UNIVERSITY OF LONDON THESIS

Degree PhD Year 2006 Name of Author BARTHEL, Friedrike Maria-Sophie

Issues in the design and analysis of clinical trials with time-to-event outcomes

This dissertation is submitted for the degree of
PhD

by

Friederike Maria-Sophie Barthel

February, 2006

University College London &
MRC Clinical Trials Unit

I, Friederike Maria-Sophie Barthel, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Issues in the design and analysis of clinical trials with time-to-event outcomes

by

Friederike M.-S. Barthel

Submitted to the University College London
on February 16, 2006, in partial fulfillment of the
requirements for the degree of
PhD

Abstract

Designers of clinical trials today face a number of different challenges. In a number of disease areas several treatments become available at any one time with only a limited number of patients available for investigation purposes. Furthermore, in disease areas such as HIV and cancer the pressure is high to find effective treatments quickly. Due to the recent advances in the human genome project more and more interaction effects between a treatment under study and the genetic make-up of a patient may be successfully analysed.

This thesis aims to evaluate existing tools for the design and analysis of clinical trials with a time-to-event outcome and provide extensions in areas where existing tools do not perform satisfactorily. Particular emphasis is placed on sample size calculations for multi-arm and multi-stage trials and other complex mechanisms such as loss to follow-up and patient withdrawal from allocated treatment. Furthermore, advances are made in the area of treatment-covariate interactions, particularly in terms of analysis tools for such interactions.

Thesis Supervisor: Patrick Royston
Title: Professor

Thesis Supervisor: Andrew Copas
Title: PhD

Contents

List of Figures

List of Tables

Acknowledgement 1 *I would like to thank the following people for lending their support and inspiration during the completion of this work:*

Patrick Royston for his supervision and advice, as well as for his patience especially in the early stages of my work. Max Parmar, my other supervisor at the MRC CTU, for providing many fruitful ideas for this research.

I would also like to thank Andrew Copas, my second supervisor at UCL, for taking over the task from Stephen Senn and for giving guidance on the organisational side of the PhD. Furthermore Abdel Babiker for his invaluable support with the mathematics regarding the sample size calculations.

Last but not least I would like to express my sincere appreciation to my husband and

my family for their wonderful co-operation, understanding and support throughout the period of this research.

Chapter 1

Introduction

1.1 The context of the research

Clinical trials are scientific investigations that examine and evaluate safety and efficacy of therapies in human subjects. Trials therefore carry a responsibility both to ensure the welfare of their participants and to be publicly accountable. If trials are to be successful and achieve their aim of improving healthcare for the public at large, then they need to be practical and relevant. This calls for trial designs which allow us to answer the right questions as quickly and efficiently as possible.

In the last 20 years there has been a major increase in our basic understanding of many diseases based on a revolution in molecular sciences. This has inevitably fuelled great hope in our potential to cure many serious diseases, such as cancer, HIV and heart disease. However, in a report in March 2004 the US Food and Drug Administration have identified a slowdown, rather than expected acceleration, in innovative medical therapies reaching patients [39]. As a consequence there is increasing concern that the hoped-for advances in improving survival and quality of life in many major diseases may not materialise.

Two factors are highlighted as being involved in this downturn including the high cost of bringing a new product to the market, estimated to be of the order of one billion US dollars, and the fact that most new treatments are not effective. The FDA have estimated that only approximately 8% of therapies entering Phase I trials reach the market.

This has happened despite the fact that in the last 10 years biomedical research spending has more than doubled in real terms in the private sector internationally and in the public sector in the USA. There have also been corresponding increases in research spending in the public sector in many countries in Europe. The FDA, in their report, emphasise the need for new approaches to reject ineffective therapies and continue testing the promising ones as rapidly and as reliably as possible. In this thesis we present some new methods that aim to achieve this goal.

1.2 Organisation and overview

This thesis is essentially divided into four main parts. The first part (Chapters 2 to 4) deals with the issues surrounding sample size for trials with survival type endpoints and extensions to non-uniform survival, multiple arms, loss to follow-up, non-proportional hazards and cross-over. The second part (Chapters 5 to 7) concentrates on multi-stage, multi-arm trials with intermediate endpoints. In Chapter 8, the third part, the impact of the variability in accruing events on the total trial time is examined both in a standard parallel group trial setting as well as in the multi-stage, multi-arm trials introduced in Chapters 5 to 7. Finally, we concentrate on the analysis of treatment-covariate interactions in the fourth part of this thesis (Chapter 9).

The results of many randomised clinical trials are inconclusive, often because insufficient numbers of patients were included. In Chapter 2 we consider the need to estimate sample sizes realistically with particular emphasis on aspects which may reduce the power of a trial in time-to-event situations. In Section 2.2 the development of sample size formulae for these types of trials over the years is examined. Extensions to more than one experimental arm in comparison with a control are described in Section 2.3 while Section 2.4 introduces more complex censoring situations.

