Adaptive Clinical Trials Incorporating Treatment Selection and Evaluation: Methodology and Applications in Multiple Sclerosis

Susan Todd[1], Tim Friede[2], Nigel Stallard[2], Nicholas Parsons[2], Elsa Valdés-Márquez[1], Jeremy Chataway[3] and Richard Nicholas[4]

[1] Applied Statistics, University of Reading, Philip Lyle Building, Reading, RG6 6BX, UK
(e-mail of corresponding author: s.c.todd@reading.ac.uk)

[2] Warwick Medical School, University of Warwick, Coventry, CV4 7AL, UK

[3] Imperial College Healthcare NHS Trust / National Hospital for Neurology & Neurosurgery, Department of Neurology, St Mary's Hospital, Praed Street, London, W2 1NY, UK

[4] Imperial College Healthcare NHS Trust, Department of Cellular & Molecular Neurosciences, Charing Cross Hospital, Fulham Palace Road, London, W6 8RF, UK

Abstract. Recent advances in statistical methodology for clinical trials have led to the development of seamless adaptive designs. These allow a number of experimental treatments to be compared with a control in the same trial, with less promising treatments dropped from the study early, whilst maintaining control of type I error rate. From a statistical viewpoint, the treatment selection may be made in any way without compromising the type I error rate control. From a practical aspect, however, the treatment selection rule must be chosen carefully; too stringent a rule increases the risk of erroneously dropping effective treatments, whereas too lax a rule leads to ineffective treatments remaining in the trial, less efficient use of resources and a reduction in power. In this paper we explore the implementation of a seamless adaptive design in the setting of multiple sclerosis. The objective is to illustrate how the statistical aspects of the trial design are brought together with the practical considerations of running a trial. Statistical simulation studies are a powerful tool for exploring the properties of different treatment selection rules and a key objective is to explore their use in this setting. Conclusions focus on the important aspects which need to be considered when planning the actual trial.

Keywords. Adaptive designs, interim analyses, phase II/III clinical trials, simulation study.

1 Introduction

The clinical evaluation of new drugs is generally divided into three phases (Chow and Liu, 2004). The first two phases may be considered exploratory, and allow drug tolerability to be explored and initial assessment of efficacy to be made, often on the basis of short-term endpoints. Phase III clinical trials are confirmatory and provide definitive evidence of treatment effect. This is usually via a comparison with a randomised placebo control group using long-term endpoints and a clinically realistic patient population.

The assessment of new treatments for multiple sclerosis (MS) presents a number of features that mean that the common phase I-II-III paradigm for clinical trial design may not be the most appropriate. The simultaneous availability of a number of drugs for evaluation means that it is desirable to compare several active treatments in the same trial rather than conducting trials with a single new treatment compared with a placebo. Furthermore, relatively long follow-up times are required to assess the clinical efficacy of new treatments in this indication (CHMP, 2005). In these circumstances small scale Phase II trials of individual drugs have to rely on early outcomes and might be inefficient.

Recently, advances in statistical methodology for clinical trials have led to the development of *seamless adaptive designs* (Maca et al, 2006; Bretz et al, 2006). These allow a number of experimental treatments to be compared with a control in the same trial, with less promising treatments dropped from the trial early. This paper explores the applicability of adaptive methodology in the setting of clinical trials for MS, and presents an approach for determining an efficient clinical trial design for the selection and evaluation of potential new MS therapies. In Section 2 the methodology will be presented briefly. In Section 3 an explanation is given of how simulations, based on a realistic model constructed from analysis of data from previous trials in MS, can be used prior to the commencement of an actual clinical trial to ensure that treatments will be dropped appropriately. Section 4 presents an illustrative example. The paper concludes with a brief discussion in Section 5.

