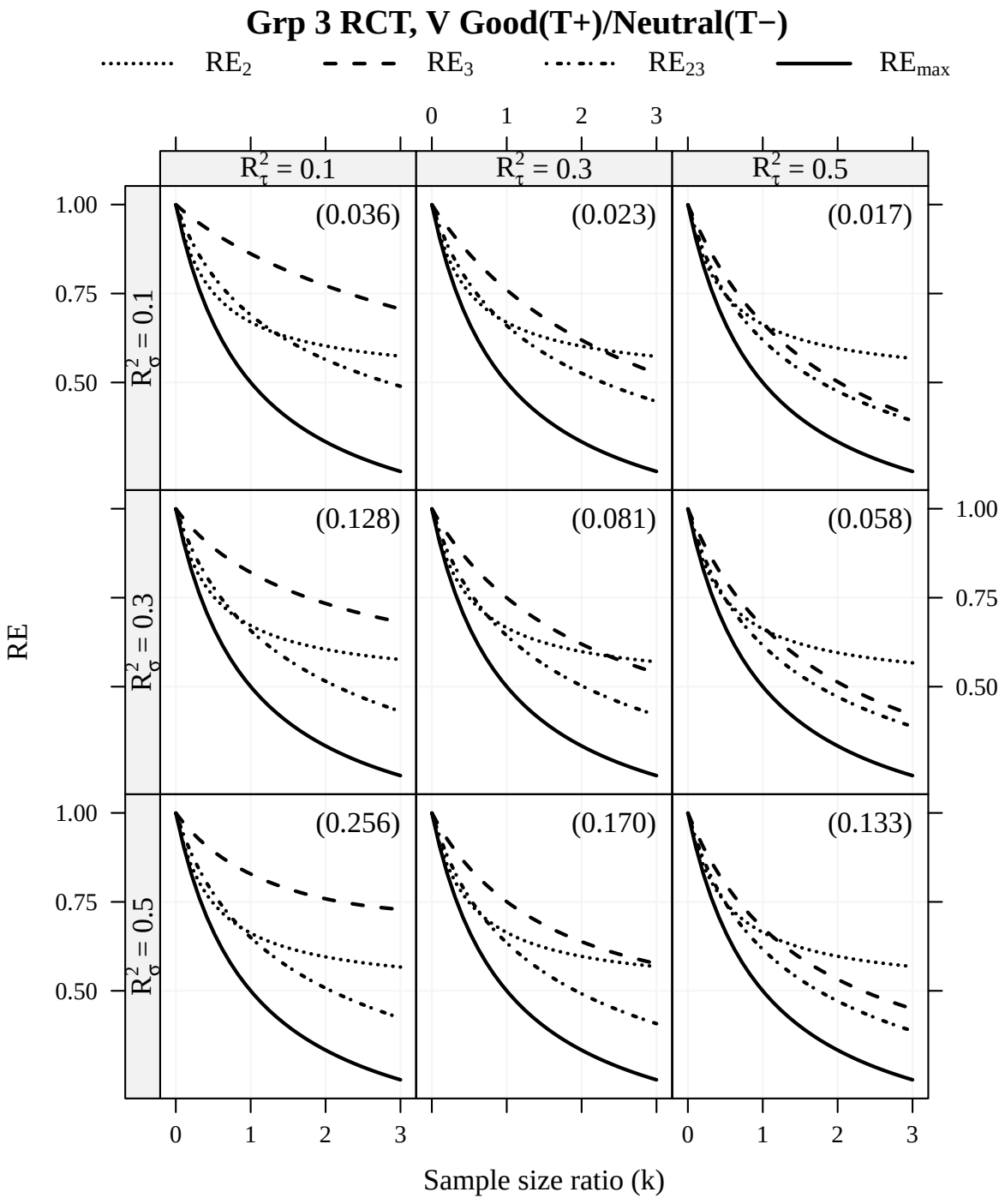

Figure 2: Comparison of efficiency gains when $\boldsymbol{\alpha} = \mathbf{0}_q$ and $\{\beta_V, \theta\} = \{1/\sqrt{5}, 2/\sqrt{5}\}$ for varying R^2_τ (columns) and R^2_σ (rows). The x-axis is the sample size of the added group as a fraction of the original sample size of group 1. In parentheses in each panel is the partial-R^2 value of θ, which is a measure of the contribution of θ to R^2_σ.

