ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2017)
66, *Part* 2, *pp.* 345–361

Subgroup selection in adaptive signature designs of confirmatory clinical trials

Zhiwei Zhang,

University of California at Riverside, USA

Meijuan Li, Min Lin and Guoxing Soon,

Food and Drug Administration, Silver Spring, USA

Tom Greene

University of Utah School of Medicine, Salt Lake City, USA

and Changyu Shen

Indiana University School of Medicine, Indianapolis, USA

[Received September 2014. Revised July 2016]

Summary. The increasing awareness of treatment effect heterogeneity has motivated flexible designs of confirmatory clinical trials that prospectively allow investigators to test for treatment efficacy for a subpopulation of patients in addition to the entire population. If a target subpopulation is not well characterized in the design stage, it can be developed at the end of a broad eligibility trial under an adaptive signature design. The paper proposes new procedures for subgroup selection and treatment effect estimation (for the selected subgroup) under an adaptive signature design. We first provide a simple and general characterization of the optimal subgroup that maximizes the power for demonstrating treatment efficacy or the expected gain based on a specified utility function. This characterization motivates a procedure for subgroup selection that involves prediction modelling, augmented inverse probability weighting and low dimensional maximization. A cross-validation procedure can be used to remove or reduce any resubstitution bias that may result from subgroup selection, and a bootstrap procedure can be used to make inference about the treatment effect in the subgroup selected. The approach proposed is evaluated in simulation studies and illustrated with real examples.

Keywords: Cross-validation; Personalized medicine; Predictive biomarker; Subgroup analysis; Treatment effect heterogeneity

1. Introduction

It is well recognized that treatment effects can be heterogeneous, i.e. the same treatment can have different effects on different patients. The increasing awareness of treatment effect heterogeneity has motivated the development of biomarkers that help to identify a target population that is particularly sensitive to a new treatment (Simon, 2008, 2010). There are different ways to incorporate such considerations in a confirmatory clinical trial that is designed to demonstrate the efficacy of a new treatment in some population of patients. If a promising subpopulation

Address for correspondence: Zhiwei Zhang, Department of Statistics, University of California at Riverside, 900 University Avenue, Riverside, CA 92521, USA.
E-mail: zhiwei_zhang@yahoo.com

 0035–9254/17/66345

is available in the design stage, this knowledge can be used to plan a subgroup analysis in a traditional broad eligibility trial (Gail and Simon, 1985; Russek-Cohen and Simon, 1997), or to restrict enrolment in a targeted trial (Simon and Maitournam, 2004; Maitournam and Simon, 2005), depending on how confident the investigator is about the subpopulation. Such a predefined subpopulation is often unavailable in the design stage; instead, the investigator may have only a vague idea that a collection of baseline covariates (e.g. biomarkers, demographics or clinical measurements) could be useful in identifying a sensitive subpopulation. In that case, a target subpopulation can be developed at the end of a broad eligibility trial under an adaptive signature design (ASD) (Freidlin and Simon, 2005; Jiang *et al*., 2007; Freidlin *et al*., 2010), or at an interim analysis (with the possibility of restricting subsequent enrolment) under an adaptive enrichment design (e.g. Follmann (1997), Liu *et al*. (2010), Rosenblum and van der Laan (2011), Simon and Simon (2013) and Wang and Hung (2013)). The various approaches have been discussed and compared by Wang *et al*. (2007), US Food and Drug Administration (2012) and Chen *et al*. (2014), among others.

This paper is concerned with subgroup selection and related estimation and testing problems in ASDs. Freidlin and Simon (2005) described a subgroup selection procedure for genomic biomarkers, and Freidlin *et al*. (2010) considered optimizing the cut point for a single continuous biomarker. Here we consider the problem of finding or approximating the optimal subgroup on the basis of an arbitrary set of baseline covariates. Our objective is to maximize the power for detecting a positive treatment effect on the selected subgroup or, more generally, the expected utility based on a specified utility function (of marker values). The latter formulation allows investigators to take into account the size of the subgroup selected as well as the clinical value (based on prognosis and available treatment options) of demonstrating treatment efficacy for the selected subgroup. Our objective here is notably different from finding an optimal treatment regime that maximizes the expected outcome for the entire population (e.g. Qian and Murphy (2011), Zhao *et al*. (2012) and Zhang *et al*. (2012)). Recognizing the importance of the latter objective, we believe that successful demonstration of treatment efficacy for a clinically important subpopulation is also an important objective, especially in regulatory settings.

Our approach to this problem is based on a simple and general characterization of the optimal subgroup. For a binary outcome, this characterization takes the form of a half-space in terms of covariate-specific response rates (in both treatment groups) and utility (if specified). Motivated by this characterization, we propose a procedure which consists of the following three steps.

Step 1: estimate the covariate-specific response rate in each treatment group.
Step 2: estimate the expected gain for each candidate half-space defined by a vector of coefficients and the estimates from step 1.
Step 3: choose the half-space with the largest estimate of the expected gain.

Step 1 is a standard prediction problem. Step 2 can be cast as a missing data problem, for which we adopt an augmented inverse probability weighting approach. Step 3 is a low dimensional optimization problem which can be solved by using standard techniques. In estimating the treatment effect for the chosen subgroup, a cross-validation approach can be used to remove or reduce any resubstitution bias that may have resulted from the subgroup selection process. A bootstrap procedure can be used to make inference about the treatment effect on the subgroup chosen.

The rest of the paper is organized as follows. In the next section, we formulate the problem and describe the proposed methodology in detail. Section 3 presents simulation results and Section

4 gives illustrative examples. The paper ends with a discussion in Section 5. Additional details are given in Web appendices.

The programs that were used to analyse the data can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. Methodology

2.1. Formulation

An ASD is just an all-comer study with a prospective plan for testing treatment efficacy for a data-driven subgroup of patients. Unlike other adaptive designs, an ASD does not (necessarily) involve an interim analysis; it is adaptive in the selection of patients for a possible subgroup analysis. Usually, the analysis of an ASD begins with a test of overall treatment efficacy on the entire study cohort (at level α_1, say). If the overall test result is significant, then the treatment is considered efficacious for the entire study population. If the test result is non-significant, a subgroup is then selected by using a prespecified procedure, and another test of treatment efficacy is performed on the subgroup selected (at level α_2, say). If the second test result is significant, then the treatment is considered efficacious for the subpopulation selected. The significance levels α_1 and α_2 should be chosen prospectively in a manner that controls the familywise error rate.

In this paper, we restrict attention to ASDs without a test of overall treatment efficacy (the rationale for removing the overall test is given later in remark 1). Thus, our analysis plan differs from that outlined in the previous paragraph in that we perform only one test of treatment efficacy (for a selected subpopulation, which could be the entire study population). Consider a randomized clinical trial with broad entry criteria, where each subject is assigned a treatment T (1, experimental; 0, control) and has an outcome Y. Let $\mathbf{X}$ denote a vector of baseline covariates (e.g. biomarkers, demographics and clinical measurements) that may be used to define a subgroup for demonstrating treatment efficacy. The question is how to choose such a subgroup and to test the associated hypothesis on the basis of $\{(\mathbf{X}_i, T_i, Y_i) : i = 1, \ldots, n\}$, which we conceptualize as independent copies of $(\mathbf{X}, T, Y)$.

Let $\mathcal{X}$ denote the support of $\mathbf{X}$, and let $\mathrm{pow}(A)$ denote the power of a given test of treatment efficacy based on a subgroup defined by $A \subset \mathcal{X}$. (We do not consider measurability issues in this paper). If the investigator is concerned about power only, then an optimal choice of A is a maximizer of $\mathrm{pow}(A)$ over $A \subset \mathcal{X}$. There may, however, be additional considerations concerning the size and content of A. It is obviously more desirable to demonstrate treatment efficacy for the entire population than for a small portion (say 10%) of the population. Moreover, successful demonstration of treatment efficacy is more important for a subgroup of patients with no alternative treatments than for a subgroup with many treatment options. A utility function $u : \mathcal{X} \to [0, \infty)$ can be used to incorporate such considerations in the optimization. With a slight abuse of notation, we write $u(A) = \int_A u(\mathbf{x})\, \mathrm{d}F(\mathbf{x})$, where F is the distribution of $\mathbf{X}$, for the realized gain when a subset A tests significantly. If $u(\mathbf{x}) \equiv 1$, then $u(A) = F(A)$ is just the proportion of patients belonging to A. In general, $u(A)$ is the product of $F(A)$ with the F-average of $u(\mathbf{x})$ over $\mathbf{x} \in A$. For a specified utility function, an optimal choice of A may be defined as a maximizer of the expected gain

$$\gamma(A) = u(A)\ \mathrm{pow}(A).$$

Remark 1. As mentioned earlier, ASDs usually involve two separate tests: a test for the entire population followed by a test for a selected subpopulation (if the first test is non-significant). The main motivation for this strategy appears to be the greater gain in establishing treatment

efficacy for the entire population than for a strict subpopulation. Under our approach, which explicitly takes $u(A)$ into account, there is no need for a separate test for the entire population. This allows us to focus on one test for a selected subpopulation, which could be the entire population.

