The Use of Group Sequential Tests with Designs which Adjust for Imbalances in Prognostic Factors

Yolanda Barbáchano[1] and Steve Coad[2]

[1] Department of Clinical Research and Development,
Royal Marsden Hospital,
Downs Road,
Sutton SM2 5PT, UK
yolanda.barbachano@rmh.nhs.uk

[2] School of Mathematical Sciences,
Queen Mary, University of London,
Mile End Road,
London E1 4NS, UK
d.s.coad@qmul.ac.uk

Abstract. Minimisation and methods that make use of optimum design theory have been suggested to balance treatment groups across prognostic factors. Although the problem of analysing a trial when one of these methods has been used has been looked at in the fixed-sample case, it has so far not been considered in the group sequential setting. In this paper, simulation is used to explore the consequences of adapting for prognostic factors in a group sequential trial. Both Pocock's test and the O'Brien and Fleming test are considered and three methods of adjusting for covariates are studied. When the variance of the response variables is unknown, the critical values are obtained using those in the known variance case and the significance level approach. The resulting tests have approximately the required type I error probability. To maintain the desired power, sample size re-estimation is incorporated. Simulation shows that the tests satisfy the power requirement for moderate sample sizes, with complete randomisation being less powerful than the adaptive methods. Repeated confidence intervals for the mean treatment effects are also calculated.

Keywords. adaptive design, D_A-optimum design, general linear model, interim analysis, minimisation, repeated confidence interval.

1 Introduction

In clinical trials, it is sometimes felt necessary, for ethical or economical reasons, to have several interim analyses rather than just one at the end. This means that, if one treatment is obviously performing better than the others, the trial can be stopped early, avoiding further patients from being treated with the inferior treatments. When planning a group sequential trial like this, one has to decide how to preserve the type I error probability until the end of the trial, by not spending the whole of it at the first analysis. Pocock's test and the O'Brien and Fleming test provide two methods of achieving this. Their tests provide a way of adjusting the critical values so that the type I error probability is maintained at the nominal level and they also indicate how to calculate the sample size in order to attain a desired power.

If the variance is unknown, the sample size has to be re-calculated at each interim analysis, using the current estimate of the variance (Morgan, 2003). In order to do this, one must first guess an estimate of the variance and use it to calculate the size of a pilot study. When all the patients needed for the first analysis have been recruited, the variance of their responses is calculated and this is used as the new estimate of the population variance. The sample size is then re-calculated with this new estimate and more patients are recruited, if necessary, before performing the first analysis. If more patients than necessary have already been enrolled, it might be appropriate to leave out the first analysis and go straight to what would have been the second or third one. After any analysis, if the trial needs to continue, the number of patients required is re-calculated using the latest estimate of the variance.

Another issue to consider when designing a trial is whether apart from the treatment, other factors, such as age or gender, are likely to affect a patient's response. If this is the case then it would be desirable to balance the treatment groups across the prognostic factors and minimisation (Taves, 1974; Pocock and Simon, 1975) is a method often used to achieve this. It works by looking at the characteristics of the current patient, checking how many patients there are so far with the same levels of the factors in

each treatment group and then giving the patient the treatment that will most reduce the imbalance. To avoid predictability, a biased coin can be used. Atkinson (1982) suggested an alternative to minimisation which makes use of optimum design theory. Every time a new patient comes along, the variance of the parameter estimates is calculated and the patient is assigned to the treatment that produces the smallest variance by some criterion.

2 Group sequential tests

2.1 The model

Suppose that we wish to compare two treatments. Let Y_i denote the response for the ith patient and let μ_j denote the mean effect for the jth treatment for $i = 1, 2, \ldots, n$ and $j = 1, 2$. Furthermore, let $\underline{x}_i = (x_{1i}, \ldots, x_{fi})^T$ denote the vector of f covariates for the ith patient and let $\underline{b} = (b_1, \ldots, b_f)^T$ denote the vector of regression coefficients of the response on the covariates. Then the linear model considered is

$$Y_i = \sum_{j=1}^{2} \delta_{ij}\mu_j + \underline{b}^T \underline{x}_i + \epsilon_i,$$

where $\delta_{ij} = 1$ if the ith patient is assigned to treatment j and $\delta_{ij} = 0$ otherwise, $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$ is the error term and the ϵ_i are independent. Obviously, $\sum_{j=1}^{2} \delta_{ij} = 1$ for all i.