2 Adaptive design methodology applied to MS

A particularly important type of adaptation in the setting of clinical trials in multiple sclerosis as described above is the ability to select promising treatments during the trial or to stop the trial early. It is assumed that the trial is conducted in two stages, and involves at least two experimental treatments and a single control treatment. In the first stage, patients are randomised to receive one of the experimental treatments or the control treatment. The treatments are then compared at the end of stage one using one of a number of possible selection rules based on an 'early' outcome measure. Based on this early outcome measure, either the trial is stopped for futility or one or more of the experimental treatments are selected to continue into stage two of the trial. The trial then continues with randomisation between the control treatment and the remaining experimental treatments. At the end of the second stage, the remaining experimental treatments are compared, using a clinically meaningful outcome measure, with the control treatment in a series of formal hypothesis tests based on data from both design stages. The hypothesis testing is carried out in such a way that the familywise type I error rate is strongly controlled, allowing both for the multiple comparisons and for the treatment selection.

At the outset of the trial assume that there are k_1 experimental treatments and a single control treatment. On completion of the first stage of the trial some experimental treatments might be dropped based upon observation of the early outcome measure, and a non-empty subset of k_2 treatment arms $(0 < k_2 \leq k_1)$ together with the control are continued into stage two. In what follows, denote by $\theta_i, i = 1,\ldots,k_1$ the treatment effects of the k_1 experimental treatments each compared to the control. Furthermore, let $Z_{i,1}$ and $Z_{i,2}$ denote standardised test statistics in the first and second stage, respectively, where $Z_{i,2}$ is based on the data of the second stage only and not on the accumulated data. The null hypotheses of interest are denoted by $H_i : \theta_i = 0$. These are tested against the one-sided alternative hypotheses $H_i' : \theta_i > 0$. The sample sizes per group in stages one and two are denoted by n_1 and n_2 respectively and the $Z_{i,j}$ follow Normal distributions with expectations $\sqrt{n_j/2}\theta_i$ and variance 1. The standardised test statistic based on the accumulated data at the end of the second stage is denoted by Z_i. It holds that $Z_i = \sqrt{n_1/n}Z_{i,1} + \sqrt{n_2/n}Z_{i,2}$ for a total sample size per group of $n = n_1 + n_2$.

Dunnett's test (1955) can be used to compare the test treatments to the control treatment based on the late outcome measure at stage one and stage two. The test is a test for many-to-one comparisons testing the null hypothesis that all θ_i are equal to 0 against the alternative hypothesis that at least one θ_i is larger than 0. The test statistic is given by $Z^{\max} = \max_{i\in\{1,\ldots,k_1\}} Z_i$. On observing a $Z^{\max} = z$ a p-value is calculated as $1 - F_{Z^{\max}}(z)$ where

$$F_{Z^{\max}}(z) = \int_{-\infty}^{\infty} [\Phi(\sqrt{2}z + x)]^{k_1} \phi(x) dx . \qquad (1)$$

This expression is the cumulative distribution function of $Z^{\max}$ with $\Phi(.)$ and $\phi(.)$ denoting the cumulative distribution function and the density function of the standard normal distribution, respectively. An objective of the adaptive design is to control the familywise type I error rate in the strong sense at a pre-specified level α. In order to do this, the closure test principle is applied (Marcus *et al.*, 1976). This means that an individual null hypothesis H_i is rejected only if all intersection hypotheses $H_S = \bigcap_{i\in S} H_i$ $(S \subseteq \{1,\ldots,k_1\})$ with index sets S that include i are rejected at level α. Considering the case of a trial with two test treatments, the intersection hypothesis $H_{\{1,2\}}$ is the null hypothesis that $\theta_1 = \theta_2 = 0$; i.e. that neither treatment is effective. An intersection hypothesis H_S is tested using

a Dunnett test which includes all treatments in S. The null hypothesis H_S is rejected if $\max_{i\in S} Z_i \geq d_s$ where the critical value d_s (s : number of elements in S) is the solution of equation (1) set to $1-\alpha$ with $k_1 = s$. The test procedure is modified to allow for dropped treatments by setting test statistics Z_i, comparing dropped treatments to the control, to $-\infty$ (Koenig *et al.*, 2008).

