



Available at
www.elsevierMathematics.com
POWERED BY SCIENCE @ DIRECT[®]
Journal of Statistical Planning and
Inference 127 (2005) 323–340

journal of
statistical planning
and inference

www.elsevier.com/locate/jspi

Robust allocation schemes for clinical trials with prognostic factors[☆]

Douglas P. Wiens*

*Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton,
Alberta, Canada T6G 2G1.*

Received 19 June 2002; accepted 9 August 2003

Abstract

We present schemes for the allocation of subjects to treatment groups, in the presence of prognostic factors. The allocations are robust against incorrectly specified regression responses, and against possible heteroscedasticity. Assignment probabilities which minimize the asymptotic variance are obtained. Under certain conditions these are shown to be minimax (with respect to asymptotic mean squared error) as well. We propose a method of sequentially modifying the associated assignment rule, so as to address both variance and bias in finite samples. The resulting scheme is assessed in a simulation study. We find that, relative to common competitors, the robust allocation schemes can result in significant decreases in the mean squared error when the fitted models are biased, at a minimal cost in efficiency when in fact the fitted models are correct.

© 2003 Elsevier B.V. All rights reserved.

MSC: primary 62K05; 62F35; 92C50; secondary 62J05

Keywords: Balance; Biased coin design; M-estimation; Minimax; Optimal design; Sequential allocation

1. Introduction

In this article we pose a generalization of the following problem of interest in clinical trials, and study its solution. Suppose that n subjects enter a clinical study and are to be assigned to one of the two treatment groups, referred to here as ‘treatment’ and ‘control’. Corresponding to each subject is a vector \mathbf{x} of prognostic factors or covariates, upon which the mean response relies through a vector $\mathbf{z}(\mathbf{x})$ of possible regressors. Upon

[☆]This research is supported by the Natural Sciences and Engineering Research Council of Canada.

* Tel.: +1-780-492-4406; fax: +1-780-492-6826.

E-mail address: doug.wiens@ualberta.ca (D.P. Wiens).

observing \mathbf{x} , the experimenter is to assign the subject to one of the groups. The aim is to obtain an efficient and robust estimate of the difference in the mean responses to the treatments.

A complicating factor calling for a robust solution is that the fitted model relating the response to treatment/covariate effects may be only approximately valid. We entertain two such approximate models:

$$\text{Model 1: } Y = \theta_1 u_1 + \theta_2 u_2 + n^{-1/2} f_{\mathbf{u}}(\mathbf{x}) + \sigma_{\mathbf{u}} \varepsilon,$$

$$\text{Model 2: } Y = \theta_1 u_1 + \theta_2 u_2 + \mathbf{z}^T(\mathbf{x})\phi + n^{-1/2} f_{\mathbf{u}}(\mathbf{x}) + \sigma_{\mathbf{u}} \varepsilon.$$

In each case $\mathbf{u} = (u_1, u_2)^T = (1, 0)^T$ for the treatment group, $\mathbf{u} = (0, 1)^T$ for the control group and ε denotes random error with zero mean and unit variance. We are interested primarily in the parameter $\theta_1 - \theta_2$. The function $f_{\mathbf{u}}(\mathbf{x})$ is unknown to the experimenter and serves to formalize the approximate nature of the fitted models, which are $\hat{Y} = \hat{\theta}_1 u_1 + \hat{\theta}_2 u_2$ and $\hat{Y} = \hat{\theta}_1 u_1 + \hat{\theta}_2 u_2 + \mathbf{z}^T(\mathbf{x})\hat{\phi}$, respectively.

Section 2 of this article concerns static, i.e. nonsequential, allocations in an asymptotic framework. In Section 3 we consider a sequential implementation scheme.

We show that the existence and properties of $f_{\mathbf{u}}$ follow from a natural identifiability condition on the regression parameters. The purpose of the $n^{-1/2}$ in the models, which has no effect in finite samples since it can be absorbed by f , is to ensure that bias and variance are of the same order asymptotically.

Throughout, $\hat{\theta}$ denotes a least-squares estimate. Our results would remain essentially unchanged, although there would be some minor complications in the derivations, if these estimates were to be replaced by (ordinary) M-estimates. This is a consequence of the asymptotics in Maronna and Yohai (1981).

The presence of $f_{\mathbf{u}}(\mathbf{x})$ biases the estimate $\hat{\theta}_1 - \hat{\theta}_2$, so that an optimal assignment of subjects to groups should aim at the minimization of the mean squared error (mse) rather than merely the variance. Since $f_{\mathbf{u}}(\cdot)$ is unknown, we shall propose a method for sequentially estimating this quantity.

Ethical considerations often dictate that there should be some randomness in the assignment of subjects to treatments. Thus, we shall derive a probability $\rho(\mathbf{x})$ according to which a subject exhibiting covariates \mathbf{x} is to be assigned to the treatment group. The actual assignment is then made after carrying out a Bernoulli trial with success probability $\rho(\mathbf{x})$.

In Section 2 we present a generalization of this problem to $p \geq 2$ treatment groups, and give a formal definition of the neighbourhood structure referred to above. To express the loss associated with this problem, we first obtain the normalized mse matrix of a complete set of contrasts of the regression parameters, conditional on the covariates and on the assignments made. We then evaluate the almost sure limit of the determinant of this matrix. Optimal assignment probabilities are defined as the minimizers of this loss function.

Atkinson (1982) obtained D-optimal designs for problems similar to those considered here, but assuming the fitted models to be exactly correct. Thus variance minimization was the only concern. Heckman (1987) considered a problem closer in nature to ours. Her approach assumed $p = 2$ and a single continuous regressor, i.e. $\mathbf{z}(\mathbf{x}) = x$ in the

notation above. She assumed a neighbourhood structure somewhat less broad than ours, and obtained allocations which were *minimax* when employed with estimates $\hat{\theta}$ which were themselves minimax, for the given neighbourhood structure, within the class of linear estimates.

In Section 2 we show that, if minimization of the asymptotic variance is the goal, then the optimal probabilities of the assignments to the treatment groups are *constant*, i.e. independent of the covariates. This is surprising in the context of Model 2, since we make no assumptions implying that the regression and treatment effect estimates are uncorrelated. It is however not unprecedented. A precedent is in the aforementioned minimax approach of Heckman (1987), discussed at the end of Section 2. Under certain side conditions determining f_u , constant assignment probabilities are minimax (with respect to asymptotic mse) for our version of this problem as well.

The variance minimizing probabilities should be viewed only as asymptotic goals, rather than as exact finite-sample prescriptions. Also, asymptotic bias is of course an equal contributor to mean squared error. Thus, in Section 3, a sequential approach to making the assignments, similar to that proposed by Atkinson (1982), is outlined. Here the goal is to adjust sequentially for bias and variation, as well as to guide the empirical assignment frequencies towards their intended asymptotic values. This is in contrast to proposals of Taves (1974), Pocock and Simon (1975), Begg and Iglewicz (1980) and others in which interest centres on the maintenance of balance across treatment groups and covariates. These proposals in turn seek to improve on the commonly used *permuted block* design, which divides the experiment into blocks and within each block assigns units to treatment in a random but balanced manner.

Efron (1971, p. 404) remarks that permuted blocks “can be quite effective in eliminating unbalanced designs but... suffer from the disadvantage that at certain points in the experiment the experimenter knows for certain whether the next subject will be assigned as a treatment or as a control”. This raise the issue of the role of *selection bias* in sequential clinical trials. Atkinson (2002) defines this as the probability of correctly guessing which treatment is to be allocated next, and compares a number of allocation methods with respect to this criterion. In this article however, we concentrate only on *estimation* bias arising from model misspecification.

The methods of Section 3, and those employed by the researchers mentioned above, yield designs which are sequential but not adaptive, viz. each allocation depends on the previous allocations and prognostic factors, but not on the responses. There is of course a very practical motivation for this: in many clinical trials, assignments must be made before the responses to previous assignments have been obtained. For a further discussion of some ethical problems which may also arise in connection with adaptive designs in clinical trials, see Bather (1995).

A consequence of this approach is that the usual fixed-sample size tests of hypotheses, carried out conditionally on the assignments and covariates, remain valid. The power of such tests typically increases with the precision in the estimation of the parameters being tested; thus optimal designs may be expected to lead to optimal tests.