Clearly, we can write the above model in the form of the general linear model given by

$$\underline{Y}_n = Z_n\underline{\theta} + \underline{\epsilon}_n,$$

where $\underline{Y}_n = (Y_1, \ldots, Y_n)^T$, $\underline{\theta} = (\mu_1, \mu_2, \underline{b}^T)^T$ and $\underline{\epsilon}_n = (\epsilon_1, \ldots, \epsilon_n)^T$, and

$$Z_n = \begin{pmatrix} \underline{\delta}_1^T & \underline{x}_1^T \\ \vdots & \vdots \\ \underline{\delta}_n^T & \underline{x}_n^T \end{pmatrix}$$

is the design matrix. Note that $\underline{\delta}_i = (\delta_{i1}, \delta_{i2})^T$ for $i = 1, 2, \ldots, n$. Since the errors are normal, the maximum likelihood estimator of $\underline{\theta}$ is

$$\underline{\hat{\theta}}_n = (Z_n^T Z_n)^{-1} Z_n^T \underline{Y}_n.$$

If σ^2 is unknown, its usual estimator is

$$\hat{\sigma}_n^2 = \frac{(\underline{Y}_n - Z_n\underline{\hat{\theta}}_n)^T(\underline{Y}_n - Z_n\underline{\hat{\theta}}_n)}{n - f - 2}$$

for $n > f + 2$.

Although the forms of the above estimators are not affected by an adaptive design, their sampling distributions are. However, when patients have been completely randomised to treatments, we know that, conditional on the $N_j = \sum_{i=1}^{n} \delta_{ij}$,

$$\underline{\hat{\theta}}_n \sim \mathrm{N}_{f+2}\left\{\underline{\theta}, (Z_n^T Z_n)^{-1}\sigma^2\right\}$$

and $(n - f - 2)\hat{\sigma}_n^2/\sigma^2 \sim \chi^2_{n-f-2}$. We wish to test whether the treatments have the same mean effect, and so the null hypothesis is $H_0 : \mu_1 = \mu_2$. Clearly, we can rewrite this as $H_0 : \underline{c}^T\underline{\theta} = 0$, where $\underline{c} = (1, -1, 0, \ldots, 0)^T$. It follows that

$$T = \frac{\underline{c}^T\underline{\hat{\theta}}_n}{\hat{\sigma}_n/\sqrt{\underline{c}^T(Z_n^T Z_n)^{-1}\underline{c}}} \sim t_{n-f-2}$$

under H_0.

2.2 Form of group sequential test

Suppose that we wish to compare the two treatments using a group sequential test with a maximum of K interim analyses. Let $\hat{\underline{\theta}}^{(k)}$ and s_k^2 denote the above estimators of $\underline{\theta}$ and σ^2 based on the data up to the kth interim analysis, and let $Z^{(k)}$ denote the design matrix based on the same information for $k = 1, 2 \ldots, K$. Then the test statistic at analysis k is

$$T_k = \frac{\underline{c}^T \hat{\underline{\theta}}^{(k)}}{s_k / \sqrt{\underline{c}^T \{Z^{(k)^T} Z^{(k)}\}^{-1} \underline{c}}} \sim t_{n_k - f - 2}$$

under H_0, where n_k denotes the total number of patients in the first k groups.

The group sequential test rejects H_0 at the kth interim analysis if $|T_k| \geq c_{k,\alpha}$ for $k = 1, 2, \ldots, K$, where the $c_{k,\alpha}$ are chosen to give an overall significance level of $100\alpha\%$. Thus, if α_k is the nominal significance level at analysis k, the $c_{k,\alpha}$ satisfy the equation

$$P(|T_k| \geq c_{k,\alpha} | H_0) = \alpha_k$$

for $k = 1, 2, \ldots, K$. The general theory of this test has been developed by Jennison and Turnbull (1997). Two tests are considered in this paper, those of Pocock and O'Brien and Fleming.