Inferences can be made using the late outcome measure, by combining stagewise p-values from stages one and two of the trial; the 'combination test approach'. Bauer and Kieser (1999) describe the fundamental ideas of the combination test approach referring to the combination function described by Bauer and Köhne (1994) to combine stagewise p-values which allows for interim adaptations and the application of the closed test principle to control the overall size of the test across multiple hypotheses. The method for combining p-values that is used here is the *weighted inverse normal method* as described by Lehmacher and Wassmer (1999). Denoting the two p-values from the different stages of the trial by p_1 and p_2, the weighted inverse normal combination function for these p-values is given by

$$C(p_1, p_2) = 1 - \Phi(w_1\Phi^{-1}(1-p_1) + w_2\Phi^{-1}(1-p_2)) \qquad (2)$$

where w_j, $j = 1,2$ are pre-specified weights with $0 < w_j \leq 1$ and $w_1^2 + w_2^2 = 1$. The most widely used option for the weights is $w_j = \sqrt{n_j/n}$. If $C(p_1, p_2) \leq \alpha$ then the null hypothesis is rejected. When a single hypothesis is tested, the inverse normal method with the weights described above is equivalent to a classical group-sequential test. The stagewise Dunnett p-values for hypothesis H_S are calculated for the test statistics of the first stage $\max_{i\in S} Z_{i,1}$ and second stage $\max_{i\in S} Z_{i,2}$ according to equation (1) and these are then combined using equation (2).

Three distinct types of selection rules have been considered. These are all based on an early outcome measure. The treatment effects of the early outcome measure for the k_1 experimental treatments are denoted by $\pi_i, i = 1,\ldots,k_1$ and assumed to follow a Normal distribution with mean π_i and variance 1. The three selection rules are as follows;

(i) A *fixed* selection rule, selected the m largest values of π_i for $m = 1$, $m = 2$ and $m = k_1$.

(ii) A *variable* (*epsilon*) selection rule, continued with all treatments with an early outcome measure where $\pi_i \geq \pi^{\max} - \varepsilon$. This includes the extremes of selecting exactly one treatment ($\varepsilon = 0$) and selecting all treatments ($\varepsilon = \infty$).

(iii) A *futility* selection rule, selected only those treatments where $\pi_i \geq C$, for threshold value C.

3 Strategy for simulations

A general framework for the comprehensive evaluation of competing options for clinical programs, trial designs and analysis methods, as a basis for decision making has been proposed by Benda et al. (2009). They introduced key terminology and definitions, and described the overall process as Clinical Scenario Evaluation (CSE); shown in Figure 1. Statistical models that describe the data generation process in a clinical trial and the specification of parameters such as treatment effects, correlations and standard deviations are called *assumptions*. The range of assumptions under consideration are referred to as the *assumption set*. Particular clinical trial designs and variants of such are called *options*. Again the range of options relevant to the comparison at hand are referred to as the *option set*. The combination of the assumption set and the option set are the *clinical scenarios*. In order to judge whether a design is efficient or robust we need to define criteria by which we judge the designs. These criteria are referred to as *metrics*. In the context of clinical trials metrics of interest are for example

statistical power and selection probabilities of particular treatments at interim. The combination of metrics to be used is called *metrics set*. Finally, *clinical scenario evaluation* is defined as the comparison of the metrics of alternative clinical trial design options for a particular assumption set. We adopt and use this convenient CSE framework and terminology throughout this paper.

For any *clinical scenario*, the methods described in Section 2 can be implemented as follows. For given treatment means for the early outcome π_i, a selection rule is used at the end of stage one to choose a number of treatments to progress to stage two. Late outcome data from stages one and two are combined together using Dunnett's test, for a many-to-one comparison against the control treatment, using the weighted inverse normal method Hypothesis testing proceeds using a closed test procedure, such that for elementary hypotheses H_i for $i = 1,2,\ldots,k_1$ and intersection hypotheses $H_S = \bigcap_{i \in S} H_i$ for $S \subseteq \{1,\ldots,k_1\}$, H_i is rejected if all intersection hypotheses H_S with $i \in S$ are rejected at level α. This closed testing procedure controls the familywise error rate in the strong sense.