In Section 4 we assess our allocations numerically and compare them with those of Atkinson (1982), modified for possibly heteroscedastic errors. We find that the robust allocation schemes can result in significant decreases in the mse when the fitted

models are biased, at a minimal cost in efficiency when in fact the fitted models are correct.

2. A generalized design problem

We consider regression models in which a subject with covariates x_1, \dots, x_t yields a response Y satisfying

$$Y = E[Y|\mathbf{u}, \mathbf{x}] + \sigma_{\mathbf{u}}\varepsilon,$$

where $\mathbf{u}_{p \times 1}$ is a vector of non-random regressors taking on, only finitely, many values and $\mathbf{x}_{t \times 1}$ is the vector of covariates ranging over a space \mathcal{S} . Our results are expressed for known scales $\sigma_{\mathbf{u}}$, with the understanding that estimates will be substituted at the implementation stage. We assume that \mathbf{x} has a density $m(\mathbf{x})$, with respect to a measure μ on \mathcal{S} . A common special case is $\mathbf{x} = (\mathbf{x}_0^T, \mathbf{x}_1^T)^T$, where \mathbf{x}_0 is a vector of discrete covariates ranging over a space \mathcal{S}_0 , and \mathbf{x}_1 is a vector of continuous covariates ranging over a space \mathcal{S}_1 . Denote counting measure on \mathcal{S}_0 by μ_0 , and Lebesgue measure on \mathcal{S}_1 by μ_1 . Then in this case $m(\mathbf{x}) = m(\mathbf{x}_0, \mathbf{x}_1)$ is the density on $\mathcal{S} = S_0 \times \mathcal{S}_1$, with respect to $\mu = \mu_0 \times \mu_1$.

The regression response is a function of \mathbf{u} and of random regressors $\mathbf{z}(\mathbf{x})_{q \times 1}$, although these regressors are not necessarily modelled by the experimenter. We consider two cases. In each,

$$E[Y|\mathbf{u}, \mathbf{x}] \approx \mathbf{v}_{\mathbf{u}}^T(\mathbf{x})\psi. \quad (1)$$

In Model 1, the fitted response is $\hat{Y} = \mathbf{u}^T \hat{\theta}$ and we take $\mathbf{v}_{\mathbf{u}}(\mathbf{x}) \equiv \mathbf{u}$, $\psi = \theta_{p \times 1}$. In Model 2 the fitted response is $\hat{Y} = \mathbf{u}^T \hat{\theta} + \mathbf{z}^T(\mathbf{x})\hat{\phi}$ and we take $\mathbf{v}_{\mathbf{u}}(\mathbf{x}) = (\mathbf{u}^T, \mathbf{z}^T(\mathbf{x}))^T$, $\psi = (\theta^T, \phi^T)^T$ for a vector $\phi_{q \times 1}$.

Define P to be p for Model 1 and $p + q$ for Model 2. The ‘true’ parameter $\psi_{P \times 1}$ is defined by

$$\psi = \arg \min_{\beta} \sum_{\mathbf{u}} \int_{\mathcal{S}} \{E[Y|\mathbf{u}, \mathbf{x}] - \mathbf{v}_{\mathbf{u}}^T(\mathbf{x})\beta\}^2 \mu(d\mathbf{x}). \quad (2)$$

We assume that the approximation (1) is sufficiently accurate that the integrals in (2) exist for all β in some open set. Now define

$$f_{\mathbf{u}}(\mathbf{x}) = \sqrt{n}\{E[Y|\mathbf{u}, \mathbf{x}] - \mathbf{v}_{\mathbf{u}}^T(\mathbf{x})\psi\},$$

so that

$$Y = \mathbf{v}_{\mathbf{u}}^T(\mathbf{x})\psi + n^{-1/2} f_{\mathbf{u}}(\mathbf{x}) + \sigma_{\mathbf{u}}\varepsilon \quad (3)$$

with, by virtue of (2),

$$\sum_{\mathbf{u}} \int_{\mathcal{S}} f_{\mathbf{u}}(\mathbf{x}) \mathbf{v}_{\mathbf{u}}(\mathbf{x}) \mu(d\mathbf{x}) = \mathbf{0}. \quad (4)$$

In order that errors due to model misspecification not swamp those due to random variation, we shall assume that

$$\int_{\mathcal{S}} f_{\mathbf{u}}^2(\mathbf{x}) \mu(d\mathbf{x}) \leq \eta_{\mathbf{u}}^2 \quad (5)$$

for given, finite bounds $\eta_{\mathbf{u}}^2$.

Example 1.1. Suppose that there are $p=2$ treatments, a discrete covariate $x_0 = \text{Gender}$ coded as -1 (Male) or 1 (Female) and a continuous covariate $x_1 = \text{Weight}$, scaled to $[-1, 1]$. Then $\mathcal{S} = \{-1, 1\} \times [-1, 1]$. Let $\mathbf{u} = (u_1, u_2)^T$ be the vector of treatment indicators taking values $\mathbf{u}_1 = (1, 0)^T$, $\mathbf{u}_2 = (0, 1)^T$. Denote by $m(x_0, x_1)$ the joint probability that $X_0 = x_0$ and density of X_1 . For Model 1, in which the covariates are observed but not fitted, the orthogonality requirements (4) become

$$\sum_{i=1,2} \int_{-1}^1 \{f_{\mathbf{u}_i}(-1, x_1) + f_{\mathbf{u}_i}(1, x_1)\} dx_1 = 0. \quad (6)$$

For Model 2 take $\mathbf{z}(\mathbf{x}) = \mathbf{x}$, so that $\mathbf{v}_{\mathbf{u}}(\mathbf{x}) = (\mathbf{u}^T, x_0, x_1)^T$. Then the orthogonality requirements become

$$\begin{aligned} 0 &= \sum_{i=1,2} \int_{-1}^1 f_{\mathbf{u}_i}(-1, x_1) dx_1 = \sum_{i=1,2} \int_{-1}^1 f_{\mathbf{u}_i}(1, x_1) dx_1 \\ &= \sum_{i=1,2} \int_{-1}^1 x_1 \{f_{\mathbf{u}_i}(-1, x_1) + f_{\mathbf{u}_i}(1, x_1)\} dx_1. \end{aligned}$$

For Model 1 only the treatment effects are estimated. The conditional bias, as derived at (A.2), is

$$E[\hat{\theta} - \theta | \mathbf{x}] = n^{-1/2} \begin{pmatrix} \text{aver } f_{\mathbf{u}_1}(\mathbf{x}^{T_1}) \\ \text{aver } f_{\mathbf{u}_2}(\mathbf{x}^{T_2}) \end{pmatrix}, \quad (7)$$

where $\text{aver } f_{\mathbf{u}_i}(\mathbf{x}^{T_i})$ denotes the average value of $f_{\mathbf{u}_i}(\mathbf{x})$, over all sampled covariates \mathbf{x} accompanying individuals receiving treatment i .

For Model 2 the mean responses are given by

Treatment	Gender	
	M	F
1	$\theta_1 - \phi_0 + \phi_1 x_1 + n^{-1/2} f_{\mathbf{u}_1}(-1, x_1)$	$\theta_1 + \phi_0 + \phi_1 x_1 + n^{-1/2} f_{\mathbf{u}_1}(1, x_1)$
2	$\theta_2 - \phi_0 + \phi_1 x_1 + n^{-1/2} f_{\mathbf{u}_2}(-1, x_1)$	$\theta_2 + \phi_0 + \phi_1 x_1 + n^{-1/2} f_{\mathbf{u}_2}(1, x_1)$

and an expression for the bias, similar to but more involved than (7) can be derived.

Example 1.2. Suppose that all covariates are continuous and that, after an affine transformation, $\mathbf{x} = \mathbf{x}_1$ follows a q -dimensional unit normal distribution restricted to a sphere

$\mathcal{S} = \mathcal{S}_1$ which is of finite measure, but large enough to contain all of the data with high probability. In practice, one could first transform to standardized covariates, viz. $\mathbf{x} \rightarrow \mathbf{S}^{-1/2}(\mathbf{x} - \mathbf{t})$, where \mathbf{S} is a scatter matrix and \mathbf{t} a location estimate, and then choose a radius at least as large as the maximum value of $\|\mathbf{S}^{-1/2}(\mathbf{x} - \mathbf{t})\|$.