2.3 Critical values and group sizes

Suppose that we wish to test $H_0 : \mu_1 - \mu_2 = 0$ with two-sided type I error probability α and power $1 - \beta$ when $\mu_1 - \mu_2 = \pm\delta$. When σ^2 is known, the fixed-sample test requires Fisher information for $\mu_1 - \mu_2$ given by

$$I_{f,2} = \frac{\{\Phi^{-1}(1 - \alpha/2) + \Phi^{-1}(1 - \beta)\}^2}{\delta^2},$$

where Φ denotes the standard normal distribution function. From Jennison and Turnbull (2000), the group sequential test must have a maximum information level for $\mu_1 - \mu_2$ of $R(K, \alpha, \beta) I_{f,2}$, where the value of $R(K, \alpha, \beta)$ depends on the test being used. Dividing this information equally between the K analyses gives the required information levels

$$I_k = \frac{k}{K} R(K, \alpha, \beta) I_{f,2}$$

for $k = 1, 2, \ldots, K$. So if n_k patients, divided equally between the two treatments, are observed by analysis k, we have

$$\mathrm{var}(\hat{\mu}_{k1} - \hat{\mu}_{k2}) \simeq \frac{4\sigma^2}{n_k},$$

where $\hat{\mu}_{kj}$ denotes the maximum likelihood estimator of μ_j based on the data up to the kth analysis, which means that $n_k = 4\sigma^2 I_k$. It follows that

$$n_1 = \frac{4\sigma^2}{K} R(K, \alpha, \beta) I_{f,2}$$

patients per group will be needed. When σ^2 is unknown, we can re-calculate the sample size as the trial proceeds using the current estimate of the variance.

Pocock's test rejects H_0 at stage k if

$$|T_k| \geq t_{n_k - f - 2, 1 - \Phi\{C_P(K,\alpha)\}},$$

where $t_{\nu,\gamma}$ denotes the upper $100\gamma\%$ point of the t distribution with ν degrees of freedom and the constants $C_P(K, \alpha)$ correspond to the critical values in the known variance case. Note that, although this test satisfies the type I error probability requirement only approximately, this approximation has

been found to be remarkably accurate (Jennison and Turnbull, 2000). The O'Brien and Fleming test rejects H_0 at stage k if

$$|T_k| \geq t_{n_k-f-2,1-\Phi\{C_B(K,\alpha)\sqrt{K/k}\}},$$

where the constants $C_B(K,\alpha)$ again correspond to the critical values in the known variance case. As above, this test is only approximate.

2.4 Repeated confidence intervals

For notational convenience, let $M_k = {Z^{(k)}}^T Z^{(k)}$ and write

$$M_k^{-1} = \begin{pmatrix} M_k^{11} & M_k^{12} \\ M_k^{21} & M_k^{22} \end{pmatrix},$$

where M_k^{11} is a 2×2 matrix. Then, when patients have been completely randomised to treatments, conditional on the $N_{kj} = \sum_{i=1}^{n_k} \delta_{ij}$, we know that

$$\hat{\mu}_{k1} - \hat{\mu}_{k2} \sim \mathrm{N}(\mu_1 - \mu_2, \underline{c}^T M_k^{11} \underline{c} \sigma^2).$$

It follows that repeated $100(1-\alpha)\%$ confidence intervals for $\mu_1 - \mu_2$ are given by

$$\mathcal{I}_k = \hat{\mu}_{k1} - \hat{\mu}_{k2} \pm c_{k,\alpha} s_k \sqrt{\underline{c}^T M_k^{11} \underline{c}}$$

for $k = 1, 2, \ldots, K$.

3 Covariate-adaptive designs

3.1 Minimisation

Minimisation was introduced by Taves (1974) and Pocock and Simon (1975). It works by calculating the treatment imbalance every time a new patient enters the trial, and then giving that patient the treatment that would most reduce this imbalance. The method that Taves (1974) introduced is described below.

At an arbitrary point in the trial, let m_{ruj} be the number of patients with level u of factor r who have been given treatment j for $r = 1, 2, ..., f$, $u = 1, 2, ..., f_r$ and $j = 1, 2, \ldots, t$, where f_r denotes the number of levels of factor r. Now suppose that the next patient that enters the trial has levels $l_1, l_2, ..., l_f$ of the prognostic factors. Then, to determine which treatment should be given to this patient, one calculates the effect that each treatment assignment would have on the overall imbalance. For each treatment j, one considers the new $\{m_{ruk}\}$, denoted $\{m^j_{ruk}\}$, that would arise if that treatment was assigned to the patient. This means that

$$m^j_{ruk} = \begin{cases} m_{ruk} & \text{if } u \neq l_r \text{ or } k \neq j, \\ m_{ruk} + 1 & \text{if } u = l_r \text{ and } k = j. \end{cases}$$

One then calculates the range of $\{m^j_{r,l_r,k}, k = 1, 2, \ldots, t\}$ for patients with level l_r of factor r to determine the treatment imbalance that would result at this level if treatment j was given next. This range is then added to those for the other levels of the patient to obtain the overall treatment imbalance. The treatment that produces the least overall imbalance would be the one given to the patient. A biased coin can be used to avoid predictability. If this is done, then the treatment that most reduces the imbalance is assigned with probability p, $0.5 < p < 1$, and the other $t-1$ treatments are assigned with probability $(1-p)/(t-1)$.