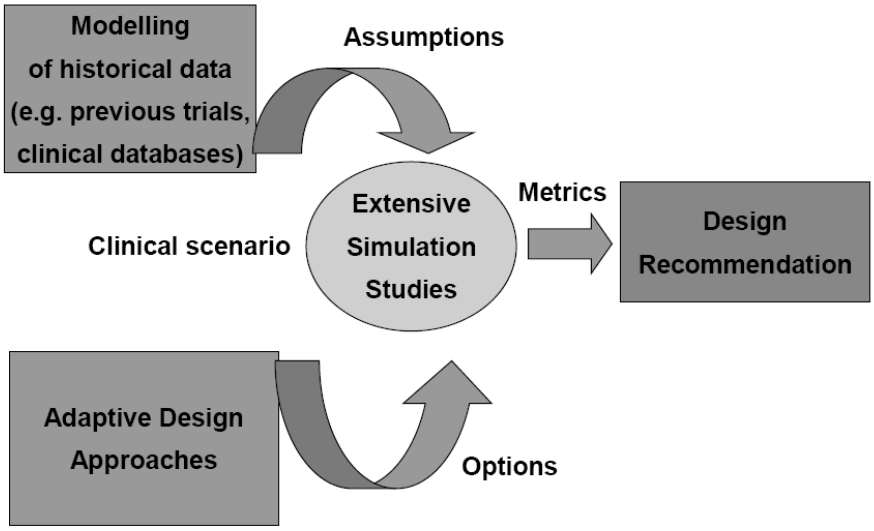

Fig. 1. Clinical Scenario Evaluation

Programs were written in R (http://www.r-project.org/) to implement the methodologies described above. Following the CSE framework, four design *options* for the MS scenario were identified. These are

(a) Number of experimental treatments

(b) Total number of patients in the trial

(c) Patient numbers at stages 1 and 2

(d) Selection rule at the interim analysis (see above)

In order to determine suitable *assumptions* to be used in the CSE framework, three key sources of information were used. First, the results from a comprehensive literature review of the MS literature; second, detailed analysis of clinical datasets supplied to the project team and third, consultation with experts in the field. The following four assumptions were identified

(a) Standardised treatment effect size

(b) Treatment effect type

(c) Early and final treatment effect sizes

(d) Correlation between early and final outcomes

A variety of settings were made for each of the options and assumptions in order to evaluate a range of potential design scenarios. Particular combinations of the options and assumptions make up clinical scenarios for evaluation.

The performance of each of the clinical scenarios was assessed by the power to reject at least one false hypothesis, based on repeated simulations of Normally distributed standardized treatment effects. This is the key *metric* of interest. Two thousand simulations were used for each of clinical scenarios described; this was chosen mainly for practical reasons, based on the available computing resources.

4 Example

In this section a number of *clinical scenarios* are selected from the available matrix of options and assumptions in order to illustrate a range of potential trial designs. The power to reject at least one false hypothesis is plotted against the total trial size (between 750 and 2000 patients in total) for the following set of assumptions/options

(i) four experimental treatments (plus a single control)
(ii) a standardized treatment effect size of $\Delta = 0.25$
(iii) a correlation between the early and late outcome measures of $\rho = 0.1$
(iv) a ratio between the early and the late outcome measures (F) of 1.25
(v) a treatment effect type scenario where only one of the experimental treatments is effective

Three selection rules are compared, for designs that allocate patients in the ratio 1:1 and the ratio 1:3, for the two stages of the trial. The three chosen selection rules are (i) a fixed rule selecting one test treatment only at the interim, (ii) a variable selection rule where $\varepsilon = 0.15$ and (iii) a futility selection rule determined by $F = 1.25$.

Plots of power to reject at least one false hypothesis are shown in Figure 2 for these clinical scenarios.