Formally, if $\phi(\mathbf{x})$ denotes the q -dimensional standard normal density and $H(\cdot)$ the χ_q^2 distribution function, then for some c slightly less than 1 one may take $\mathcal{S} = \{\mathbf{x} \mid \|\mathbf{x}\|^2 \leq H^{-1}(c)\}$ and $m(\mathbf{x}) = \phi(\mathbf{x})/c$, $\mathbf{x} \in \mathcal{S}$.

Suppose now that \mathbf{u} takes on only p possible values $\mathbf{u}_i = (0, \dots, 0, \overset{i}{1}, 0, \dots, 0)^T$ ($i = 1, \dots, p$) and that a subject with covariates \mathbf{x} is assigned to a group defined by $\mathbf{u} = \mathbf{u}_i$ with probability

$$\rho_i(\mathbf{x}) = P(i|\mathbf{x}),$$

independent of all other assignments. Then the joint density of the covariates and probability that a covariate results in an assignment to group i is given by

$$m_i(\mathbf{x})\mu(d\mathbf{x}) := \rho_i(\mathbf{x})m(\mathbf{x})\mu(d\mathbf{x}).$$

Also put $f_i(\mathbf{x}) = f_{\mathbf{u}_i}(\mathbf{x})$, $\mathbf{v}_i(\mathbf{x}) = \mathbf{v}_{\mathbf{u}_i}(\mathbf{x})$, $\eta_i = \eta_{\mathbf{u}_i}$ and $\sigma_i = \sigma_{\mathbf{u}_i}$.

The experimenter is assumed to be interested in estimating a complete set of $p - 1$ orthonormal contrasts $\mathbf{W}_0\theta$. Define $\mathbf{W}_{p-1 \times p} = \mathbf{W}_0$ for Model 1, $= (\mathbf{W}_0 \ \mathbf{0}_{p-1 \times q})$ for Model 2. Then $\mathbf{W}_0\theta = \mathbf{W}\psi$ in both models. We shall derive the mean squared error matrix $\text{MSE}(\mathbf{W}_0\hat{\theta})$ of $\mathbf{W}_0\hat{\theta}$, conditional on the covariates and the treatment assignments. The loss function is taken to be the almost sure limit of the normalized determinant of this matrix:

$$\mathcal{L}(\rho_1, \dots, \rho_p; f_1, \dots, f_p) = \lim_{n \rightarrow \infty} |n\text{MSE}(\mathbf{W}_0\hat{\theta})|.$$

Theorem 1. Assume that

$$\int_{\mathcal{S}} \|\mathbf{v}_i(\mathbf{x})\|^2 m(\mathbf{x})\mu(d\mathbf{x}) < \infty \quad \text{for } i = 1, \dots, p.$$

Of the n subjects entering the study, let n_i be the number assigned to group i . Assume that all $n_i \rightarrow \infty$ in such a way that n_i/n approaches a limit. Define $P \times 1$ vectors and $P \times P$ matrices

$$\mathbf{b}_i(f_i) = \int_{\mathcal{S}} \mathbf{v}_i(\mathbf{x})f_i(\mathbf{x})m_i(\mathbf{x})\mu(d\mathbf{x}),$$

$$\mathbf{A}_i = \int_{\mathcal{S}} \mathbf{v}_i(\mathbf{x})\mathbf{v}_i^T(\mathbf{x})\mu(d\mathbf{x}),$$

$$\mathbf{B}_i = \int_{\mathcal{S}} \mathbf{v}_i(\mathbf{x})\mathbf{v}_i^T(\mathbf{x})m_i(\mathbf{x})\mu(d\mathbf{x}),$$

and

$$\mathbf{A} = \sum_{i=1}^p \mathbf{A}_i, \quad \mathbf{B} = \sum_{i=1}^p \mathbf{B}_i, \quad \mathbf{Q} = \sum_{i=1}^p \sigma_i^2 \mathbf{B}_i, \quad \mathbf{b}(f_1, \dots, f_p) = \sum_{i=1}^p \mathbf{b}_i(f_i). \quad (8)$$

Assume that \mathbf{A} , \mathbf{B} and \mathbf{Q} are positive definite. Then the limiting determinant of the normalized mean squared error matrix is

$$\mathcal{L}(\rho_1, \dots, \rho_p; f_1, \dots, f_p) =$$

$$|\mathbf{WB}^{-1}\mathbf{QB}^{-1}\mathbf{W}^T| \cdot \left(1 + \|(\mathbf{WB}^{-1}\mathbf{QB}^{-1}\mathbf{W}^T)^{-1/2}\mathbf{WB}^{-1}\mathbf{b}(f_1, \dots, f_p)\|^2\right). \quad (9)$$

In the proof of Theorem 2, and elsewhere, we will make use of the following result.

Lemma 1. If the rows of $\mathbf{W}_{p-1 \times p}$ are mutually orthonormal contrasts, so that $\mathbf{T} := ((1/\sqrt{p})\mathbf{1} \ \mathbf{W}^T)^T$ is an orthogonal matrix, and if $\mathbf{U}_{p \times p}$ is a non-singular matrix, then

- (i) $(\mathbf{WU}^{-1}\mathbf{W}^T)^{-1} = \mathbf{WUW}^T - \frac{\mathbf{WU}\mathbf{1}\mathbf{1}^T\mathbf{UW}^T}{\mathbf{1}^T\mathbf{U}\mathbf{1}}$,
- (ii) $\mathbf{W}^T(\mathbf{WU}^{-1}\mathbf{W}^T)^{-1}\mathbf{W} = \mathbf{U} - \frac{\mathbf{U}\mathbf{1}\mathbf{1}^T\mathbf{U}}{\mathbf{1}^T\mathbf{U}\mathbf{1}}$,
- (iii) $|\mathbf{WU}^{-1}\mathbf{W}^T| = \frac{\mathbf{1}^T\mathbf{U}\mathbf{1}}{p|\mathbf{U}|}$.

Define $\rho = (\rho_1, \dots, \rho_p)^T$. The first factor in (9), i.e.

$$\mathcal{L}_0(\rho) := |\mathbf{WB}^{-1}\mathbf{QB}^{-1}\mathbf{W}^T|$$

arises from variance alone. It turns out that this term is minimized by constant assignment probabilities

$$r_i = \int_{\mathcal{S}} \rho_i(\mathbf{x}) m(\mathbf{x}) \mu(d\mathbf{x}). \quad (10)$$

Theorem 2. Constant assignment probabilities $\rho(\mathbf{x}) \equiv \mathbf{r} := (r_1, \dots, r_p)^T$ minimize $\mathcal{L}_0(\rho)$. Define matrices $\mathbf{D}_r = \text{diag}(r_1, \dots, r_p)$, $\mathbf{D}_{\sigma^2} = \text{diag}(\sigma_1^2, \dots, \sigma_p^2)$, $\mathbf{D} = \mathbf{D}_{\sigma^2}^{-1}\mathbf{D}_r$. If the $\rho_i(\mathbf{x})$ are constant then

$$\mathcal{L}(\rho_1, \dots, \rho_p) = \frac{\text{tr } \mathbf{D}}{p|\mathbf{D}|}, \quad (11)$$

and so optimal probabilities r_1, \dots, r_p must minimize $\text{tr } \mathbf{D}/|\mathbf{D}|$ subject to $0 \leq r_i \leq 1$, $\sum_{i=1}^p r_i = 1$.

The minimizing r_i are obtained from the following theorem.

Theorem 3. Suppose that $\rho_i(\mathbf{x}) \equiv r_i$. Assume that the groups have been re-labelled, if necessary, so that $\sigma_1^2 = \max_{1 \leq i \leq p} \sigma_i^2$ and hence $\tau_i := \sigma_1^2/\sigma_i^2 \geq 1$ for $i = 2, \dots, p$. Define a function

$$h(\alpha) = \alpha - 1 - \sum_{i=2}^p (\tau_i - 1)(p - 1 + \tau_i \alpha^{-1})^{-1}, \quad \alpha \geq 0.$$

Table 1

Optimal probabilities r_2^* (upper value) and r_3^* (lower value) when $p = 3$

	1	2	3	4
1	0.333			
	0.333			
2	0.360	0.309		
	0.281	0.309		
3	0.375	0.327	0.297	
	0.250	0.278	0.297	
4	0.385	0.339	0.310	0.289
	0.229	0.257	0.275	0.289

Then:

- (i) There is a unique zero α_* of h , and $\alpha_* \geq 1$.
- (ii) The probabilities minimizing (11) are $r_i^* = (p - 1 + \alpha_*^{-1} \tau_i)^{-1}$ for $i = 2, \dots, p$ and $r_1^* = 1 - \sum_{i=2}^p r_i^*$.