Minimisation was not designed for a group sequential trial, but it can be applied in this setting. In this paper, minimisation was employed in the usual way by assigning each patient to a treatment when they joined the trial, rather than when the whole group of patients has been enrolled. After each interim analysis, if the trial needs to continue, the next patient is assigned taking into account the treatment imbalance produced by those already in the trial.

3.2 Atkinson's method

Atkinson (1982) introduced a procedure of the biased coin type, which made use of optimum design theory, and could be used to balance over prognostic factors with any number of treatments. This method is based on the model

$$E(Y) = \underline{z}^T \underline{\beta} = \underline{z}_1^T \underline{\beta}_1 + \underline{z}_2^T \underline{\beta}_2,$$

where Y is the treatment response, $\underline{z}_1$ is the vector of indicator variables for the treatments, $\underline{\beta}_1$ is the vector of treatment mean effects, $\underline{z}_2$ is the vector of f prognostic factors and $\underline{\beta}_2$ is the vector of regression coefficients.

Interest is in the contrasts between treatment effects, which can be written as s linear combinations which are the components of $A\underline{\beta}_1$, where A is a $s \times t$ contrast matrix of rank $s < t$. The nuisance parameters $\underline{\beta}_2$ are not of interest, and so we concentrate on the linear combination $C\underline{\beta}$, where $C = (A : 0)$ and 0 is a $s \times f$ matrix of zeroes. The covariance matrix of the least squares estimator $C\underline{\hat{\beta}}$ is proportional to $CM_n^{-1}C^T$, where $M_n = n^{-1}(Z_n^T Z_n)$ and $Z_n^T Z_n$ is the $(t+f) \times (t+f)$ information matrix which results from the n observations. This symmetric matrix can be partitioned according to the above model as

$$M_n^{-1} = \begin{pmatrix} M_n^{11} & M_n^{12} \\ M_n^{21} & M_n^{22} \end{pmatrix}.$$

Now, D_A-optimality maximises $\det\{(CM_n^{-1}C^T)^{-1}\}$ in order to minimise the generalised variance of $C\underline{\hat{\beta}}$. According to the equivalence theorem for D_A-optimality, this can be achieved by giving the $(n+1)$st patient the treatment for which the standardised variance of the predicted response is at a maximum. For the model given above, the standardised variance for treatment j is

$$d_A(j,n) = \underline{z}_1^T M_n^{11} B_n M_n^{11} \underline{z}_1 + 2\underline{z}_1^T M_n^{11} B_n M_n^{12} \underline{z}_2 + \underline{z}_2^T M_n^{21} B_n M_n^{12} \underline{z}_2,$$

where $\underline{z}_1$ has one for its jth component and zeroes elsewhere, $\underline{z}_2$ contains the values of the prognostic factors for the $(n+1)$st patient and $B_n = A^T(AM_n^{11}A^T)^{-1}A$. Therefore, the deterministic design would allocate the next patient to the treatment with the largest $d_A(j,n)$.

A biased coin can be introduced to reduce the predictability. Atkinson (1982) suggested choosing treatment j with probability

$$p_j = \frac{d_A(j,n)}{\sum_{k=1}^t d_A(k,n)}$$

instead of choosing the treatment with the largest $d_A(j,n)$ straight away.

Ball et al. (1993) introduced a biased coin which takes into account the balance between the variance that one is aiming to reduce and the entropy, a measure of predictability. Since this biased coin design was derived within a decision-theoretic Bayesian framework, we call it a Bayesian biased coin design. The probabilities of treatment selection are chosen to maximise a utility which combines both the variance of the parameter estimates and randomness. In the D_A-optimality case, the probabilities are

$$p_j = \frac{\{1 + d_A(j,n)\}^{1/\xi}}{\sum_{k=1}^t \{1 + d_A(k,n)\}^{1/\xi}},$$

where the design parameter $\xi \geq 0$ is a trade-off coefficient between efficient inference when $\xi = 0$ and complete randomisation when $\xi \to \infty$. So the Bayesian biased coin design can be regarded as a variant of Atkinson's method.