To fix ideas, we now consider a binary outcome with $Y=1$ being the desired result, and we define $p_t(\mathbf{x})=P(Y=1|T=t,\mathbf{X}=\mathbf{x})$ and $p_t(A)=P(Y=1|T=t,\mathbf{X}\in A)$ for $t=0,1$, $\mathbf{x}\in\mathcal{X}$ and $A\subset\mathcal{X}$. Because T is randomized and therefore independent of $\mathbf{X}$, we have $p_t(A)=\int_A p_t(\mathbf{x})\,\mathrm{d}F(\mathbf{x})/F(A)$ if $F(A)>0$. For a given subset A, $p_t(A)$ can be estimated by the sample proportion

$$\hat{p}_t^{\mathrm{emp}}(A)=\frac{\sum_{i=1}^{n} I(\mathbf{X}_i\in A, T_i=t, Y_i=1)}{\sum_{i=1}^{n} I(\mathbf{X}_i\in A, T_i=t)},$$

where $I(\cdot)$ is the indicator function. A Wald test of $H_0:\delta_p(A)\leqslant\delta_0$ against $H_1:\delta_p(A)>\delta_0$, where $\delta_p(A)=p_1(A)-p_0(A)$ and δ_0 is a prespecified constant, can be constructed by using the difference $\hat{p}_1^{\mathrm{emp}}(A)-\hat{p}_0^{\mathrm{emp}}(A)$, which in large samples is approximately normally distributed with mean $p_1(A)-p_0(A)$ and variance

$$\begin{aligned}\mathrm{var}\{\hat{p}_1^{\mathrm{emp}}(A)-\hat{p}_0^{\mathrm{emp}}(A)\}&=\frac{p_1(A)\{1-p_1(A)\}}{n_1(A)}+\frac{p_0(A)\{1-p_0(A)\}}{n_0(A)}\\&\approx\frac{p_1(A)\{1-p_1(A)\}}{\varpi_1 F(A)n}+\frac{p_0(A)\{1-p_0(A)\}}{\varpi_0 F(A)n},\end{aligned}$$

where $n_t(A)=\Sigma_{i=1}^{n} I(\mathbf{X}_i\in A, T_i=t)$ and $\varpi_t=P(T=t)$, $t=0,1$. The power of a level α Wald test can be approximated by

$$\widetilde{\mathrm{pow}}(A)=\Phi\left(\{p_1(A)-p_0(A)-\delta_0\}\left[\frac{p_1(A)\{1-p_1(A)\}}{\varpi_1 F(A)n}+\frac{p_0(A)\{1-p_0(A)\}}{\varpi_0 F(A)n}\right]^{-1/2}-z_\alpha\right),$$

where Φ is the standard normal distribution function and z_α is the upper α-quantile of the standard normal distribution. This allows us to approximate $\gamma(A)$ by

$$g(A)=u(A)\,\widetilde{\mathrm{pow}}(A),$$

which may be easier to maximize because an analytical expression is available.

Remark 2. Before attempting to maximize $g(A)$, it may be helpful to consider the following subgroups:

$$\begin{aligned}\mathcal{X}_+&=\{\mathbf{x}\in\mathcal{X}: p_1(\mathbf{x})>p_0(\mathbf{x})\},\\\mathcal{X}_-&=\{\mathbf{x}\in\mathcal{X}: p_1(\mathbf{x})<p_0(\mathbf{x})\}.\end{aligned}$$

If a treatment decision (between 1 and 0) must be made on the basis of $\mathbf{X}$ alone, then patients in $\mathcal{X}_+$ and $\mathcal{X}_-$ should receive treatment 1 and 0 respectively. For any $A\subset\mathcal{X}$, we can define the conditional probabilities

$$\begin{aligned}\psi_+(A)&=P(\mathbf{X}\in A|\mathbf{X}\in\mathcal{X}_+),\\\psi_-(A)&=P(\mathbf{X}\notin A|\mathbf{X}\in\mathcal{X}_-),\end{aligned}$$

assuming that the conditioning events have positive probabilities. It may be tempting to call these conditional probabilities sensitivity and specificity, but these terms can be misleading for two reasons. First, our objective is not to estimate $\mathcal{X}_+$ but rather to find a promising subgroup

and to demonstrate treatment efficacy for the selected subgroup. Second, $\mathcal{X}_+$ is not the 'positive treatment effect' subgroup (i.e. the subgroup of patients who will benefit from treatment 1 *versus* treatment 0). The positive treatment effect subgroup can be defined precisely by comparing the potential outcomes for the two treatments (e.g. Zhang *et al.* (2013)). Typically, the two potential outcomes cannot both be observed, which makes it impossible to determine, without strong assumptions, whether a given patient belongs to the positive treatment effect subgroup. For each $\mathbf{x} \in \mathcal{X}_+$, patients with $\mathbf{X} = \mathbf{x}$ fare better on average under treatment 1 than under treatment 0. Within the stratum $\mathbf{X} = \mathbf{x}$, there may be patients who individually fare better under treatment 0 than under treatment 1. In general, when a new treatment is found efficacious for a (sub)population of patients, it is quite possible that some patients in that (sub)population do not benefit from the new treatment.

2.2. Subgroup selection

If $\mathbf{X}$ is discrete, then the subsets of $\mathcal{X}$ can be enumerated, at least in principle. We assume that $\mathbf{X}$ has at least one continuous component, and that the functions (u, p_0, p_1) are continuous in the continuous components of $\mathbf{X}$. Under appropriate conditions, we show in the Web appendix A that a subset A_{opt} that maximizes $g(A)$ consists of $\mathbf{x} \in \mathcal{X}$ such that

$$(1, u(\mathbf{x}), p_0(\mathbf{x}), p_1(\mathbf{x}))\, \mathbf{c}(A_{\text{opt}}) \leqslant 0, \tag{1}$$

where $\mathbf{c}(A_{\text{opt}})$ is a 4-vector that depends on A_{opt} but not on $\mathbf{x}$. If the utility function is constant or if there is no utility function (so the objective function is simply $\widetilde{\text{pow}}(A)$), the same characterization of A_{opt} applies after removing $u(\mathbf{x})$ from expression (1) and changing the definition of $\mathbf{c}(A_{\text{opt}})$ in an obvious way. An analogous result for a continuous outcome is given in the Web appendix B.

Although expression (1) does not describe A_{opt} in an explicit form (because $\mathbf{c}(A_{\text{opt}})$ depends on A_{opt}), it does suggest a practical approach for estimating or approximating A_{opt}. First, obtain an estimate $\hat{p}_t(\mathbf{x})$ of $p_t(\mathbf{x})$ based on a model for $P(Y = 1 | \mathbf{X}, T)$. Next, consider the class of subsets $\mathcal{A} = \{A(\mathbf{c}) : \|\mathbf{c}\| = 1\}$, where

$$A(\mathbf{c}) = \{\mathbf{x} \in \mathcal{X} : (1, u(\mathbf{x}), \hat{p}_0(\mathbf{x}), \hat{p}_1(\mathbf{x}))\mathbf{c} \leqslant 0\} \tag{2}$$

and the unit norm constraint on $\mathbf{c}$ is for uniqueness. If $u(\mathbf{x})$ is constant or unspecified, it will be removed from expression (2). Our proposal is to estimate A_{opt} by $\hat{A}_{\text{opt}} = A(\hat{\mathbf{c}}_{\text{opt}})$, where $\hat{\mathbf{c}}_{\text{opt}}$ maximizes an estimate of $g\{A(\mathbf{c})\}$ over the unit sphere. If $\hat{p}_t(\mathbf{x})$ estimates $p_t(\mathbf{x})$ consistently, then we can expect $\hat{\mathbf{c}}_{\text{opt}}$ and $\hat{A}_{\text{opt}}$ to approach $\mathbf{c}(A_{\text{opt}})$ and A_{opt} respectively in large samples. If $\hat{p}_t(\mathbf{x})$ is inconsistent for $p_t(\mathbf{x})$ (e.g. because of model misspecification), then $\hat{A}_{\text{opt}}$ estimates a local optimum (within the class $\mathcal{A}$).

Remark 3. This approach allows there to be collinearity in $\mathbf{X}$, because the sole purpose of regressing Y on $(\mathbf{X}, T)$ is to obtain $\hat{p}_t(\mathbf{x})$.

Remark 4. It is possible to detect a severe departure of $\hat{A}_{\text{opt}}$ from the global optimum A_{opt}. If $\hat{A}_{\text{opt}}$ approaches A_{opt}, the associated $\hat{\mathbf{c}}_{\text{opt}}$ should be close to $\mathbf{c}(A_{\text{opt}}) \approx \mathbf{c}(\hat{A}_{\text{opt}})$, with $\mathbf{c}(\cdot)$ defined in the Web appendix A and normalized for comparability with $\hat{\mathbf{c}}_{\text{opt}}$. A large difference between $\hat{\mathbf{c}}_{\text{opt}}$ and $\mathbf{c}(\hat{A}_{\text{opt}})$ would then suggest that the class $\mathcal{A}$ does not contain A_{opt}. We could even perform a formal test based on a specified norm for $\hat{\mathbf{c}}_{\text{opt}} - \mathbf{c}(\hat{A}_{\text{opt}})$ by using the following parametric bootstrap procedure: fix $\{(\mathbf{X}_i, T_i) : i = 1, \ldots, n\}$, generate $\{Y_i : i = 1, \ldots, n\}$ from the estimated model $\hat{p}_t(\mathbf{x})$, compute $D_b = \|\hat{\mathbf{c}}^{(b)}_{\text{opt}} - \mathbf{c}(\hat{A}^{(b)}_{\text{opt}})\|$ for each bootstrap sample and take as the p-value the proportion of the D_bs exceeding the observed value $D_{\text{obs}} = \|\hat{\mathbf{c}}_{\text{opt}} - \mathbf{c}(\hat{A}_{\text{opt}})\|$.