Corollary 1. When $p = 2$, optimal constant assignment probabilities are $\rho_i(\mathbf{x}) \equiv \sigma_i / (\sigma_1 + \sigma_2)$ for $i = 1, 2$. When $p = 3$ they are obtained from

$$\alpha_* = \sqrt{\frac{\tau_2 \tau_3 + \tau_2 + \tau_3}{3}} \cos\left(\frac{1}{3} \arctan \sqrt{\frac{(\tau_2 \tau_3 + \tau_2 + \tau_3)^3}{27 \tau_2^2 \tau_3^2} - 1}\right).$$

When all σ_i^2 are equal, optimal constant assignment probabilities are $\rho_i(\mathbf{x}) \equiv p^{-1}$ for $i = 1, \dots, p$.

For $p = 3$ some representative values of the r_i^* are given in Table 1.

Remark. Under a different set of side conditions determining the nature of contamination, the assignment probabilities of Theorem 3 are *minimax*, in that they minimize $\max_{f_1, \dots, f_p} \mathcal{L}(\rho_1, \dots, \rho_p; f_1, \dots, f_p)$. To see this, suppose that (4) is replaced by the requirement that $\int_{\mathcal{S}'} f_i(\mathbf{x}) \mathbf{v}_i(\mathbf{x}) m(\mathbf{x}) \mu(d\mathbf{x}) = \mathbf{0}$ for each i . The intuitive content of this requirement is that $f_i(\mathbf{x})$ has mean zero and (in Model 2) is uncorrelated with the regressors. Then in Theorem 1 we have that $\mathbf{b}_i(f_i) \equiv \mathbf{0}$ if the $\rho_i(\mathbf{x})$ are constant; thus the bias vanishes and constants minimizing the variance component of mse are a fortiori minimax.

For $p = 2$ and $q = 1$, it is possible to compare our results with those of Heckman (1987), who obtained minimax allocation schemes in the case of a single continuous covariate $x \in [-1, 1]$, as briefly described in Section 1. In our notation the neighbourhood structure used was as at (3) with $f_i(x)$ replaced by $f_i(x', x)$

satisfying $|n^{-1/2} f_i(x', x)| \leq \eta_i(x')|x - x'|$ for specified $x' \in [-1, 1]$ and functions $\eta_i(\cdot)$. The allocation probability $\rho_H(x) = P(\text{treatment}|x)$ was obtained as

$$\rho_H(x) = \frac{1}{1 + [\sigma_2^2 \kappa(x)/\sigma_1^2]^{1/3}}$$

with $\kappa(x) = \eta_2(x)/\eta_1(x)$ assumed known. These assignment probabilities are constant if $\eta_2(x) \propto \eta_1(x)$. In particular $\rho_H(x) \equiv r_1^*$ if $\eta_2(x)/\sigma_2 = \eta_1(x)/\sigma_1$.

3. Sequential assignments

Efron (1971) introduced the biased coin design to clinical trials, the intention being to achieve balance by giving preference in a sequential assignment to an under-represented treatment. Pocock and Simon (1975) gave an extension of the method to account for prognostic factors. Atkinson (1982) proposed a modification whose purpose was to increase the efficiency of the design while hastening the convergence of the sequence of assignments to a balanced state. All assumed the fitted response function to be correct and the errors to be homoscedastic.

We propose an extension of Atkinson's design to allow for departures from the fitted response, and from homoscedasticity, as outlined in Section 2. We give the details only for discrete covariates; in the continuous case some grouping will be necessary. Suppose then that there are L levels of the covariates, denoted $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(L)}$. Suppose that n subjects have been assigned to treatment groups, and that the $(n+1)$ th subject, with covariates $\mathbf{x}_* \in \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(L)}\}$, is to be assigned. Our method assigns this subject to group k with probability

$$P(k|\mathbf{x}_*) = \frac{r_k^* d_k^* b_k^*}{\sum_{i=1}^p r_i^* d_i^* b_i^*}, \quad k = 1, \dots, p, \quad (12)$$

where r_k^* is as in Theorem 3, $d_k^* = d_k(\mathbf{x}_*)$ is a measure of the sequential reduction in the determinant of the covariance matrix of $\mathbf{W}_0 \hat{\theta}$ to be realized by an assignment to group k , and $b_k^* = b_k(\mathbf{x}_*)$ is a measure of the inverse bias to be effected by such an assignment. Atkinson (1982) assumes that the model is correctly specified and that the errors are homoscedastic; he then takes $r_k^* = b_k^* \equiv 1$.

Each of r_k^* , d_k^* and b_k^* must be estimated. The first of these is estimated by substituting estimates $\hat{\sigma}_i$ into the expressions of Theorem 3. To express and estimate d_k^* , let the n_i subjects assigned to group i have covariates \mathbf{x}_{ij} , $j = 1, \dots, n_i$, $i = 1, \dots, p$. Let \mathbf{V}_n be the $n \times P$ model matrix determined by the first n assignments, with rows $\mathbf{v}_{ij}^\top := (\mathbf{u}_i^\top, \mathbf{z}^\top(\mathbf{x}_{ij}))$ ordered lexicographically. For Model 1, of course $\mathbf{v}_{ij} = \mathbf{u}_i$. For Model 2 denote by $\mathbf{Z}_{n \times q}$ the last q columns of \mathbf{V}_n . Let \mathbf{f}_n be the $n \times 1$ vector with elements $f_i(\mathbf{x}_{ij})$, ordered compatibly and estimated as in (14) below. Define $\mathbf{B}_n = \mathbf{V}_n^\top \mathbf{V}_n$, $\mathbf{Q}_n = \mathbf{V}_n^\top \Sigma_n \mathbf{V}_n$, where $\Sigma_n = \bigoplus_{i=1}^p \sigma_i^2 \mathbf{I}_{n_i}$. If the $(n+1)$ th assignment is to group k these become

$$\begin{aligned} \mathbf{V}_{(k)} &= \begin{pmatrix} \mathbf{V}_n \\ \mathbf{v}_k^{*\top} \end{pmatrix}, & \mathbf{f}_{(k)} &= \begin{pmatrix} \mathbf{f}_n \\ f_k^* \end{pmatrix}, \\ \mathbf{B}_{(k)} &= \mathbf{B}_n + \mathbf{v}_k^* \mathbf{v}_k^{*\top}, & \mathbf{Q}_{(k)} &= \mathbf{Q}_n + \sigma_k^2 \mathbf{v}_k^* \mathbf{v}_k^{*\top}, \end{aligned}$$

where $\mathbf{v}_k^* := \mathbf{v}_k(\mathbf{x}_*)$ and $f_k^* := f_k(\mathbf{x}_*)$. Now partition $\mathbf{B}_n^{-1} \mathbf{Q}_n \mathbf{B}_n^{-1}$ as

$$\mathbf{B}_n^{-1} \mathbf{Q}_n \mathbf{B}_n^{-1} = \begin{pmatrix} \mathbf{U}_n^{-1} & * \\ * & * \end{pmatrix} \quad (13)$$

with \mathbf{U}_n being $p \times p$ and (in Model 2) with * indicating terms whose specific values are not important. Define $\mathbf{U}_{(k)}$ via a corresponding partitioning of $\mathbf{B}_{(k)}^{-1} \mathbf{Q}_{(k)} \mathbf{B}_{(k)}^{-1}$. From (A.1) and (A.2), based on n observations $\mathbf{W}_0 \hat{\theta}$ has mse matrix

$$\text{MSE}(\mathbf{W}_0 \hat{\theta}) = \mathbf{W} \mathbf{B}_n^{-1} \mathbf{Q}_n \mathbf{B}_n^{-1} \mathbf{W}^T + \frac{1}{n} \mathbf{W} \mathbf{B}_n^{-1} \mathbf{V}_n^T \mathbf{f}_n \mathbf{f}_n^T \mathbf{V}_n \mathbf{B}_n^{-1} \mathbf{W}^T.$$

Let $\text{COV}(\mathbf{W}_0 \hat{\theta})$ and $\text{COV}_{(k)}(\mathbf{W}_0 \hat{\theta})$ denote the covariance matrix before and after the assignment to group k . We take

$$d_k^* = \left(\frac{|\text{COV}(\mathbf{W}_0 \hat{\theta})|}{|\text{COV}_{(k)}(\mathbf{W}_0 \hat{\theta})|} - 1 \right)^+,$$

(the positive part); by Lemma 1(ii) and (iii) this becomes

$$d_k^* = \left(\frac{|\mathbf{U}_{(k)}|}{|\mathbf{U}_n|} \cdot \frac{\mathbf{1}^T \mathbf{U}_n \mathbf{1}}{\mathbf{1}^T \mathbf{U}_{(k)} \mathbf{1}} - 1 \right)^+.$$

We estimate d_k^* by replacing the variances in $\mathbf{Q}_{(k)}$ by estimates.