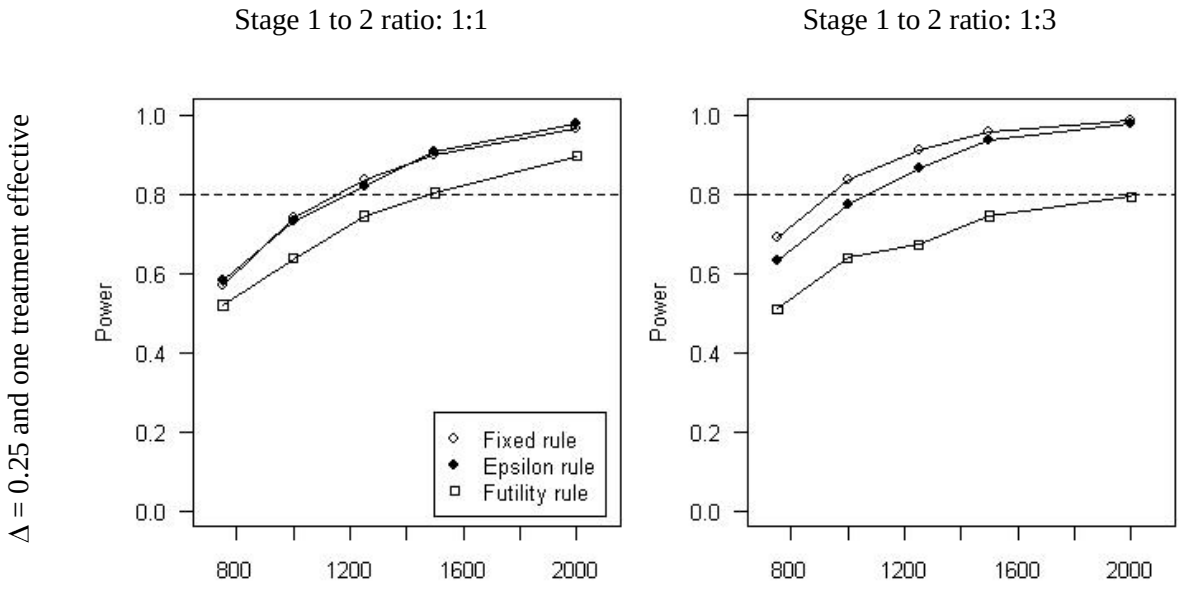

Fig. 2. Power to reject at least one false hypothesis against total trial sample size (patients) for **four** test treatments, for a **fixed selection rule** (select 1), a **variable selection rule** ($\varepsilon = 0.15$) and a **futility rule**

5 Conclusions

The aim of this work has been to describe an application of adaptive designs for a clinical trial in MS. The methodology has been presented briefly, together with an explanation of how simulations, based on a realistic model constructed from analysis of data from previous trials in MS, relevant literature and expert opinion, can be used prior to the commencement of the actual clinical trial to ensure that treatments will be dropped appropriately. Much of the challenge of implementation of a treatment selection design as described above arises exactly because of its flexibility. It is important to consider carefully the range of *options* and *assumptions* likely to arise and then to use these in a range of simulation evaluations. Our work has brought together statisticians and clinicians, all of whom, as a result of the process described in this paper, now have a greater understanding of the important factors impacting the design of trials in this therapeutic area. However, it is only when definitive statements can be made on the number of test treatments and the likely treatment effect types that a more definitive statement indicating the optimal settings of the design parameters for a future study be made.

Acknowledgements

This work was supported by a grant from the Multiple Sclerosis (MS) Society of Great Britain and Northern Ireland.

References

Bauer, P. and Kieser M. (1999). Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine*, **18**, 1833-1848.

Bauer, P. and Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics*, **50,** 1029-1041.

Benda N, Branson M, Maurer W, Friede T. (2009). Aspects of modernizing drug development using clinical scenario planning and evaluation. Submitted.

Bretz F, Schmidli H, Koenig F, Racine A, Maurer W. (2006). Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: General concepts. *Biometrical Journal*, **48**, 623-634.

Committee for Medicinal Products for Human Use (CHMP). Guideline on clinical investigation of medicinal products for the treatment of multiple sclerosis. 2005, London. Doc. Ref. EMEA/CHMP/EWP/561/98 Rev 1.

Chow, S.-C. and Liu, J.-P. (2004). Design and Analysis of Clinical Trials: Concepts and Methodologies. Wiley, Hoboken. Second Edition.

Dunnett, C.W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, **50**, 1096-1121.

Koenig F, Brannath W, Bretz F, Posch M (2008) Adaptive Dunnett tests for treatment selection. *Statistics in Medicine* **27**, 1612–1625.

Lehmacher, W. and Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics*, **55**, 1286-1290.

Maca J., Bhattacharya S., Dragalin V., Gallo P., Krams M. (2006) Adaptive seamless phase II/III designs: Background, operational aspects, and examples. *Drug Information Journal*, **40**, 463-473.

Marcus, R., Peritz, E., Gabriel, K.R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, **63**, 655-660.