For a given $\mathbf{c}$, an estimate of $g\{A(\mathbf{c})\}$ can be obtained by substituting estimates of $(F(A), u(A), p_0(A), p_1(A))$, where $A = A(\mathbf{c})$ is considered fixed in this paragraph. It is natural to estimate $(F(A), u(A))$ by

$$\hat{F}(A) = \frac{1}{n}\sum_{i=1}^{n} I(\mathbf{X}_i \in A),$$
$$\hat{u}(A) = \frac{1}{n}\sum_{i=1}^{n} I(\mathbf{X}_i \in A)\, u(\mathbf{X}_i).$$

Estimation of $p_t(A)$ $(t = 0, 1)$ can be cast as a missing data problem by considering the subjects in A with $T \neq t$ as having missing data on Y (Rubin, 1974). This perspective allows us to draw on existing techniques and insights in the missing data literature. It is straightforward to estimate $p_t(A)$ with the sample proportion $\hat{p}_t^{\text{emp}}(A)$ that was defined in Section 2.1; however, this estimator does not incorporate the available covariate information and therefore may be inefficient. Alternatively, we could use the following regression estimator:

$$\hat{p}_t^{\text{reg}}(A) = \frac{\sum_{i=1}^{n} I(\mathbf{X}_i \in A)\, \hat{p}_t(\mathbf{X}_i)}{\sum_{i=1}^{n} I(\mathbf{X}_i \in A)},$$

where $\hat{p}_t(\mathbf{x})$ is a generic estimate of $p_t(\mathbf{x})$ that was mentioned earlier. This regression estimator does incorporate covariate information, but its consistency depends on the consistency of $\hat{p}_t(\mathbf{x})$. To improve the efficiency of $\hat{p}_t^{\text{emp}}(A)$ without compromising its consistency, we propose to estimate $p_t(A)$ with the following augmented estimator:

$$\hat{p}_t^{\text{aug}}(A) = \hat{p}_t^{\text{emp}}(A) - \frac{\sum_{i=1}^{n} I(\mathbf{X}_i \in A)\{I(T_i = t) - \hat{\varpi}_t(A)\}\, \hat{p}_t(\mathbf{X}_i)}{\sum_{i=1}^{n} I(\mathbf{X}_i \in A, T_i = t)},$$

where $\hat{\varpi}_t(A) = n_t(A)/\{n_0(A) + n_1(A)\}$ is the proportion of subjects in A that receive treatment t. To the extent that $\hat{p}_t^{\text{emp}}(A)$ can be considered an inverse probability weighting estimator, $\hat{p}_t^{\text{aug}}(A)$ is the corresponding augmented inverse probability weighting estimator (Robins *et al.*, 1994). Because of randomization, $\hat{p}_t^{\text{aug}}(A)$ is guaranteed to be consistent for $p_t(A)$, and it attains the semiparametric information bound if $p_t(\mathbf{x})$ is consistent for $p_t(\mathbf{x})$ (Tsiatis, 2006).

Once an estimate of $g\{A(\mathbf{c})\}$, say $\hat{g}\{A(\mathbf{c})\}$, is available for every $\mathbf{c}$, standard optimization techniques can be used to maximize $\hat{g}\{A(\mathbf{c})\}$ over $\mathbf{c}$ or a suitable transformation of $\mathbf{c}$. For example, it may be advantageous to fix one component of $\mathbf{c}$ at 1 (and to remove the unit norm constraint) if a genetic algorithm is used (Zhang *et al.*, 2012). If $u(\mathbf{x})$ is constant or unspecified, $\mathbf{c}$ is three dimensional with 2 degrees of freedom, so it is possible to perform a grid search for $(\eta \in [-\pi/2, \pi/2], \theta \in [0, 2\pi))$ under the following representation of $\mathbf{c} = (c_1, c_2, c_3)$:

$$c_1 = \cos(\eta)\cos(\theta),$$
$$c_2 = \cos(\eta)\sin(\theta),$$
$$c_3 = \sin(\eta).$$

2.3. Treatment effect estimation

Once a subgroup $\hat{A}_{\text{opt}}$ has been chosen, the next question is how to estimate the associated treatment difference $p_1(\hat{A}_{\text{opt}}) - p_0(\hat{A}_{\text{opt}})$. For a fixed A, estimators of $p_t(A)$ have been described in Section 2.2. In the Web appendix C, we give conditions under which an estimator $\hat{p}_t(\hat{A}_{\text{opt}})$ of $p_t(\hat{A}_{\text{opt}})$ is consistent and asymptotically normal. Under such conditions, valid asymptotic inference on $p_t(\hat{A}_{\text{opt}})$ can be made despite the fact that $\hat{A}_{\text{opt}}$ and $\hat{p}_t(\cdot)$ are obtained from the same set of data.

In finite samples, however, a resubstitution bias can arise when the same sample is used to develop $\hat{A}_{\text{opt}}$ and to estimate the treatment effect in this subgroup. In general, the maximum of an estimated objective function is an overestimator of the true objective function evaluated at the maximizer (Dawid, 1994; Senn, 2008; Simon, 2008). In the present setting, $\hat{A}_{\text{opt}}$ is developed for the objective function $g(A)$, which is different from but closely related to the treatment difference $p_1(A) - p_0(A)$. Because a large treatment difference tends to increase the power and hence $g(A)$, some resubstitution bias is likely to exist in the present setting. If this is a serious problem, a cross-validation approach can be used to remove or reduce the resubstitution bias. Specifically, we propose to partition the study cohort randomly into a specified number, say K, of subsamples that are roughly equal in size. For each $k \in \{1, \ldots, K\}$, we use the kth subsample as the validation sample and combine the other subsamples into a training sample. From the training sample we obtain $\hat{A}_{\text{opt}}^{(-k)} = \arg\max_{A \in \mathcal{A}^{(-k)}} \hat{g}^{(-k)}(A)$ by using the exact same method for obtaining $\hat{A}_{\text{opt}}$. The superscript $(-k)$ in $\mathcal{A}^{(-k)}$ and $\hat{g}^{(-k)}$ emphasizes that both objects are based on the training sample alone. Next, we apply $\hat{A}_{\text{opt}}^{(-k)}$ to the validation sample and obtain $\hat{p}_t^{(k)}(\hat{A}_{\text{opt}}^{(-k)})$, where $\hat{p}_t^{(k)}(\cdot)$ is based on the validation sample alone. Depending on the size of the kth subsample, $\hat{p}_t^{(k)}(\cdot)$ could be the empirical estimator (i.e. the sample proportion) or the augmented estimator; in the latter case, the required estimate of $p_t(\mathbf{x})$ must also be based on the kth subsample. Because $\hat{A}_{\text{opt}}^{(-k)}$ and $\hat{p}_t^{(k)}(\cdot)$ are based on independent data, $\hat{p}_t^{(k)}(\hat{A}_{\text{opt}}^{(-k)})$ is free of resubstitution bias. The final cross-validated estimator of $p_t(\hat{A}_{\text{opt}})$ is given by

$$\hat{p}_t^{\text{cv}}(\hat{A}_{\text{opt}}) = \frac{1}{K} \sum_{k=1}^{K} \hat{p}_t^{(k)}(\hat{A}_{\text{opt}}^{(-k)}).$$

The choice of K represents a trade-off between computational demand and potential bias because each training sample is smaller than the whole sample by a factor of $1/K$—a larger K decreases that bias at the expense of more computation. Once K has been chosen, the size of a subsample is determined as roughly n/K. If this number is sufficiently large for reliable estimation of $p_t(\mathbf{x})$, the augmented inverse probability weighting method can be applied to each validation sample to estimate $p_t(\hat{A}_{\text{opt}}^{(-k)})$; otherwise we can use the empirical estimator.

Remark 5. If an empirical estimator is used to estimate $p_t(\hat{A}_{\text{opt}}^{(-k)})$ for each k, the definition of $\hat{p}_t^{\text{cv}}(\hat{A}_{\text{opt}})$ can be refined as

$$\frac{\sum_{i=1}^{n} S_i^{\text{cv}} I(T_i = t) Y_i}{\sum_{i=1}^{n} S_i^{\text{cv}} I(T_i = t)},$$

where S_i^{cv} is the indicator for $\mathbf{X}_i \in \hat{A}_{\text{opt}}^{(-k)}$ if the ith subject belongs to the kth subsample. This refined definition may be preferable as it accounts for possible differences between subsamples in the number of subjects receiving treatment t. Furthermore, the indicator S_i^{cv} can be used to define

$$\hat{F}^{\mathrm{cv}}(\hat{A}_{\mathrm{opt}}) = \frac{1}{n}\sum_{i=1}^{n} S_i^{\mathrm{cv}},$$

$$\hat{u}^{\mathrm{cv}}(\hat{A}_{\mathrm{opt}}) = \frac{1}{n}\sum_{i=1}^{n} S_i^{\mathrm{cv}} u(\mathbf{X}_i).$$

Together with $\hat{p}_t^{\mathrm{cv}}(\hat{A}_{\mathrm{opt}})$ $(t=0,1)$, these estimators can be used to obtain cross-validated estimators of $\widetilde{\mathrm{pow}}(\hat{A}_{\mathrm{opt}})$ and $g(\hat{A}_{\mathrm{opt}})$, which can be helpful in assessing the resubstitution bias and the effect of cross-validation (see Section 3).