The bias component of the MSE matrix is proportional to $\mathbf{W} \mathbf{B}_n^{-1} \mathbf{V}_n^T \mathbf{f}_n \mathbf{f}_n^T \mathbf{V}_n \mathbf{B}_n^{-1} \mathbf{W}^T$. Since this has rank 1, the Euclidean norm coincides with the trace and with the maximum eigenvalue; their common value has

$$\begin{aligned} \|\mathbf{W} \mathbf{B}_n^{-1} \mathbf{V}_n^T \mathbf{f}_n \mathbf{f}_n^T \mathbf{V}_n \mathbf{B}_n^{-1} \mathbf{W}^T\| &= \mathbf{f}_n^T \mathbf{V}_n \mathbf{B}_n^{-1} \mathbf{W}^T \cdot \mathbf{W} \mathbf{B}_n^{-1} \mathbf{V}_n^T \mathbf{f}_n \\ &= \mathbf{f}_n^T \mathbf{V}_n \mathbf{B}_n^{-1} \begin{pmatrix} \mathbf{I}_p - \frac{1}{p} \mathbf{1} \mathbf{1}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{B}_n^{-1} \mathbf{V}_n^T \mathbf{f}_n. \end{aligned}$$

Some approximations yield the ad hoc estimate

$$\hat{f}_i(\mathbf{x}^{(l)}) = \text{sign}(\tilde{e}_{i,l}) \sqrt{\tilde{e}_{i,l}^2 + \frac{\hat{\sigma}_i^2}{n_{i,l}}}, \quad (14)$$

where $n_{i,l}$ is the number of times, in the first n assignments, that a subject with covariates $\mathbf{x}^{(l)}$ has been assigned to group i and $\tilde{e}_{i,l}$ is the median of the residuals at these data points. We then take

$$b_k^* = \left\{ \hat{\mathbf{f}}_{(k)}^T \mathbf{V}_{(k)} \mathbf{B}_{(k)}^{-1} \begin{pmatrix} \mathbf{I}_p - \frac{1}{p} \mathbf{1} \mathbf{1}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{B}_{(k)}^{-1} \mathbf{V}_{(k)}^T \hat{\mathbf{f}}_{(k)} \right\}^{-2}.$$

In the rare case that the term in braces is 0 we set this $b_k^* = 1$ and all others = 0.

Instead of our d_k^* Atkinson (1982) uses the quantity $d_k^A := n\mathbf{v}_k^{*\top} \mathbf{B}_n^{-1} \mathbf{W}^T [\mathbf{W} \mathbf{B}_n^{-1} \mathbf{W}^T]^{-1} \mathbf{W} \mathbf{B}_n^{-1} \mathbf{v}_k^*$. In the case of homoscedastic errors some algebra reveals that for Model 2,

$$d_k^A = n \left\{ 1 + \mathbf{z}^T(\mathbf{x}_*) (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{z}(\mathbf{x}_*) + \frac{(1 - \mathbf{1}^T \mathbf{Z} (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{z}(\mathbf{x}_*))^2}{n - \mathbf{1}^T \mathbf{Z} (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{1}} \right\} d_k^*.$$

For Model 1 this reduces to $d_k^A = (n+1)d_k^*$. In either case d_k^A/d_k^* does not depend on k and so both forms yield the same values of (12).

For Model 1 there is considerable simplification in d_k^* and b_k^* even for heteroscedastic errors. If the first n subjects have n_i assignments to treatment i then we find that

$$d_k^* = \frac{\sum_{i \neq k} n_i / \sigma_i^2}{n_k \left(\sum_{i \neq k} n_i / \sigma_i^2 + (n_k + 1) / \sigma_k^2 \right)}.$$

For homoscedastic errors this becomes $d_k^* = ((n/n_k) - 1)/(n+1)$. Now let $n_{i,l}^{(k)}$ and $n_i^{(k)}$ denote the frequencies after the $(n+1)$ th subject is assigned to group k , and let $\zeta_i^{(k)}$ be the weighted f_i -estimate $\zeta_i^{(k)} = \sum_{l=1}^L n_{i,l}^{(k)} \hat{f}_i(\mathbf{x}^{(l)}) / n_i^{(k)}$, with an average $\bar{\zeta}^{(k)} = \sum_{i=1}^p \zeta_i^{(k)} / p$. Then

$$b_k^* = \left\{ \sum_{i=1}^p (\zeta_i^{(k)} - \bar{\zeta}^{(k)})^2 \right\}^{-2}.$$

A reduced MSE may be attained at a cost of increased imbalance, and so we propose a measure of the imbalance after the n th assignment. This measure is defined through the estimated variances of the $\rho_i(\mathbf{x})$. First define $n_{\cdot l} = \sum_{i=1}^p n_{i,l}$, and $n_{i\cdot} = n_i$ analogously. We adopt an empirical probability model under which $\mathbf{x}^{(l)}$ occurs with probability $P(\mathbf{x}^{(l)}) = n_{\cdot l}/n$. We estimate $\rho_i(\mathbf{x}^{(l)})$ by $\hat{\rho}_i(\mathbf{x}^{(l)}) = n_{i,l}/n_{\cdot l}$, with $0/0 := 0$. Then the expected value r_i of ρ_i , as at (10), is estimated by

$$\hat{r}_i = \sum_{l=1}^L \hat{\rho}_i(\mathbf{x}^{(l)}) P(\mathbf{x}^{(l)}) = \frac{n_i}{n},$$

and $\text{VAR}_{\mathbf{x}}[\rho_i(\mathbf{x})]$ by

$$S_i^2 = \sum_{l=1}^L (\hat{\rho}_i(\mathbf{x}^{(l)}) - \hat{r}_i)^2 P(\mathbf{x}^{(l)}) = \frac{1}{n} \sum_{l=1}^L n_{\cdot l} \left(\frac{n_{i,l}}{n_{\cdot l}} - \frac{n_{i\cdot}}{n} \right)^2.$$

Note that S_i^2 is a natural measure of the imbalance in the allocation of the covariates to the i th treatment. Our overall measure of imbalance is

$$S_{\rho}^2 = \sum_{i=1}^p S_i^2.$$

4. Simulations

We consider two examples.

Example 4.1. If there are $p = 2$ treatments and a single covariate X taking values $\{-1, 0, 1\}$ (on each of which points μ places mass $1/3$) then the constraints on f_1 and f_2 are

$$\begin{aligned} \sum_{i=1,2} \{f_i(-1) + f_i(0) + f_i(1)\} &= 0, \\ f_i^2(-1) + f_i^2(0) + f_i^2(1) &\leq \eta_i^2 \end{aligned} \quad (15)$$

in Model 1, and in addition $\sum_{i=1,2} \{f_i(-1) - f_i(1)\} = 0$ in Model 2. All three constraints, with equality in (15), are satisfied by

$$f_i(x) = \frac{(-1)^i \eta_i}{\sqrt{2}} (2 - 3x^2), \quad i = 1, 2. \quad (16)$$

In Model 1 the indicators of the treatments are the only regressors; in Model 2 the assumed response is

$$E[Y|x] = \sum_{i=1,2} \theta_i I(\text{Treatment } i) + \phi x.$$

In each case design robustness is sought against possible contamination of these responses by $f_i(x)$, and against heteroscedasticity.

In our simulations we have used $\theta_1 = \theta_2 = \phi = 1$ and normally distributed errors, with $\sigma_1^2 = 1$, $\sigma_2^2 = 1/4$. Thus $r_1^* = 2/3$, $r_2^* = 1/3$. The standard deviations σ_1 and σ_2 are estimated by the median absolute deviations of the corresponding residuals, normalized in the usual way for consistency at the normal distribution. We investigate the cases $\eta_1 = \eta_2 = 0$ and $\eta_1 = \eta_2 = 3$.