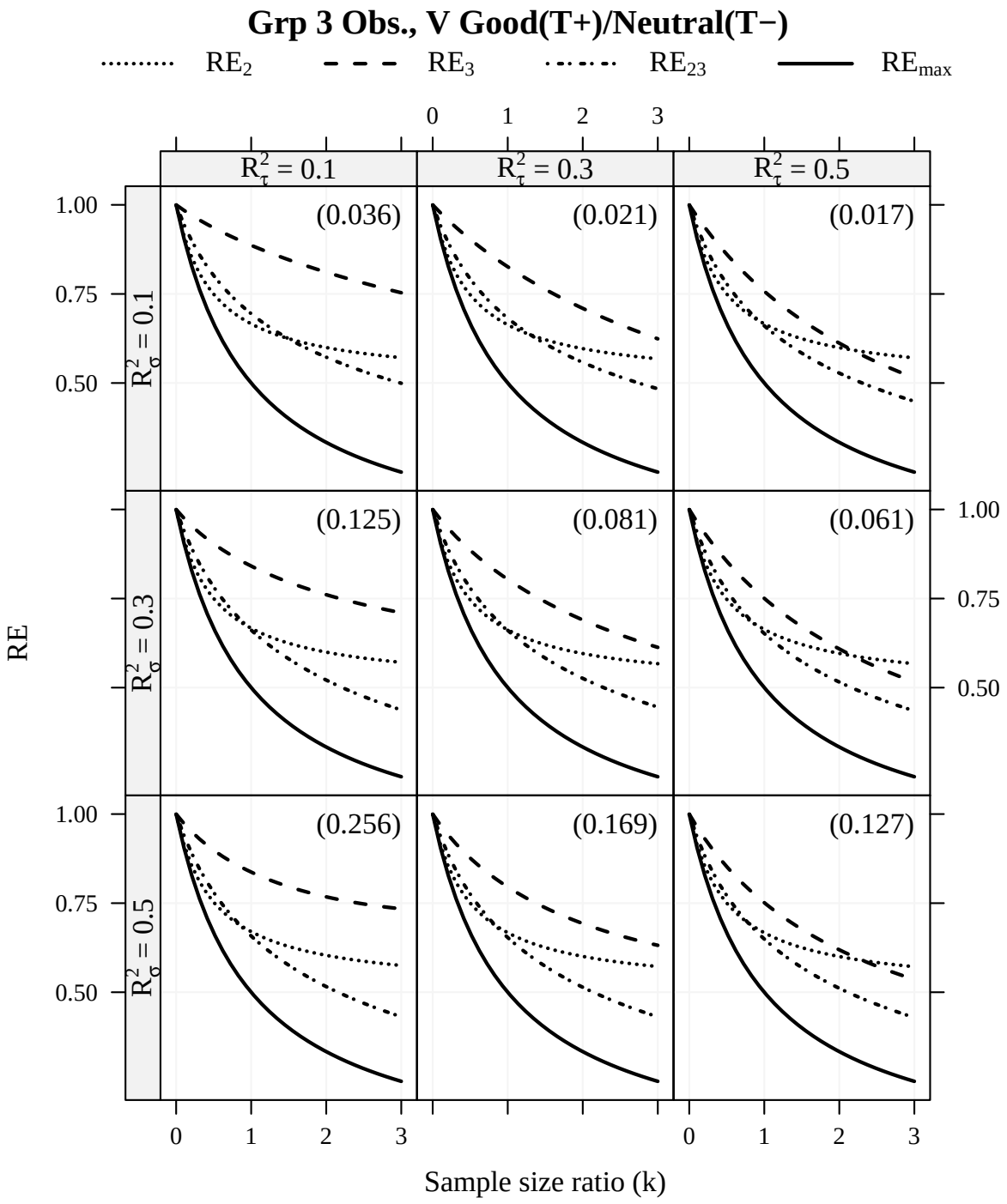

Figure 3: Comparison of efficiency gains when $\boldsymbol{\alpha} = (2q^{-1/2})\mathbf{1}_q$ and $\{\beta_V, \theta\} = \{1/\sqrt{5}, 2/\sqrt{5}\}$ for varying R^2_τ (columns) and R^2_σ (rows).. The x-axis is the sample size of the added group as a fraction of the original sample size of group 1. In parentheses in each panel is the partial-R^2 value of θ, which is a measure of the contribution of θ to R^2_σ.

Model	Parameter	Values Used
$Y\|\boldsymbol{X}$ (1)	β_T	0
	$\boldsymbol{\beta_W}$	$q^{-1/2}\mathbf{1}_q$
	$\{\beta_V, \theta\}$	$\big\{\{0,1\}, \{1/\sqrt{5}, 2/\sqrt{5}\}, \{2/\sqrt{5}, 1/\sqrt{5}\}\big\}$
	R^2_σ	$\{0.1, 0.3, 0.5\}$
$V\|\boldsymbol{W}$ (2)	$\mathrm{E}[V]$	0
	$\boldsymbol{\gamma}$	$q^{-1/2}\mathbf{1}_q$
	R^2_τ	$\{0.1, 0.3, 0.5\}$
$T\|\boldsymbol{W}$ (3)	$\mathrm{E}[T]$	0
	$\boldsymbol{\alpha}$	$\{\mathbf{0}_q, (2q^{-1/2})\mathbf{1}_q\}$
$\boldsymbol{W} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma_W})$	$\boldsymbol{\Sigma_W}$	$\mathrm{diag}\{1, \ldots, 1\}$
	$\boldsymbol{\mu}$	$\mathbf{0}_q$

Table 1: Summary of parameter values in numerical studies. The parameter $\boldsymbol{\alpha}$ varies for group 3 only; for group 1, it is fixed at $\boldsymbol{\alpha} = \mathbf{0}_q$. R^2_σ is used to determine σ by solving $R^2_\sigma = \boldsymbol{\beta}^\top \mathrm{Var}[\boldsymbol{X}]\boldsymbol{\beta}/(\boldsymbol{\beta}^\top \mathrm{Var}[\boldsymbol{X}]\boldsymbol{\beta} + \sigma^2)$. Similarly, τ^2 is found by solving $R^2_\tau = \boldsymbol{\gamma}^\top \boldsymbol{\Sigma_W} \boldsymbol{\gamma}/(\boldsymbol{\gamma}^\top \boldsymbol{\Sigma_W} \boldsymbol{\gamma} + \tau^2)$. The marginal expectations $\mathrm{E}[V]$ and $\mathrm{E}[T]$ respectively induce values for γ_0 and α_0.