Because the terms $\hat{p}_t^{(k)}(\hat{A}_{\mathrm{opt}}^{(-k)})$ $(t=0,1;\, k=1,\ldots,K)$ are not independent of each other, it is not straightforward to derive the variance of the cross-validated treatment difference estimator, $\hat{p}_1^{\mathrm{cv}}(\hat{A}_{\mathrm{opt}}) - \hat{p}_0^{\mathrm{cv}}(\hat{A}_{\mathrm{opt}})$. Nonetheless, inference on $p_1(\hat{A}_{\mathrm{opt}}) - p_0(\hat{A}_{\mathrm{opt}})$ can be based on non-parametric bootstrap standard errors and confidence intervals. A non-parametric bootstrap sample is created by sampling n subjects with replacement from $\{(\mathbf{X}_i, T_i, Y_i) : i = 1,\ldots,n\}$. A cross-validated estimate of $p_1(\hat{A}_{\mathrm{opt}}) - p_0(\hat{A}_{\mathrm{opt}})$ can be obtained from each bootstrap sample, repeating all steps in subgroup selection and cross-validation. This procedure can be repeated many times to produce a collection of bootstrap estimates, from which a bootstrap standard error BSE can be obtained as a sample standard deviation. A level α confidence interval can then be obtained as $\hat{p}_1^{\mathrm{cv}}(\hat{A}_{\mathrm{opt}}) - \hat{p}_0^{\mathrm{cv}}(\hat{A}_{\mathrm{opt}}) \pm z_{\alpha/2}$ BSE, and a level α test of H_0 against H_1 has p-value $1 - \Phi[\{\hat{p}_1^{\mathrm{cv}}(\hat{A}_{\mathrm{opt}}) - \hat{p}_0^{\mathrm{cv}}(\hat{A}_{\mathrm{opt}}) - \delta_0\}/\mathrm{BSE}]$.

Remark 6. Hypothesis testing in ASDs is sometimes conducted by using a permutation test based on permutations of the treatment labels (Jiang *et al.*, 2007; Freidlin *et al.*, 2010). For the permutation distribution to be a valid null distribution, T must be independent of $(\mathbf{X}, Y)$ under the null hypothesis. This independence implies that $p_0(\mathbf{x}) = p_1(\mathbf{x})$ for all $\mathbf{x}$, which further implies that $p_0(A) = p_1(A)$ for all $A \subset \mathcal{X}$. Thus, the permutation test is really a test of the sharp null hypothesis of no treatment effect in any subpopulation. A significant result from the permutation test indicates that a treatment effect exists in some subpopulation, without specifying which subpopulation it is. It is arguably more desirable to test the specific null hypothesis of no treatment effect in the selected subpopulation. The bootstrap procedure that was described in the preceding paragraph provides a specific test as well as a confidence interval for the treatment effect in the subpopulation selected.

3. Simulation

Simulation experiments are performed to evaluate the proposed methodology for subgroup selection and treatment effect estimation. Each experiment is based on 1000 simulated trials, each of which consists of 1000 subjects. In the first two experiments, $\mathbf{X}$ is a 20-vector whose elements are independently distributed as standard normal, and T is a Bernoulli variable with $P(T=1) = 0.5$, independently of $\mathbf{X}$. Given $(\mathbf{X}, T)$, a binary outcome Y is generated from a logistic regression model:

$$P(Y=1 | \mathbf{X}=\mathbf{x}, T=t) = \mathrm{expit}\{(1, \mathbf{x}')\boldsymbol{\beta}_t\}, \tag{3}$$

where $\mathrm{expit}(z) = \{1+\exp(-z)\}^{-1}$ and the true value of $(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1)$ varies across experiments. The analytic objective is to use the trial data to select a promising subpopulation (in terms of $\mathbf{X}$) and to demonstrate a beneficial treatment effect (with $\delta_0 = 0$) on the selected subpopulation at level $\alpha = 0.025$ (one sided). Throughout the simulation study, we work with a constant utility function, $u(\mathbf{x}) \equiv 1$, so that $u(A) = F(A)$ for all $A \subset \mathcal{X}$. This implies that $u(\mathbf{x})$ should be removed from expressions (1) and (2).

Each simulated data set is analysed by using a working model, which may be model (3) or a misspecified model (to be described later), to obtain $\hat{p}_t(\mathbf{x})$, which defines the working class $\mathcal{A}$ through equation (2). For each $A \in \mathcal{A}$, the objective function $g(A)$ is estimated by substituting the estimates $\hat{F}(A)$ and $\hat{p}_t^{\text{aug}}(A)$ (which are defined in Section 2.2) into the definition of $g(A)$. The resulting estimate $\hat{g}(A)$ is then maximized by using the grid search method (with 10^4 grid points) described at the end of Section 2.2. This procedure produces a subset $\hat{A}_{\text{opt}}$ together with the naive (i.e. without cross-validation) estimates $\hat{F}(\hat{A}_{\text{opt}})$, $\widetilde{\text{pow}}(\hat{A}_{\text{opt}})$, $\hat{g}(\hat{A}_{\text{opt}})$ and $\hat{p}_t(\hat{A}_{\text{opt}})$ $(t = 0, 1)$. Next, we perform a tenfold cross-validation as indicated in Section 2.3 and follow remark 5 to obtain the cross-validated estimates $\hat{F}^{\text{cv}}(\hat{A}_{\text{opt}})$, $\widetilde{\text{pow}}^{\text{cv}}(\hat{A}_{\text{opt}})$, $\hat{g}^{\text{cv}}(\hat{A}_{\text{opt}})$ and $\hat{p}_t^{\text{cv}}(\hat{A}_{\text{opt}})$ $(t = 0, 1)$. The two sets of estimates are compared with each other and also with a third set of estimates, which we call the reference estimates. The reference estimates are obtained from an external validation sample, which is an external sample of size 10^5 from the same distribution, on which the same calculations for the naive estimates are repeated, treating $\hat{A}_{\text{opt}}$ as a predefined subset. The reference estimates, which are denoted by $\hat{F}^{\text{ref}}(\hat{A}_{\text{opt}})$, $\widetilde{\text{pow}}^{\text{ref}}(\hat{A}_{\text{opt}})$, $\hat{g}^{\text{ref}}(\hat{A}_{\text{opt}})$ and $\hat{p}_t^{\text{ref}}(\hat{A}_{\text{opt}})$ $(t = 0, 1)$, are not exactly true values; however, they are unbiased (exactly for $F(\hat{A}_{\text{opt}})$ and $p_t(\hat{A}_{\text{opt}})$, and approximately for the other quantities) and much less variable than the other two sets of estimates.

In the first experiment, we set $\beta_0 = \beta_1 = \beta$, which implies the sharp null hypothesis that no treatment effect exists in any subpopulation that is defined by $\mathbf{X}$. We consider two scenarios:

(a) $\beta = \mathbf{0}$;
(b) $\beta = (0, 0.5, 0.5, \ldots, 0.5)'$.

In both scenarios, the optimal subset is easily seen to be $A_{\text{opt}} = \mathcal{X} = \mathbb{R}^{20}$, whose attributes are shown in Table 1 under 'true values'. The optimal subset is estimated by using working models that take the form of equation (3) but possibly exclude some elements of $\mathbf{X}$. Specifically, we compare models based on the first 1, 2, 3, 5, 10 and 20 elements of $\mathbf{X}$. All these working models are correct in scenario (a) because of zero coefficients. In scenario (b), the working models based on a strict subvector of $\mathbf{X}$ are misspecified. Table 1 summarizes the simulation results in terms of the empirical means and standard deviations of the reference, naive and cross-validated estimates of $F(A)$, $\widetilde{\text{pow}}(A)$, $g(A)$, $p_0(A)$, $p_1(A)$ and $\delta_p(A) = p_1(A) - p_0(A)$, where A is $\hat{A}_{\text{opt}}$ based on a specified number of covariates. In Table 1, the reference estimates confirm the sharp null hypothesis, the naive estimates demonstrate the resubstitution bias (for all quantities except $F(A)$) and the cross-validated estimates show that most of the resubstitution bias can be removed through cross-validation. In the case of $\delta_p(A)$, the cross-validated estimates are slightly biased in the negative direction, suggesting that cross-validation could result in overcorrection. Nonetheless, the results in Table 1 indicate that the procedure proposed will be valid (though a little conservative) in terms of type I error control as long as the bootstrap standard error works well. The bootstrap standard error is too time consuming to include in this simulation study, but it has been shown to work well in many situations with $n \leqslant 1000$.

In the second experiment, we simulate data under the alternative hypothesis with

$$\beta_0 = (0, 0.5, 0.5, 0.5, 0, 1, 0.5, 0.5, 0.5, 0.5, 0.5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)',$$
$$\beta_1 = (0, -0.5, -0.5, -0.5, 1, 0, 0.5, 0.5, 0.5, 0.5, 0.5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)'.$$

These values indicate that the first three elements of $\mathbf{X}$ are effect modifiers, the next two are both prognostic and effect modifying, the next five are prognostic only and the last 10 are not related to the outcome in any way. Since both β_0 and β_1 have 0 as the first element, we have $P(Y = 1 | T = t) = 0.5$ $(t = 0, 1)$, so the overall treatment difference is 0. However, a substantial treatment difference may exist in a subpopulation because of effect modifiers. In this experiment,

Table 1. Results of the first simulation experiment (under the sharp null hypothesis): empirical means and standard deviations of reference, naive and cross-validated estimates of $F(A)$, $\widetilde{\text{pow}}(A)$, $g(A)$, $p_0(A)$, $p_1(A)$ and $\delta_p(A) = p_1(A) - p_0(A)$, where A is $\hat{A}_{\text{opt}}$ (based on a specified number of covariates), as well as the true values of these quantities for $A = A_{\text{opt}}$†