Of course the experimenter is unaware of the true value of the η_i . If he believes strongly that his response function is exactly correct, i.e. that all $\eta_i = 0$, but wishes to guard against heteroscedasticity, then a modification of the technique of Atkinson (1982) would seem to be appropriate. The modification consists of using non-constant values of r_k^* , as in Theorem 3, and in (12) to account for possible heteroscedasticity. Then use constant values $b_k^* \equiv 1$ so that bias is discounted.

The results using both methods—our robust method, and this modification of Atkinson's method—are illustrated in Fig. 1 (a)–(d), where we plot the estimated root-mse

$$\sqrt{\frac{2}{J} \sum_{j=1}^J \left| \mathbf{W}_0(\hat{\theta}_j - \theta) \right|^2} = \sqrt{\frac{1}{J} \sum_{j=1}^J (\hat{\delta}_j - \delta)^2}$$

($\delta = \theta_1 - \theta_2$), from $J = 200$ simulated runs, after each new subject is assigned. As an indication of simulation variability we have also calculated the standard errors of the differences of these root-mse values. Their averages over the 30 new subjects appear in parentheses in the plots of Fig 1. The corresponding imbalance measures are plotted in Fig. 2 (a)–(d).

Example 4.2. In this example, we consider the case of 2 treatments, with two covariates X_1, X_2 taking labels ± 1 . In Model 1, the assumed response is the same as that in

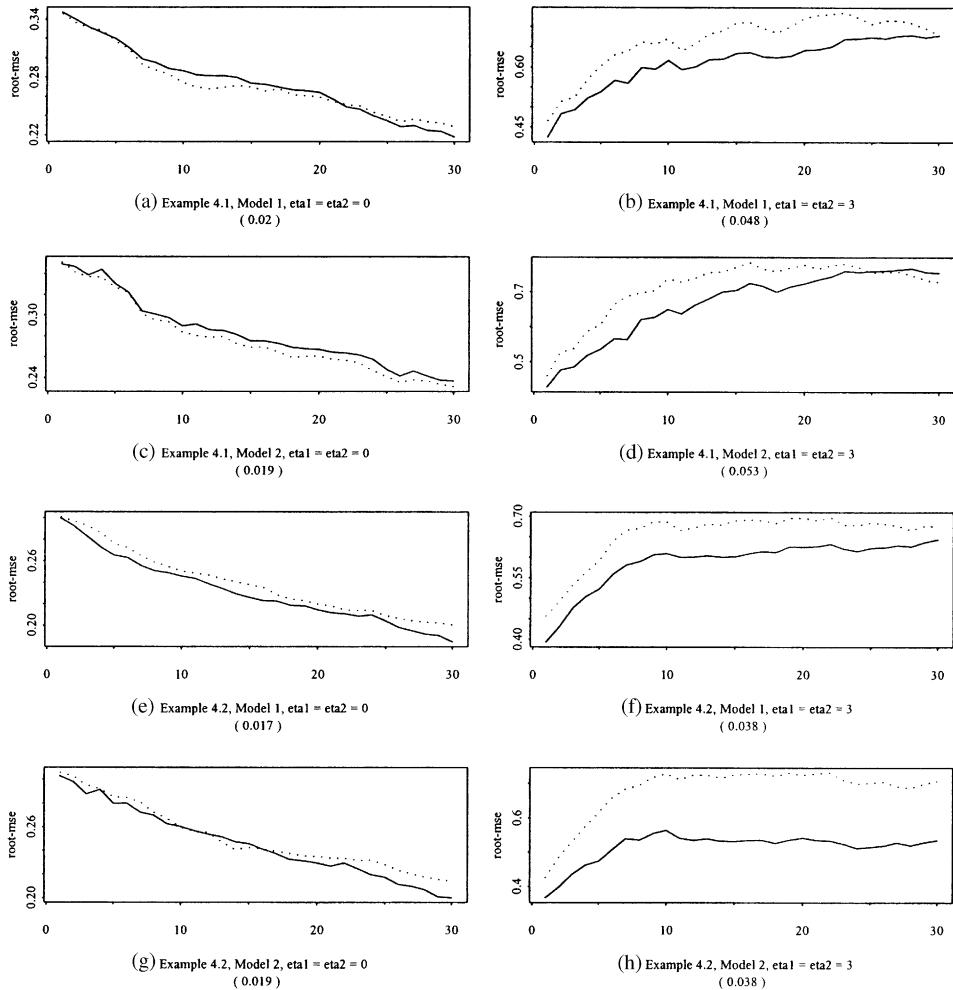


Fig. 1. Root-mse versus new subjects for the examples of Section 4. Dotted line is for Atkinson's method ($b_k^* \equiv 1$) modified for heteroscedasticity; solid line is for the robust method. Average (over the 30 subjects) standard error of the difference in parentheses.

Example 4.1; in Model 2 it is

$$E[Y|x] = \sum_{i=1,2} \theta_i I(\text{Treatment } i) + \phi_1 I(x_1 = 1) + \phi_2 I(x_2 = 1).$$

The constraints imposed by Model 2, hence by Model 1 as well, are satisfied by $f_i(x_1, x_2) = (-1)^i \eta_i x_1 x_2 / \sqrt{2}$. See Figs. 1 and 2(e)–(h).

Both examples lead to similar conclusions. Allocation schemes which model possible biases arising from model uncertainties can significantly decrease the mean squared error of the estimates when such biases actually exist. When there are in fact no such

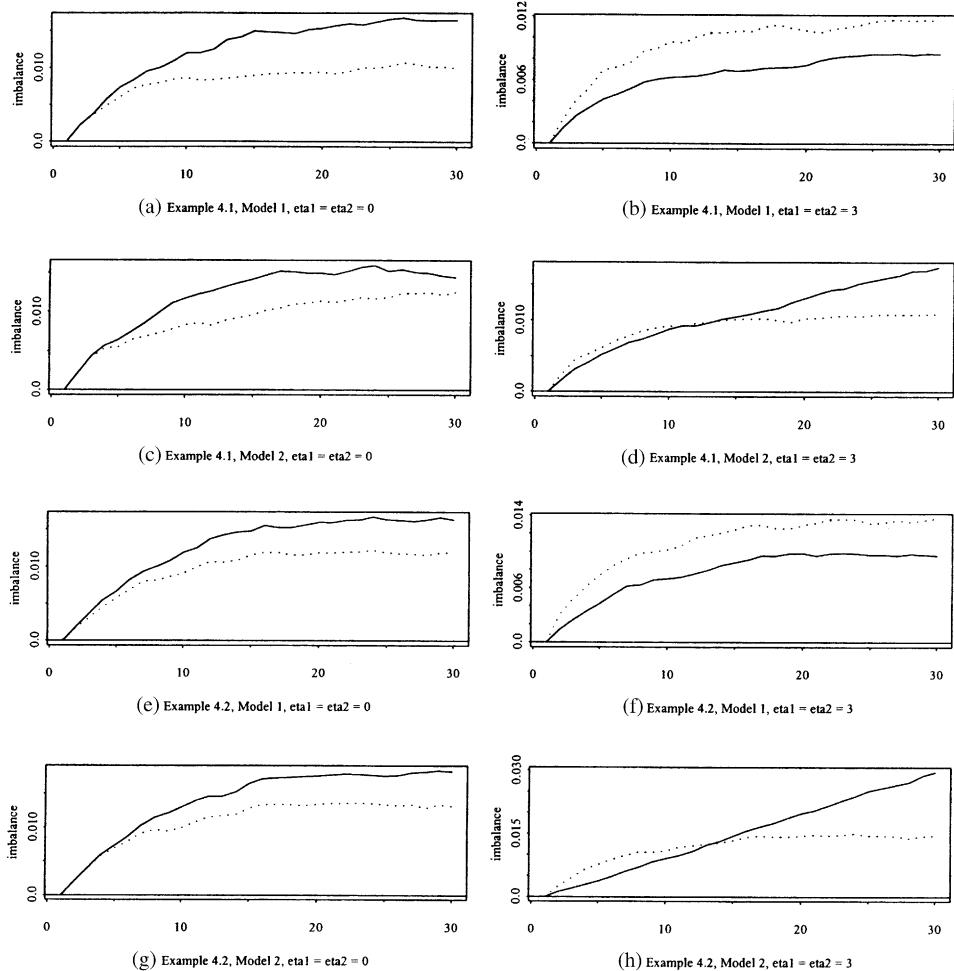


Fig. 2. Imbalance versus new subjects for the examples of Section 4. Dotted line is for Atkinson's method ($b_k^* \equiv 1$) modified for heteroscedasticity; solid line is for the robust method.

biases these schemes perform as well as the competing method of Atkinson (1982). The gains are often, but not always, achieved at a cost of some increased imbalance across prognostic factors.