Although these designs were not developed with a group sequential trial in mind, they can be easily used in this case. After each interim analysis, if the trial needs to continue, the information matrix generated by all of the patients already in the trial is used to calculate the treatment allocation for the next patient.

4 Simulation results

4.1 Choice of design parameters

Simulation was used to study the effect of using the above three covariate-adaptive designs in a group sequential trial. For comparison purposes, results were also obtained for complete randomisation. For Atkinson's method, we used $A = (1, -1)$, so that $s = 1$. Minimisation was used with $p = 0.7$ and the Bayesian biased coin design had $\xi = 0.01$, since these values reduce the predictability considerably without increasing too much the treatment imbalance (Barbáchano et al., 2008). We had a maximum of four interim analyses, chose $\alpha = 0.05$ and aimed for 90% power when $\mu_1 - \mu_2 = \pm\delta$. Assuming that $\sigma^2 = 1$, when $\delta = 0.9$, 16 patients per group were needed for Pocock's test and 14 for the O'Brien and Fleming test. Similarly, when $\delta = 0.5$, the respective numbers of patients per group were 50 and 43. The significance level and power for a range of mean treatment effect differences were estimated after 10,000 simulations for each design. In each case, we took $\underline{X} \sim \mathrm{N}_f(\underline{0}, I_f)$, where I_f is the identity matrix of order f, so that the prognostic factors are uncorrelated.

4.2 Significance level and power

When the preliminary estimate of the variance, s_P^2, is equal to the true value and the sample size is not re-estimated, the results show that, when there are two prognostic factors, all the allocation methods achieve a valid significance level. However, the O'Brien and Fleming test seems to be more powerful than Pocock's test and Atkinson's method with either biased coin tends to be more powerful than minimisation, and this method in turn has more power than complete randomisation. In the case of six prognostic factors, the actual significance levels are slightly higher than the nominal value of 0.05, especially for Pocock's test. The differences in power between the allocation rules are also more obvious in this case, with differences as large as 0.06 in some instances. The expected sample size is often higher for complete randomisation than for the covariate-adaptive methods, which reflects the variability of the rules.

When the variance is unknown and sample size re-estimation is being used, the difference in treatment effects that we are trying to detect will have an impact on the power. For example, if we aim to detect a treatment effect difference of 0.9 with 90% power, only about 80% power is obtained when $s_P^2/\sigma^2 = 0.5$. This is because the sample size needed to detect this difference is quite small, and therefore there is not enough time to correct the variance estimate before the trial ends. However, if we aim to detect a difference in treatment effects of 0.5 with 90% power, we obtain more than 86% power in all cases. The different allocation methods produce very similar results, though Atkinson's method tends to be slightly more powerful. It does not seem to make much difference to the power whether s_P^2 underestimates or overestimates σ^2.

References

Atkinson, A. C. (1982). Optimum Biased Coin Designs for Sequential Clinical Trials with Prognostic Factors. *Biometrika*, **69**, 61–67.

Ball, F. G., Smith, A. F. M. and Verdinelli, I. (1993). Biased Coin Designs with a Bayesian Bias. *Journal of Statistical Planning and Inference*, **34**, 403–421.

Barbáchano, Y., Coad, D. S. and Robinson, D. R. (2008). Predictability of Designs which Adjust for Imbalances in Prognostic Factors. *Journal of Statistical Planning and Inference*, **138**, 756–767.

Jennison, C. and Turnbull, B. W. (1997). Distribution Theory of Group Sequential t, χ^2 and F Tests for General Linear Models. *Sequential Analysis*, **16**, 295–317.

Jennison, C. and Turnbull, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials*. Chapman and Hall, London.

Morgan, C. C. (2003). Sample Size Re-Estimation in Group-Sequential Response-Adaptive Clinical Trials. *Statistics in Medicine*, **22**, 3843–3857.

Pocock, S. J. and Simon, R. (1975). Sequential Treatment Assignment with Balancing for Prognostic Factors in the Controlled Clinical Trial. *Biometrics*, **31**, 103–115.

Taves, D. R. (1974). Minimisation: A New Method of Assigning Patients to Treatment and Control Groups. *Clinical Pharmacology and Therapeutics*, **15**, 443–453.