Subset A (number of covariates)	*Empirical means*						*Empirical standard deviations*					
	$F(A)$	$\widetilde{\text{pow}}(A)$	$g(A)$	$p_0(A)$	$p_1(A)$	$\delta_p(A)$	$F(A)$	$\widetilde{\text{pow}}(A)$	$g(A)$	$p_0(A)$	$p_1(A)$	$\delta_p(A)$
Scenario (a): $\beta_0 = \beta_1 = \mathbf{0}$												
True values												
A_{opt}	1.00	0.025	0.025	0.50	0.50	0.000						
Reference estimates												
$\hat{A}_{\text{opt}}(1)$	0.59	0.026	0.015	0.50	0.50	0.000	0.28	0.01	0.01	0.00	0.00	0.01
$\hat{A}_{\text{opt}}(2)$	0.57	0.026	0.015	0.50	0.50	0.000	0.25	0.01	0.01	0.00	0.00	0.01
$\hat{A}_{\text{opt}}(3)$	0.56	0.026	0.014	0.50	0.50	0.000	0.23	0.01	0.01	0.00	0.00	0.00
$\hat{A}_{\text{opt}}(5)$	0.58	0.026	0.015	0.50	0.50	0.000	0.21	0.01	0.01	0.00	0.00	0.00
$\hat{A}_{\text{opt}}(10)$	0.58	0.026	0.015	0.50	0.50	0.000	0.19	0.01	0.01	0.00	0.00	0.00
$\hat{A}_{\text{opt}}(20)$	0.63	0.026	0.016	0.50	0.50	0.000	0.15	0.01	0.01	0.00	0.00	0.00
Naive estimates												
$\hat{A}_{\text{opt}}(1)$	0.59	0.264	0.155	0.47	0.53	0.057	0.28	0.20	0.15	0.03	0.03	0.04
$\hat{A}_{\text{opt}}(2)$	0.57	0.348	0.198	0.46	0.54	0.071	0.25	0.20	0.16	0.04	0.03	0.04
$\hat{A}_{\text{opt}}(3)$	0.56	0.402	0.227	0.46	0.54	0.077	0.23	0.21	0.17	0.03	0.03	0.04
$\hat{A}_{\text{opt}}(5)$	0.58	0.481	0.282	0.46	0.54	0.084	0.21	0.21	0.18	0.03	0.03	0.03
$\hat{A}_{\text{opt}}(10)$	0.58	0.614	0.363	0.45	0.55	0.099	0.19	0.19	0.18	0.03	0.03	0.03
$\hat{A}_{\text{opt}}(20)$	0.63	0.752	0.477	0.45	0.56	0.110	0.15	0.15	0.16	0.03	0.03	0.02
Cross-validated estimates												
$\hat{A}_{\text{opt}}(1)$	0.58	0.096	0.065	0.50	0.49	−0.011	0.23	0.15	0.12	0.04	0.04	0.06
$\hat{A}_{\text{opt}}(2)$	0.56	0.086	0.060	0.51	0.49	−0.011	0.21	0.14	0.11	0.04	0.04	0.06
$\hat{A}_{\text{opt}}(3)$	0.56	0.085	0.061	0.51	0.49	−0.011	0.19	0.14	0.12	0.03	0.03	0.05
$\hat{A}_{\text{opt}}(5)$	0.57	0.089	0.065	0.51	0.50	−0.009	0.18	0.15	0.12	0.03	0.03	0.05
$\hat{A}_{\text{opt}}(10)$	0.59	0.088	0.066	0.50	0.50	−0.008	0.15	0.14	0.12	0.03	0.03	0.05
$\hat{A}_{\text{opt}}(20)$	0.63	0.090	0.069	0.50	0.50	−0.005	0.13	0.14	0.12	0.03	0.03	0.05
Scenario (b): $\beta_0 = \beta_1 = (0, 0.5, 0.5, \ldots, 0.5)'$												
True values												
A_{opt}	1.00	0.025	0.025	0.50	0.50	0.000						
Reference estimates												
$\hat{A}_{\text{opt}}(1)$	0.57	0.026	0.015	0.50	0.50	0.000	0.27	0.01	0.01	0.06	0.06	0.00
$\hat{A}_{\text{opt}}(2)$	0.55	0.026	0.014	0.50	0.50	0.000	0.24	0.01	0.01	0.08	0.08	0.00
$\hat{A}_{\text{opt}}(3)$	0.55	0.026	0.014	0.50	0.50	0.000	0.24	0.01	0.01	0.09	0.09	0.00
$\hat{A}_{\text{opt}}(5)$	0.55	0.026	0.014	0.50	0.50	0.000	0.22	0.01	0.01	0.10	0.10	0.00
$\hat{A}_{\text{opt}}(10)$	0.55	0.026	0.014	0.50	0.50	0.000	0.19	0.01	0.01	0.13	0.13	0.00
$\hat{A}_{\text{opt}}(20)$	0.53	0.026	0.014	0.50	0.50	0.000	0.16	0.01	0.01	0.21	0.21	0.00
Naive estimates												
$\hat{A}_{\text{opt}}(1)$	0.57	0.276	0.161	0.47	0.53	0.059	0.27	0.20	0.16	0.07	0.07	0.04
$\hat{A}_{\text{opt}}(2)$	0.55	0.344	0.193	0.47	0.53	0.068	0.24	0.21	0.16	0.08	0.09	0.04
$\hat{A}_{\text{opt}}(3)$	0.55	0.387	0.213	0.46	0.54	0.075	0.24	0.21	0.17	0.10	0.10	0.04
$\hat{A}_{\text{opt}}(5)$	0.55	0.446	0.246	0.46	0.54	0.082	0.22	0.21	0.17	0.11	0.11	0.04
$\hat{A}_{\text{opt}}(10)$	0.55	0.535	0.294	0.45	0.54	0.088	0.19	0.19	0.16	0.14	0.14	0.03
$\hat{A}_{\text{opt}}(20)$	0.53	0.645	0.341	0.45	0.55	0.092	0.16	0.16	0.13	0.23	0.23	0.02
Cross-validated estimates												
$\hat{A}_{\text{opt}}(1)$	0.57	0.100	0.068	0.50	0.49	−0.010	0.23	0.16	0.13	0.06	0.06	0.07
$\hat{A}_{\text{opt}}(2)$	0.56	0.097	0.066	0.50	0.49	−0.010	0.21	0.15	0.12	0.07	0.07	0.06
$\hat{A}_{\text{opt}}(3)$	0.55	0.088	0.061	0.51	0.49	−0.013	0.20	0.15	0.12	0.07	0.07	0.06
$\hat{A}_{\text{opt}}(5)$	0.55	0.090	0.062	0.50	0.49	−0.009	0.18	0.15	0.11	0.08	0.08	0.05
$\hat{A}_{\text{opt}}(10)$	0.55	0.083	0.056	0.51	0.50	−0.010	0.15	0.14	0.11	0.09	0.09	0.05
$\hat{A}_{\text{opt}}(20)$	0.55	0.081	0.053	0.51	0.50	−0.006	0.13	0.14	0.10	0.15	0.15	0.04

†The models and the estimates are defined in Section 3.

the true optimum A_{opt} is not available in closed form even though the true distribution of $(\mathbf{X}, T, Y)$ is known. To approximate A_{opt} and its relevant attributes (i.e. F, $\widetilde{\text{pow}}$, g and p_t), we generate another external sample of size 10^5, which we call the external training sample. The same procedure for obtaining $\hat{A}_{\text{opt}}$, with $\hat{p}_t(\mathbf{x})$ replaced by the true values, is applied to the external training sample to obtain $\tilde{A}_{\text{opt}}$. Obviously, $\tilde{A}_{\text{opt}}$ is a random subset due to the randomness in the external training sample, which is manifested in the grid search algorithm. To assess the randomness in $\tilde{A}_{\text{opt}}$, we use the methods that were described earlier to obtain naive estimates (based on the external training sample) and reference estimates (based on the external validation sample) of $F(\tilde{A}_{\text{opt}})$, $\widetilde{\text{pow}}(\tilde{A}_{\text{opt}})$, $g(\tilde{A}_{\text{opt}})$ and $p_t(\tilde{A}_{\text{opt}})$ $(t = 0, 1)$. This procedure is repeated 100 times for 100 independent instances of $\tilde{A}_{\text{opt}}$. If $\tilde{A}_{\text{opt}}$ stays close to A_{opt} with little variability, then the naive and reference estimates should be close to each other and both sets of estimates should have small variability.

Table 2 shows the empirical means and standard deviations of the reference, naive and cross-validated estimates of $F(A)$, $\widetilde{\text{pow}}(A)$, $g(A)$, $p_0(A)$, $p_1(A)$ and $\delta_p(A)$, where A is either $\tilde{A}_{\text{opt}}$ or $\hat{A}_{\text{opt}}$ (based on some working model). The working model for $\hat{A}_{\text{opt}}$ takes the form of equation (3) but possibly excludes some elements of $\mathbf{X}$. Specifically, we consider working models based on $\mathbf{X}_{[1:20]} = \mathbf{X}$, $\mathbf{X}_{[1:10]}$ (the first 10 elements of $\mathbf{X}$), $\mathbf{X}_{[1:5]}$, $\mathbf{X}_{[1:3]}$, $\mathbf{X}_{[1:2]}$, $\mathbf{X}_{[1]}$, $\mathbf{X}_{[4:20]}$, $\mathbf{X}_{[6:20]}$ and $\mathbf{X}_{[11:20]}$. The first two models are correct, and the rest are incorrect. In what follows, we interpret the results in Table 2 in a stepwise manner.

First, we note that the reference and naive estimates for $\tilde{A}_{\text{opt}}$ are quite close to each other on average, and that both sets of estimates have very small standard deviations in comparison with the corresponding estimates for $\hat{A}_{\text{opt}}$ (with any working model). As indicated earlier, these observations support the use of $\tilde{A}_{\text{opt}}$ as an approximation to A_{opt} and as a gold standard in Table 2.