Acknowledgements

This research is supported by the Natural Sciences and Engineering Research Council of Canada. The comments of the referees are also gratefully acknowledged.

Appendix. Derivations

Proof of Theorem 1. We give the details of the derivation of MSE for Model 2, and then specialize the results to Model 1. Denote by \mathbf{y} the $n \times 1$ data vector, with elements y_{ij} ordered lexicographically. Then

$$\mathbf{y} = \mathbf{V}_n \psi + n^{-1/2} \mathbf{f}_n + \boldsymbol{\Sigma}_n^{1/2} \varepsilon,$$

where \mathbf{V}_n , $\boldsymbol{\Sigma}_n$ and \mathbf{f}_n are as defined in Section 4. Conditional on the covariates and the treatment assignments, the least squares estimate $\hat{\psi} = (\mathbf{V}_n^T \mathbf{V}_n/n)^{-1} \mathbf{V}_n^T \mathbf{y}/n$ has covariance matrix

$$\text{COV}[\hat{\psi}] = \frac{1}{n} \left(\frac{\mathbf{V}_n^T \mathbf{V}_n}{n} \right)^{-1} \left(\frac{\mathbf{V}_n^T \boldsymbol{\Sigma}_n \mathbf{V}_n}{n} \right) \left(\frac{\mathbf{V}_n^T \mathbf{V}_n}{n} \right)^{-1} \quad (\text{A.1})$$

and bias vector

$$E[\hat{\psi} - \psi] = \frac{1}{\sqrt{n}} \left(\frac{\mathbf{V}_n^T \mathbf{V}_n}{n} \right)^{-1} \frac{\mathbf{V}_n^T \mathbf{f}_n}{n}. \quad (\text{A.2})$$

Put $P_n^i = n_i/n$. Then

$$\begin{aligned} \frac{\mathbf{V}_n^T \mathbf{V}_n}{n} &= \sum_{i=1}^p P_n^i \begin{pmatrix} \mathbf{u}_i & \mathbf{u}_i \cdot \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{z}^T(\mathbf{x}_{ij}) \\ * & \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{z}(\mathbf{x}_{ij}) \mathbf{z}^T(\mathbf{x}_{ij}) \end{pmatrix}, \\ \frac{\mathbf{V}_n^T \boldsymbol{\Sigma}_n \mathbf{V}_n}{n} &= \sum_{i=1}^p P_n^i \sigma_i^2 \begin{pmatrix} \mathbf{u}_i & \mathbf{u}_i \cdot \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{z}^T(\mathbf{x}_{ij}) \\ * & \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{z}(\mathbf{x}_{ij}) \mathbf{z}^T(\mathbf{x}_{ij}) \end{pmatrix}, \\ \frac{\mathbf{V}_n^T \mathbf{f}_n}{n} &= \sum_{i=1}^p P_n^i \begin{pmatrix} \mathbf{u}_i \cdot \frac{1}{n_i} \sum_{j=1}^{n_i} f_i(\mathbf{x}_{ij}) \\ \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{z}(\mathbf{x}_{ij}) f_i(\mathbf{x}_{ij}) \end{pmatrix}. \end{aligned}$$

By the strong law of large numbers these expressions of the form $P_n^i \sum_{j=1}^{n_i} A(\mathbf{x}_{ij})/n_i$ tend, with probability 1 as $n_i \rightarrow \infty$, to

$$P(\text{group } i) \cdot E[A(\mathbf{x})|i] = \int_{\mathcal{S}} A(\mathbf{x}) m_i(\mathbf{x}) \mu(d\mathbf{x})$$

and so, with probability 1, $\mathbf{V}_n^T \mathbf{V}_n/n \rightarrow \mathbf{B}$, $\mathbf{V}_n^T \Sigma_n \mathbf{V}_n/n \rightarrow \mathbf{Q}$ and $\mathbf{V}_n^T \mathbf{f}_n/n \rightarrow \mathbf{b} = \mathbf{b}(f_1, \dots, f_p)$ as defined at (8). Thus

$$\begin{aligned} & n\text{MSE}(\hat{\psi}) \\ &= n\text{COV}[\hat{\psi}] + (\sqrt{n}E[\hat{\psi} - \psi])(\sqrt{n}E[\hat{\psi} - \psi])^T \xrightarrow{\text{a.s.}} \mathbf{B}^{-1}(\mathbf{Q} + \mathbf{b}\mathbf{b}^T)\mathbf{B}^{-1}, \end{aligned} \quad (\text{A.3})$$

and $\mathbf{W}_0 \hat{\theta} = \mathbf{W} \hat{\psi}$ satisfies

$$|n\text{MSE}(\mathbf{W}_0 \hat{\theta})| \xrightarrow{\text{a.s.}} |\mathbf{W} \mathbf{B}^{-1} \mathbf{Q} \mathbf{B}^{-1} \mathbf{W}^T| (1 + \mathbf{b}^T \mathbf{B}^{-1} \mathbf{W}^T (\mathbf{W} \mathbf{B}^{-1} \mathbf{Q} \mathbf{B}^{-1} \mathbf{W}^T)^{-1} \mathbf{W} \mathbf{B}^{-1} \mathbf{b}).$$

These expressions are valid for Model 1 as well, with the sole change $\mathbf{v}_i(\mathbf{x}) = \mathbf{u}_i$. \square

Proof of Lemma 1. For (i), note that $\mathbf{T} \mathbf{U}^{-1} \mathbf{T}^T = (\mathbf{T} \mathbf{U} \mathbf{T}^T)^{-1}$. The lower right-hand block of $\mathbf{T} \mathbf{U}^{-1} \mathbf{T}^T$ is $\mathbf{W} \mathbf{U}^{-1} \mathbf{W}^T$; that of $(\mathbf{T} \mathbf{U} \mathbf{T}^T)^{-1}$ is the inverse of the matrix on the right-handside of (i). Now (ii) is a direct calculation. Statement (iii) is obtained by writing $\mathbf{T} \mathbf{U}^{-1} \mathbf{T}^T$ in partitioned form and substituting (ii) into

$$|\mathbf{U}^{-1}| = |\mathbf{T} \mathbf{U}^{-1} \mathbf{T}^T| = \frac{|\mathbf{W} \mathbf{U}^{-1} \mathbf{W}^T|}{p} \{ \mathbf{1}^T \mathbf{U}^{-1} \mathbf{1} - \mathbf{1}^T \mathbf{U}^{-1} \mathbf{W}^T (\mathbf{W} \mathbf{U}^{-1} \mathbf{W}^T)^{-1} \mathbf{W} \mathbf{U}^{-1} \mathbf{1} \}.$$

\square

Proof of Theorem 2. We must establish that, subject to (10), constant $\rho_i(\mathbf{x})$ minimize $\mathcal{L}_0(\rho)$. To obtain $\mathcal{L}_0(\rho)$ we first define matrices

$$\mathbf{P} = \int_{\mathcal{S}} \rho(\mathbf{x}) \mathbf{z}^T(\mathbf{x}) m(\mathbf{x}) \mu(d\mathbf{x}) : p \times q,$$

$$\mathbf{N} = \int_{\mathcal{S}} \mathbf{z}(\mathbf{x}) \mathbf{z}^T(\mathbf{x}) m(\mathbf{x}) \mu(d\mathbf{x}) : q \times q,$$

$$\mathbf{N}_\sigma = \int_{\mathcal{S}} \mathbf{z}(\mathbf{x}) \mathbf{z}^T(\mathbf{x}) \left(\sum_{i=1}^p \sigma_i^2 \rho_i(\mathbf{x}) \right) m(\mathbf{x}) \mu(d\mathbf{x}) : q \times q.$$

If $\mathbf{B}^{-1} \mathbf{Q} \mathbf{B}^{-1}$ is partitioned as at (13), then by Lemma 1(iii)

$$\mathcal{L}_0(\rho) = \frac{\mathbf{1}^T \mathbf{U} \mathbf{1}}{p |\mathbf{U}|}. \quad (\text{A.4})$$