Second, we compare the reference estimates in Table 2 to assess the performance of the subgroup selection procedure proposed and the effect of covariate omission. Among the various versions of $\hat{A}_{\text{opt}}$ in Table 2, the first three have the highest mean values of $\hat{g}^{\text{ref}}(\hat{A}_{\text{opt}})$. These mean values (0.69–0.71) are, of course, lower than the corresponding value (0.78) for $\tilde{A}_{\text{opt}}$; however, the difference is not unacceptably large considering that $\tilde{A}_{\text{opt}}$ is based on the true value of $p_t(\mathbf{x})$ and an external training sample that is 100 times as large as the sample from which $\hat{A}_{\text{opt}}$ is obtained. The first two models (based on $\mathbf{X}$ and $\mathbf{X}_{[1:10]}$) are correctly specified, with the second being more parsimonious. The third model (based on $\mathbf{X}_{[1:5]}$) is technically incorrect but includes all effect modifiers. For these three models, the mean reference estimates of $F(\hat{A}_{\text{opt}})$, $p_0(\hat{A}_{\text{opt}})$, $p_1(\hat{A}_{\text{opt}})$ and $\delta_p(\hat{A}_{\text{opt}})$ are quite similar to the corresponding values for $\tilde{A}_{\text{opt}}$, indicating that the procedure performs reasonably well provided that all effect modifiers are included. The performance of the procedure deteriorates when effect modifiers are excluded from the working model, and appears insensitive to inclusion or exclusion of other variables. The last two models (based on $\mathbf{X}_{[6:20]}$ and $\mathbf{X}_{[11:20]}$) contain no effect modifiers and prove useless for subgroup selection.

Next, we compare the naive estimates with the reference estimates for the various versions of $\hat{A}_{\text{opt}}$ to gauge the resubstitution bias in a sample of size 1000. The differences (in mean) between the two groups of estimates are relatively large for $\widetilde{\text{pow}}$, g and δ_p, indicating that the naive estimates tend to be overly optimistic with respect to the power, the expected gain and the treatment effect size for the subgroup selected.

Finally, we compare the cross-validated estimates with the reference estimates for the various versions of $\hat{A}_{\text{opt}}$ to assess the effectiveness of the cross-validation procedure for reducing the resubstitution bias. This comparison is focused on $\widetilde{\text{pow}}$, g and δ_p, which are the quantities that are most susceptible to a resubstitution bias. For all three quantities, Table 2 shows clearly that,

Table 2. Results of the second simulation experiment (under the alternative hypothesis): empirical means and standard deviations of reference, naive and cross-validated estimates of $F(A)$, $\widetilde{\text{pow}}(A)$, $g(A)$, $p_0(A)$, $p_1(A)$ and $\delta_p(A) = p_1(A) - p_0(A)$, where A is either A_{opt} or $\hat{A}_{\text{opt}}$ (based on a specified collection of covariates; for example, 1:10 refers to the first 10 elements of $\mathbf{X}$)†

Subset A (covariates)	*Empirical means*						*Empirical standard deviations*					
	$F(A)$	$\widetilde{pow}(A)$	$g(A)$	$p_0(A)$	$p_1(A)$	$\delta_p(A)$	$F(A)$	$\widetilde{pow}(A)$	$g(A)$	$p_0(A)$	$p_1(A)$	$\delta_p(A)$
Reference estimates												
$\tilde{A}_{\text{opt}}$	0.81	0.96	0.78	0.43	0.56	0.130	0.01	0.01	0.01	0.00	0.00	0.01
$\hat{A}_{\text{opt}}(1{:}20)$	0.82	0.85	0.69	0.44	0.56	0.113	0.05	0.15	0.10	0.02	0.02	0.03
$\hat{A}_{\text{opt}}(1{:}10)$	0.82	0.87	0.71	0.44	0.56	0.118	0.05	0.14	0.09	0.02	0.02	0.03
$\hat{A}_{\text{opt}}(1{:}5)$	0.81	0.87	0.69	0.44	0.56	0.120	0.06	0.15	0.10	0.02	0.02	0.03
$\hat{A}_{\text{opt}}(1{:}3)$	0.76	0.81	0.61	0.44	0.56	0.114	0.08	0.18	0.10	0.02	0.02	0.04
$\hat{A}_{\text{opt}}(1{:}2)$	0.72	0.76	0.53	0.45	0.55	0.108	0.10	0.20	0.10	0.02	0.02	0.04
$\hat{A}_{\text{opt}}(1)$	0.64	0.63	0.38	0.45	0.55	0.097	0.14	0.21	0.08	0.02	0.02	0.04
$\hat{A}_{\text{opt}}(4{:}20)$	0.74	0.67	0.48	0.45	0.55	0.091	0.08	0.18	0.10	0.03	0.03	0.03
$\hat{A}_{\text{opt}}(6{:}20)$	0.57	0.03	0.01	0.50	0.50	0.000	0.17	0.01	0.01	0.07	0.07	0.00
$\hat{A}_{\text{opt}}(11{:}20)$	0.58	0.03	0.01	0.50	0.50	0.000	0.18	0.01	0.01	0.00	0.00	0.00
Naive estimates												
$\tilde{A}_{\text{opt}}$	0.81	0.96	0.78	0.43	0.57	0.131	0.01	0.00	0.01	0.00	0.00	0.00
$\hat{A}_{\text{opt}}(1{:}20)$	0.82	0.97	0.80	0.44	0.57	0.133	0.05	0.01	0.05	0.02	0.02	0.01
$\hat{A}_{\text{opt}}(1{:}10)$	0.82	0.97	0.79	0.43	0.57	0.132	0.05	0.02	0.06	0.02	0.02	0.01
$\hat{A}_{\text{opt}}(1{:}5)$	0.81	0.97	0.78	0.43	0.57	0.133	0.06	0.02	0.06	0.02	0.02	0.01
$\hat{A}_{\text{opt}}(1{:}3)$	0.76	0.94	0.72	0.44	0.56	0.129	0.09	0.04	0.09	0.02	0.02	0.01
$\hat{A}_{\text{opt}}(1{:}2)$	0.72	0.91	0.65	0.44	0.56	0.125	0.10	0.06	0.11	0.02	0.02	0.01
$\hat{A}_{\text{opt}}(1)$	0.64	0.79	0.51	0.44	0.56	0.115	0.14	0.14	0.16	0.02	0.02	0.02
$\hat{A}_{\text{opt}}(4{:}20)$	0.74	0.91	0.68	0.44	0.56	0.124	0.08	0.05	0.10	0.04	0.03	0.01
$\hat{A}_{\text{opt}}(6{:}20)$	0.58	0.64	0.37	0.46	0.56	0.100	0.17	0.17	0.16	0.09	0.09	0.03
$\hat{A}_{\text{opt}}(11{:}20)$	0.58	0.61	0.36	0.45	0.55	0.098	0.18	0.19	0.17	0.03	0.03	0.03
Cross-validated estimates												
$\hat{A}_{\text{opt}}(1{:}20)$	0.82	0.86	0.71	0.45	0.56	0.111	0.05	0.13	0.12	0.02	0.02	0.02
$\hat{A}_{\text{opt}}(1{:}10)$	0.82	0.89	0.73	0.44	0.56	0.116	0.05	0.11	0.11	0.02	0.02	0.02
$\hat{A}_{\text{opt}}(1{:}5)$	0.81	0.91	0.74	0.44	0.56	0.118	0.06	0.07	0.09	0.02	0.02	0.01
$\hat{A}_{\text{opt}}(1{:}3)$	0.76	0.84	0.65	0.45	0.56	0.110	0.08	0.11	0.13	0.02	0.02	0.01
$\hat{A}_{\text{opt}}(1{:}2)$	0.72	0.77	0.57	0.45	0.55	0.104	0.10	0.15	0.15	0.02	0.02	0.02
$\hat{A}_{\text{opt}}(1)$	0.64	0.61	0.41	0.45	0.54	0.090	0.13	0.23	0.20	0.02	0.02	0.03
$\hat{A}_{\text{opt}}(4{:}20)$	0.74	0.63	0.48	0.46	0.54	0.085	0.08	0.23	0.19	0.03	0.03	0.02
$\hat{A}_{\text{opt}}(6{:}20)$	0.59	0.09	0.06	0.51	0.50	−0.007	0.14	0.14	0.11	0.06	0.06	0.05
$\hat{A}_{\text{opt}}(11{:}20)$	0.59	0.08	0.06	0.51	0.50	−0.009	0.15	0.14	0.11	0.03	0.03	0.05

†The models and the estimates are defined in Section 3.

compared with the naive estimates, the cross-validated estimates are generally closer (in mean) to the reference estimates. Although the cross-validated estimates may be slightly biased in some cases, they are usually less biased than the naive estimates.

The Web appendix D presents the results of a third simulation experiment, which indicate that mild misspecification of the working model for $p_t(\mathbf{x})$ has minimal effect on subgroup selection. For the second and third simulation experiments, the Web appendix D also provides information on $\psi_+(A)$ and $\psi_-(A)$ (which were defined in remark 2) with $A = \tilde{A}_{\text{opt}}$ or $\hat{A}_{\text{opt}}$, and the results suggest that $\tilde{A}_{\text{opt}}$ (or a 'good' $\hat{A}_{\text{opt}}$) usually comprises nearly all patients in $\mathcal{X}_+$ as well as a fair proportion of patients in $\mathcal{X}_-$. As noted in remark 2, a new treatment that benefits

a (sub)population of patients as a whole does not necessarily benefit every individual patient in that (sub)population. That *caveat* should be kept in mind when interpreting a significant test result from the approach proposed (or any other approach) establishing treatment efficacy for a (sub)population of patients.

In summary, the simulation results show that the proposed procedure for subgroup selection is effective and fairly robust, although its performance can be adversely affected by the omission of important effect modifiers and by severe misspecification of the working model for $p_t(\mathbf{x})$. The simulation results also demonstrate that estimation of $\widetilde{\text{pow}}$, g and δ_p for the subgroup selected is susceptible to a possibly substantial resubstitution bias, which can be reduced by using the cross-validation procedure proposed. In all cases that were considered, it takes less than 1 min to analyse a sample (with cross-validation but without bootstrapping) on an ordinary laptop computer.

4. Examples

4.1. The 'Magnesium in coronaries' study

As an illustrative example, consider the 'Magnesium in coronaries' (MAGIC) study, which was a randomized double-blind clinical trial that was designed to investigate, in high-risk patients with ST-elevation myocardial infarction, the effect of supplemental administration of intravenous magnesium on short-term mortality (MAGIC Trial Investigators, 2002). The MAGIC study enrolled 6213 patients with acute ST-elevation myocardial infarction at 278 sites in 14 countries and randomized them in a 1:1 ratio to receive a 2-g intravenous bolus of magnesium sulphate administered over 15 min, followed by a 17-g infusion of magnesium sulphate over 24 h, or matching placebo. Patients in both groups also received standard care. The randomization was stratified by patient eligibility for reperfusion therapy. Specifically, stratum 1 included patients who were at least 65 years old and eligible for reperfusion therapy, and stratum 2 included patients of any age who were not eligible for reperfusion therapy. The primary end point of the MAGIC study was all-cause mortality within 30 days of randomization. Vital status at 30 days was ascertained from direct patient contact, their medical record or a death certificate. The observed mortality rate was 15.3% in the magnesium group and 15.2% in the placebo group, with an odds ratio of 1.0 (95% confidence interval, 0.9–1.2). No benefit or harm from magnesium was observed in eight prespecified and 15 exploratory subgroup analyses. An empirical Bayes analysis of the MAGIC study data also failed to find a subgroup with a significant treatment effect (Shen *et al.*, 2015).

In our retrospective analysis, the baseline covariate vector $\mathbf{X}$ consists of stratum (as defined above), age, gender, systolic blood pressure, heart rate, a simple risk index (Morrow *et al.*, 2001), a modified thrombolysis in myocardial infarction score, history of hypertension, diabetes, myocardial infarction, congestive heart failure, coronary artery bypass grafting or percutaneous transluminal coronary angioplasty, stroke, clinical evidence of pulmonary congestion, infarct location and time from myocardial infarction symptom onset to start of bolus injection. Three subjects with missing covariate data are excluded from our analysis. To comply with the notation in Section 2, we work with 30-day survival as the outcome of interest (so $Y = 1$ if a patient is alive at 30 days). We estimate $p_t(\mathbf{x})$ under the logistic regression model (3) and estimate $p_t(A)$ with the augmented estimator $\hat{p}_t^{\text{aug}}(A)$. Our analysis is based on superiority hypotheses (with $\delta_0 = 0$), one-sided $\alpha = 0.025$, and a constant utility function, and involves grid search and 20-fold cross-validation (as in remark 5). As indicated in remark 4, a severe departure of $\hat{A}_{\text{opt}}$ from A_{opt} could be detected by comparing $\hat{\mathbf{c}}_{\text{opt}}$ with $\mathbf{c}(\hat{A}_{\text{opt}})$. The Euclidean distance between the two vectors is 0.047 in this example, with a non-parametric bootstrap standard error of 0.104

Table 3. Analysis of the MAGIC study data: naive and cross-validated estimates of $F(\hat{A}_{\text{opt}})$, $\widetilde{\text{pow}}(\hat{A}_{\text{opt}})$, $g(\hat{A}_{\text{opt}})$, $p_0(\hat{A}_{\text{opt}})$, $p_1(\hat{A}_{\text{opt}})$ and $\delta_p(\hat{A}_{\text{opt}}) = p_1(\hat{A}_{\text{opt}}) - p_0(\hat{A}_{\text{opt}})$, together with bootstrap standard errors based on 1000 bootstrap samples†

Quantity of interest	*Point estimate (%)*		*Standard error (%)*	
	Naive	*Cross-validated*	*Naive*	*Cross-validated*
$F(\hat{A}_{\text{opt}})$	63.1	63.8	13.8	12.1
$\widetilde{\text{pow}}(\hat{A}_{\text{opt}})$	73.8	0.7	12.2	23.2
$g(\hat{A}_{\text{opt}})$	46.5	0.5	14.9	20.1
$p_0(\hat{A}_{\text{opt}})$	86.7	87.5	1.9	1.6
$p_1(\hat{A}_{\text{opt}})$	89.4	86.9	1.8	1.3
$\delta_p(\hat{A}_{\text{opt}})$	2.7	−0.5	0.6	0.9

†See Section 4.1 for details.

(based on 1000 bootstrap samples). The parametric bootstrap test that was described in remark 4 yields a p-value of 0.87 (based on 1000 bootstrap samples), which does not indicate a severe departure of $\hat{A}_{\text{opt}}$ from A_{opt} due to model misspecification.

Table 3 presents the naive and cross-validated estimates of $F(\hat{A}_{\text{opt}})$, $\widetilde{\text{pow}}(\hat{A}_{\text{opt}})$, $g(\hat{A}_{\text{opt}})$, $p_0(\hat{A}_{\text{opt}})$, $p_1(\hat{A}_{\text{opt}})$ and $\delta_p(\hat{A}_{\text{opt}}) = p_1(\hat{A}_{\text{opt}}) - p_0(\hat{A}_{\text{opt}})$, together with boostrap standard errors based on 1000 bootstrap samples. The naive estimates in Table 3 are generally more optimistic than the corresponding cross-validated estimates, which are consistent with the simulation results in Section 3. The naive estimate of the treatment difference (magnesium − placebo) in the selected subpopulation is 2.7%, which is significantly greater than 0 ($p < 0.0001$). On cross-validation, the treatment difference estimate decreases to −0.5%, which is not significantly greater than 0 ($p = 0.71$). Thus, our analysis of the MAGIC trial data is consistent with previous analyses in that no significant treatment benefit is found in any subgroup.

4.2. The haemodialysis study

Our second example is the haemodialysis study, which was a randomized multicentre clinical trial to evaluate the effects of the dose of dialysis and the level of flux of the dialyser membrane on mortality and morbidity among patients undergoing maintenance haemodialysis (Eknoyan *et al.*, 2002). The haemodialysis study enrolled 1846 patients undergoing thrice-weekly dialysis at 15 clinical centres (associated with 72 dialysis units) and randomized them to a standard or high dose of dialysis (1:1) and to a low flux or high flux dialyser (1:1) under a 2×2 factorial design. The standard and high doses of dialysis were defined by target values of 1.05 and 1.45 respectively for the equilibrated Kt/V. The flux of a dialyser was considered low if the mean beta$_2$-microglobulin clearance was less than 10 ml min^{-1}, and as high if the ultrafiltration coefficient was more than 14 ml h^{-1} mm Hg^{-1} and the mean beta$_2$-microglobulin clearance was more than 20 ml min^{-1}. The primary end point of the haemodialysis study was the time to death from any cause, which was not significantly influenced by the dose or flux assignment: the hazard ratio for high *versus* standard dose was estimated to be 0.96 (95% confidence interval, 0.84–1.10; $p = 0.53$), and the hazard ratio for high *versus* low flux was estimated to be 0.92 (95% confidence interval 0.81–1.05; $p = 0.23$). However, possible interactions were identified between dose and sex (unadjusted $p = 0.01$) and between flux and prior years of dialysis (3.7 years or less *versus* more than 3.7 years; unadjusted $p = 0.005$). The corresponding subgroup analyses

suggested that women might benefit from a high dose of dialysis and that patients with a longer history of dialysis might benefit from high flux. Although definitive answers to these questions would require preplanned analyses, we present a retrospective analysis here mainly to illustrate the methodology proposed.

In this illustration, the treatment of interest is the level of flux (1, high; 0, low), and the baseline covariate vector $\mathbf{X}$ consists of the seven covariates prespecified for subgroup analyses and also included in the primary (Cox regression) analysis: sex (1, male; 2, female), duration (prior years of dialysis), race (1, black; 0, non-black), diabetic status (1, yes; 0, no), age (1, greater than 58 years; 0, 58 years or less), albumin (1, 3.6 mg dL^{-1} or more; 0 less than 3.6 mg dL^{-1}) and comorbidity (1, high; 0, low). We work with survival status (1, alive; 0, dead) at 3 years post randomization as the outcome variable, and we restrict attention to the 1414 subjects who were randomized at least 3 years before the administrative end of the study. These 1414 subjects have complete covariate, treatment and outcome data. A logistic regression model of the form (3) is used to estimate $p_t(\mathbf{x})$, separately for each $t \in \{0, 1\}$. Our analysis of this example is similar to the previous analysis except for the use of one-sided $\alpha = 0.05$ and tenfold cross-validation. The selected subset is given by

$$\hat{A}_{\text{opt}} = \{\mathbf{x} \in \mathcal{X} : (1, \hat{p}_0(\mathbf{x}), \hat{p}_1(\mathbf{x}))\hat{\mathbf{c}}_{\text{opt}} \leqslant 0\},$$

where $\hat{\mathbf{c}}_{\text{opt}} \approx (0.20, 0.49, -0.85)'$ and $\hat{p}_t(\mathbf{x}) = \text{expit}\{(1, \mathbf{x}')\hat{\boldsymbol{\beta}}_t\}$ with

$$\hat{\boldsymbol{\beta}}_0 \approx (1.41, 0.00, -0.06, 0.31, -0.14, -0.96, 0.65, -0.78)',$$
$$\hat{\boldsymbol{\beta}}_1 \approx (0.66, 0.28, 0.00, 0.57, -0.22, -1.03, 0.56, -0.53)'.$$

The Euclidean distance between $\hat{\mathbf{c}}_{\text{opt}}$ and $\mathbf{c}(\hat{A}_{\text{opt}})$, which is approximately $(0.01, 0.69, -0.72)'$, is 0.30, with a non-parametric bootstrap standard error of 0.31 (based on 1000 bootstrap samples). The parametric bootstrap test that was described in remark 4 yields a p-value of 0.24 (based on 1000 bootstrap samples), which does not indicate a severe departure of $\hat{A}_{\text{opt}}$ from A_{opt} due to model misspecification.