If $\mathbf{B} \mathbf{Q}^{-1} \mathbf{B}$ is partitioned as

$$\mathbf{B} \mathbf{Q}^{-1} \mathbf{B} = \begin{pmatrix} \mathbf{F}_{11} & \mathbf{F}_{12} \\ \mathbf{F}_{21} & \mathbf{F}_{22} \end{pmatrix},$$

then $\mathbf{U} = \mathbf{F}_{11} - \mathbf{F}_{12} \mathbf{F}_{22}^{-1} \mathbf{F}_{21}$. We find that

$$\mathbf{B} = \begin{pmatrix} \mathbf{D}_r & \mathbf{P} \\ \mathbf{P}^T & \mathbf{N} \end{pmatrix}, \mathbf{Q} = \begin{pmatrix} \mathbf{D}_{\sigma^2} \mathbf{D}_r & \mathbf{D}_{\sigma^2} \mathbf{P} \\ \mathbf{P}^T \mathbf{D}_{\sigma^2} & \mathbf{N}_\sigma \end{pmatrix},$$

and we then calculate

$$\mathbf{U} = \mathbf{D} - \mathbf{D}_{\sigma^2}^{-1} \mathbf{P} \mathbf{F}_{22}^{-1} \mathbf{P}^T \mathbf{D}_{\sigma^2}^{-1}.$$

If $\rho_i(\mathbf{x}) \equiv r_i$ then with $\mathbf{n}_{q \times 1} := \int_{\mathcal{S}} \mathbf{z}(\mathbf{x}) m(\mathbf{x}) \mu(d\mathbf{x})$ we obtain

$$\mathbf{U} = \mathbf{D} - \mathbf{D} \mathbf{1} (\mathbf{n}^T \mathbf{F}_{22}^{-1} \mathbf{n}) \mathbf{1}^T \mathbf{D}$$

and then

$$\mathbf{1}^T \mathbf{U} \mathbf{1} = (\text{tr } \mathbf{D})(1 - (\text{tr } \mathbf{D})(\mathbf{n}^T \mathbf{F}_{22}^{-1} \mathbf{n})),$$

$$|\mathbf{U}| = |\mathbf{D}|(1 - (\text{tr } \mathbf{D})(\mathbf{n}^T \mathbf{F}_{22}^{-1} \mathbf{n})).$$

Thus (A.4) reduces to

$$\mathcal{L}_0(\mathbf{r}) = \frac{\mathbf{1}^T \mathbf{D} \mathbf{1}}{p |\mathbf{D}|},$$

establishing (11).

Now define $\mathbf{M} = \mathbf{D}_{\sigma^2}^{-1} \mathbf{P} \mathbf{F}_{22}^{-1} \mathbf{P}^T \mathbf{D}_{\sigma^2}^{-1}$ and for $t \in [0, 1]$ put

$$\mathbf{U}_t = \mathbf{D} - \mathbf{M} + t\mathbf{M},$$

$$l(t) = \frac{\mathbf{1}^T \mathbf{U}_t \mathbf{1}}{p |\mathbf{U}_t|}.$$

Then $\mathbf{U} = \mathbf{U}_0$, $\mathbf{D} = \mathbf{U}_1$ and we will have

$$l(0) = \mathcal{L}_0(\mathbf{r}) \geq \mathcal{L}_0(\mathbf{r}) = l(1), \quad (\text{A.5})$$

thus completing the proof, if $l(\cdot)$ is a decreasing function of t . Note that \mathbf{U}_t is positive definite. With ch_{\max} denoting the maximum characteristic root we find

$$\frac{d}{dt} \log l(t) = \frac{\mathbf{1}^T \mathbf{M} \mathbf{1}}{\mathbf{1}^T \mathbf{U}_t \mathbf{1}} - \text{tr}(\mathbf{U}_t^{-1} \mathbf{M}) \leq ch_{\max} \mathbf{U}_t^{-1} \mathbf{M} - \text{tr}(\mathbf{U}_t^{-1} \mathbf{M}) \leq 0,$$

establishing (A.5). \square

Proof of Theorem 3. To minimize $\text{tr } \mathbf{D} / |\mathbf{D}|$ subject to the given side conditions on the r_i , it is convenient to introduce $\tilde{r}_i = r_i / r_1$, $i = 2, \dots, p$ and $\tilde{r}_1 = r_1 = \left(1 + \sum_{j=2}^p \tilde{r}_j\right)^{-1}$. With respect to these variables $\text{tr } \mathbf{D} / |\mathbf{D}| = (\prod_{i=2}^p \sigma_i^2) g(\tilde{r}_2, \dots, \tilde{r}_p)$, where

$$g(\tilde{r}_2, \dots, \tilde{r}_p) = \left(1 + \sum_{j=2}^p \tilde{r}_j\right)^{p-1} \left(1 + \sum_{j=2}^p \tilde{r}_j \tau_j\right) / \left(\prod_{j=2}^p \tilde{r}_j\right), \quad 0 \leq \tilde{r}_2, \dots, \tilde{r}_p < \infty.$$

Since $g \rightarrow \infty$ as any $\tilde{r}_i \rightarrow 0, \infty$ there is a minimum, corresponding to a critical point. We have

$$\frac{\partial \log g}{\partial \tilde{r}_i} = \frac{1}{\tilde{r}_i} \left(\frac{(p-1)\tilde{r}_i}{1 + \sum_{j=2}^p \tilde{r}_j} + \frac{\tilde{r}_i \tau_i}{1 + \sum_{j=2}^p \tilde{r}_j \tau_j} - 1 \right).$$

In terms of the original variables a critical point then satisfies the equations

$$r_i = (p - 1 + \alpha^{-1} \tau_i)^{-1}, \quad i = 2, \dots, p \quad (\text{A.6})$$

$$\alpha = 1 + \sum_{j=2}^p (\tau_j - 1)r_j. \quad (\text{A.7})$$

Substituting (A.6) into (A.7) gives the equation $h(\alpha)=0$. Note that $h(0)=-1$, $h(\infty)=\infty$ so that there is a zero, necessarily ≥ 1 since $h(\alpha) \leq \alpha - 1$. Since

$$h''(\alpha) = 2(p-1) \sum_{j=2}^p \tau_j(\tau_j - 1)((p-1)\alpha + \tau_i)^{-3} \geq 0,$$

h is convex and the zero α_* is unique. With r_i^* given by (A.6) with $\alpha = \alpha^*$ we have

$$r_1^* = 1 - \sum_{i=2}^p r_i^* \geq 1 - \sum_{i=2}^p (p-1 + \alpha_*^{-1})^{-1} = (\alpha^*(p-1) + 1)^{-1} > 0,$$

so that r_1^*, \dots, r_p^* satisfy the required side conditions. \square

References

- Atkinson, A.C., 1982. Optimum biased coin designs for sequential clinical trials with prognostic factors. *Biometrika* 69, 61–67.
- Atkinson, A.C., 2002. The comparison of designs for sequential clinical trials with covariate information. *J. Roy. Statist. Soc. A* 165, 349–373.
- Bather, J., 1995. Response adaptive allocation and selection bias. In: Flournoy, N. Rosenberger, W.F., (Eds.), *Adaptive Designs*, Institute of Mathematical Statistics, Hayward, CA, pp. 23–35.
- Begg, C.B., Iglewicz, B., 1980. A treatment allocation procedure for sequential clinical trials. *Biometrics* 36, 81–90.
- Efron, B., 1971. Forcing a sequential experiment to be balanced. *Biometrika* 58, 403–417.
- Heckman, N.E., 1987. Robust design in a two treatment comparison in the presence of a covariate. *J. Statist. Plann. Inference* 16, 75–81.
- Maronna, R.A., Yohai, V.J., 1981. Asymptotic behaviour of general M-estimators for regression and scale with random carriers. *Z. Wahrscheinlichkeitstheorie und Verwandte Gebiete* 58, 7–20.
- Pocock, S.J., Simon, R., 1975. Sequential treatment assignments with balancing for prognostic factors in the controlled clinical trial. *Biometrics* 31, 103–115.
- Taves, D.R., 1974. Minimization: A new method of assigning patients to treatment and control groups. *Clin. Pharmacol. Ther.* 15, 443–453.