Table 4 presents the naive and cross-validated estimates of $F(\hat{A}_{\text{opt}})$, $\widetilde{\text{pow}}(\hat{A}_{\text{opt}})$, $g(\hat{A}_{\text{opt}})$, $p_0(\hat{A}_{\text{opt}})$, $p_1(\hat{A}_{\text{opt}})$ and $\delta_p(\hat{A}_{\text{opt}})$, together with boostrap standard errors based on 1000 bootstrap samples. The naive estimate of the treatment difference in the selected subpopulation is

Table 4. Analysis of the haemodialysis study data: naive and cross-validated estimates of $F(\hat{A}_{\text{opt}})$, $\widetilde{\text{pow}}(\hat{A}_{\text{opt}})$, $g(\hat{A}_{\text{opt}})$, $p_0(\hat{A}_{\text{opt}})$, $p_1(\hat{A}_{\text{opt}})$ and $\delta_p(\hat{A}_{\text{opt}}) = p_1(\hat{A}_{\text{opt}}) - p_0(\hat{A}_{\text{opt}})$, together with bootstrap standard errors based on 1000 bootstrap samples†

Quantity of interest	*Point estimate (%)*		*Standard error (%)*	
	Naive	*Cross-validated*	*Naive*	*Cross-validated*
$F(\hat{A}_{\text{opt}})$	76.2	79.8	12.5	10.3
$\widetilde{\text{pow}}(\hat{A}_{\text{opt}})$	63.9	33.7	9.6	21.9
$g(\hat{A}_{\text{opt}})$	48.7	26.9	13.6	21.1
$p_0(\hat{A}_{\text{opt}})$	65.9	65.8	3.5	2.6
$p_1(\hat{A}_{\text{opt}})$	71.6	69.2	3.4	2.4
$\delta_p(\hat{A}_{\text{opt}})$	5.6	3.4	1.2	2.0

†See Section 4.2 for details.

5.6%, which is significantly greater than 0 ($p < 0.0001$). On cross-validation, the treatment difference estimate decreases to 3.4% and the associated standard error increases, but the difference is still significantly greater than 0 ($p = 0.044$). Thus, if this analysis had been planned prospectively, it would have successfully demonstrated the benefit of high flux for the selected subpopulation.

5. Discussion

When designing a confirmatory clinical trial, investigators who are interested in planning a subgroup analysis prospectively often do not have enough information to define the subgroup *a priori*. Adaptive signature designs offer much flexibility in that regard by allowing investigators to use the trial data to select a subgroup and to test for treatment efficacy for the selected subgroup at the end of a broad eligibility trial. In this paper, we have provided a simple characterization of the optimal subgroup, which maximizes the power for demonstrating treatment efficacy or the expected gain based on a specified utility function. We have also proposed a three-step procedure for estimating or approximating the optimal subgroup, as well as a cross-validation procedure for estimating the treatment effect on the subgroup selected. Simulation results show that the subgroup selection procedure proposed is effective and fairly robust against model misspecification, and that the cross-validation procedure proposed reduces the resubstitution bias effectively. Unlike the traditional approach to adaptive signature designs, which involves two separate tests, the approach allows us to focus on one test for a selected subpopulation, which could be the entire population (see remark 1). Other advantages of the approach include its insensitivity to any collinearity in $\mathbf{X}$ (see remark 3) and the use of augmented inverse probability weighting to incorporate covariate information and to improve precision.

Our simulation studies and examples do not involve variable selection or model selection. In practice, some applications may require the use of a variable selection procedure to choose between a large number of candidate covariates (i.e. gene expression data). A model selection procedure may also be required if the model for $p_t(\mathbf{x})$ cannot be fully specified in the design stage. Our proposed approach can incorporate variable and model selection procedures as long as they are fully specified in the analysis plan. In such cases, it is important to include the variable or model selection in the cross-validation and bootstrap procedures.

Acknowledgements

We thank two reviewers for insightful and constructive comments that have improved the manuscript greatly. The views that are expressed in this paper do not represent the official position of the US Food and Drug Administration. This manuscript was prepared by using MAGIC research materials that were obtained from the National Heart, Lung and Blood Institute Biologic Specimen and Data Repository Information Coordinating Center and does not necessarily reflect the opinions or views of the MAGIC Trial Investigators or the National Heart, Lung and Blood Institute.

References

Chen, J. J., Lu, T.-P., Chen, D. T. and Wang, S.-J. (2014) Biomarker adaptive designs in clinical trials. *Transltnl Cancer Res.*, **3**, 279–292.

Dawid, A. P. (1994) Selection paradoxes of Bayesian inference. *Multiv. Anal. Appl.*, **24**, 211–220.

Eknoyan, G., Beck, G. J., Cheung, A. K., Daugirdas, J. T., Greene, T., Kusek, J. W., Allon, M., Bailey, J., Delmez, J. A., Depner, T. A., Dwyer, J. T., Levey, A. S., Levin, N. W., Milford, E., Ornt, D. B., Rocco, M. V., Schulman, G., Schwab, S. J., Teehan, B. P. and Toto, R. for the Hemodialysis (HEMO) Study Group (2002) Effect of dialysis dose and membrane flux in maintenance hemodialysis. *New Engl. J. Med.*, **347**, 2010–2019.

Follmann, D. (1997) Adaptively changing subgroup proportions in clinical trials. *Statist. Sin.*, **7**, 1085–1102.
Freidlin, B., Jiang, W. and Simon, R. (2010) The cross-validated adaptive signature design. *Clin. Cancer Res.*, **16**, 691–698.
Freidlin, B. and Simon, R. (2005) Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clin. Cancer Res.*, **11**, 7872–7878.
Gail, M. and Simon, R. (1985) Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*, **41**, 361–372.
Jiang, W., Freidlin, B. and Simon, R. (2007) Biomarker-adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subject effect. *J. Natn. Cancer Inst.*, **13**, 1036–1043.
Liu, A., Li, Q., Liu, C., Yu, K. F. and Yuan, V. W. (2010) A threshold sample enrichment approach in a clinical trial with heterogeneous subpopulations. *Clin. Trials*, **7**, 537–545.
MAGIC Trial Investigators (2002) Early administration of intravenous magnesium to high-risk patients with acute myocardial infarction in the Magnesium in Coronaries (MAGIC) Trial: a randomised controlled trial. *Lancet*, **360**, 1189–1196.
Maitournam, A. and Simon, R. (2005) On the efficiency of targeted clinical trials. *Statist. Med.*, **24**, 329–339.
Morrow, D. A., Antman, E. M., Giugliano, R. P., Cairns, R., Charlesworth, A., Murphy, S. A., de Lemos, J. A., McCabe, C. H. and Braunwald, E. (2001) A simple risk index for rapid initial triage of patients with ST-elevation myocardial infarction: an InTIMEII substudy. *Lancet*, **358**, 1571–1575.
Qian, M. and Murphy, S. A. (2011) Performance guarantees for individualized treatment rules. *Ann. Statist.*, **39**, 1180–1210.
Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994) Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Ass.*, **89**, 846–866.
Rosenblum, M. and van der Laan, M. J. (2011) Optimizing randomized trial designs to distinguish which subpopulations benefit from treatment. *Biometrika*, **98**, 845–860.
Rubin, D. B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, **66**, 688–701.
Russek-Cohen, E. and Simon, R. M. (1997) Evaluating treatments when a gender by treatment interaction may exist. *Statist. Med.*, **16**, 455–464.
Senn, S. (2008) A note concerning a selection "paradox" of Dawid's. *Am. Statistn*, **62**, 206–210.
Shen, C., Li, X. and Jeong, J. (2015) Estimation of treatment effect in a subpopulation: an empirical Bayes approach. *J. Biopharm. Statist.*, **26**, 507–518.
Simon, R. (2008) Development and validation of biomarker classifiers for treatment selection. *J. Statist. Planng Inf.*, **138**, 308–320.
Simon, R. (2010) Clinical trials for predictive medicine: new challenges and paradigms. *Clin. Trials*, **7**, 516–524.
Simon, R. and Maitournam, A. (2004) Evaluating the efficiency of targeted designs for randomized clinical trials. *Clin. Cancer Res.*, **10**, 6759–6763.
Simon, N. and Simon, R. (2013) Adaptive enrichment designs for clinical trials. *Biostatistics*, **14**, 613–625.
Tsiatis, A. A. (2006) *Semiparametric Theory and Missing Data*. New York: Springer.
US Food and Drug Administration (2012) *Guidance for Industry: Enrichment Strategies for Clinical Trials to Support Approval of Human Drugs and Biological Products*. Washington DC: Department of Health and Human Services.
Wang, S.-J. and Hung, H. M. J. (2013) Adaptive enrichment with subpopulation selection at interim: methodologies, applications and design considerations. *Contemp. Clin. Trials*, **36**, 673–681.
Wang, S.-J., O'Neill, R. T. and Hung, H. M. J. (2007) Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharm. Statist.*, **6**, 227–244.
Zhang, B., Tsiatis, A. A., Laber, E. B. and Davidian, M. (2012) A robust method for estimating optimal treatment regimes. *Biometrics*, **68**, 1010–1018.
Zhang, Z., Wang, C., Nie, L. and Soon, G. (2013) Assessing the heterogeneity of treatment effects via potential outcomes of individual patients. *Appl. Statist.*, **62**, 687–704.
Zhao, Y., Zeng, D., Rush, J. and Kosorok, M. R. (2012) Estimating individualized treatment rules using outcome weighted learning. *J. Am. Statist. Ass.*, **107**, 1106–1118.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Web-based supplementary materials for "Subgroup selection in adaptive signature designs of confirmatory clinical trials"'.