We present a general framework for sample size calculation in survival studies based on comparing two or more survival distributions using any one of a class of tests including the logrank test in Chapter 3. The fundamentals of this method originated from work done by Professor A. Babiker. Incorporated within the method are the possible presence of non-uniform staggered patient entry, non-proportional hazards, loss to follow-up and treatment changes including cross-over between treatment arms which are discussed in Section 3.3. Further extensions to the methodology such as non-local alternatives for the logrank test are also considered. Their validity is explored using simulation studies in Section 3.4.

The sample size framework described in Chapter 3 has been implemented in the freely available program ART (Analysis of Resources for Trials) for Stata which is discussed in Chapter 4. Our investigations suggest that ART is the first software to allow incorporation of all these elements. Characteristics of ART and other sample size programs available to the public are compared in Section 4.4.

In phase II / III cancer trials, it is undesirable to stop a study early when the test treatment is promising. On the other hand, it is desirable to stop the study as early as possible when the test treatment is not effective or only likely to be minimally effective. Consequently, we propose a multi-stage design to determine at particular points during the trial whether a study drug holds sufficient promise to warrant further testing. In addition, it may not always be appropriate or possible for a randomised trial of a new treatment to be conducted on the clinical endpoint of primary interest. As a consequence, replacing the clinical endpoint of primary interest, such as overall survival, with a surrogate variable, which can be measured earlier, more frequently, easier and with lower costs, has been frequently advocated [31] [91] [29] [111] [136] [59] [23]. The lively and sometimes adversarial debate surrounding the use of surrogate markers is reflected in Chapter 5. Section 5.2.2 outlines the often cited Prentice criterion for a surrogate marker and the discussion surrounding its use. Further approaches to the validation of surrogate markers are examined in Sections 5.2.3 and 5.2.4. The second part of the chapter concentrates on multiple stage designs, starting from the early literature concerning sequential designs in Section 5.3.1.

Through a series of empirical illustrations and discussions Chapter 6 formulates our approach to combining intermediate markers, which do not have to fulfill the stringent criteria of the Prentice criterion, and a multi-arm, multi-stage selection design. This provides an extension to the two-stage design proposed by Royston et al. [103]. The main aims of this design are to quickly reject any new therapies unlikely to provide an advantage over control in the primary outcome measure as early and reliably as possible, while continuing with those therapies which are likely to provide an advantage over control on this measure. Section 6.2 deals with the extension of the design to more than two stages with consideration of the calculation of overall power and significance level for the trial. Necessary changes to the correlation structure are considered in Section 6.2.7. Assumptions underlying the design are examined in Section 6.3. Section 6.4 provides two actual trial examples employing the extension to more than two stages in cancer, one of which has just started patient accrual.

In Chapter 7 we discuss the performance of the design, and in particular its implementation in Stata, using simulation studies. For this purpose, the literature surrounding bivariate exponential distributions is surveyed in Section 7.2 and a new bivariate exponential distribution based on the bivariate normal distribution is proposed. The assessment of the robustness of the design also covers the occurrence of 'shocks' to the design in Section 7.4, such as the mis-specification of key parameters at the planning stage.

Chapter 8 seeks to explore possible strategies to preempt the inherent variability in trial time and / or the number of events. Such variability, especially in the case of trial time, has a direct impact on the time at which the primary analysis can be carried out and as a consequence may have an impact on the overall cost of the trial. Furthermore, the variability in the length of the first stage in a two-stage trial is important for the viability of the design as outlined in Chapter 6. We provide tools to assess the variability at the beginning of the trial in Section 8.3 as well as update these estimates throughout patient and event accrual in Section 8.4. A Stata tool is available which implements these methods.

The objective of a statistical interaction investigation is to assess whether the joint contribution of two or more factors is the same as the sum of the contributions from each factor when considered alone. An interaction test can be used to investigate whether the effectiveness of treatment is homogenous across groups of patients with different characteristics [123]. It is therefore important in the interpretation and inference of trial results. Interaction tests are introduced in Section 9.2 of Chapter 9. The following Section 9.3 illustrates the analysis of interaction using two substudies from cancer trials. A new Stata tool to aid the analysis and interpretation of treatment-covariate interactions is presented in Section 9.4.

Our conclusions are presented in Chapter 10.

Chapter 2

Sample size calculations for trials with time-to-event outcomes - a review

2.1 Motivation

Many researchers reach the end of their study to find out that they cannot make the conclusions with the reliability that they were hoping to, because their study did not have enough "power". This is not a simple problem to fix, but it is a simple problem to avoid. The power of a study is the ability of a study to demonstrate the targeted difference if in fact it does exist. The frequency of the event being studied, the size of the effect or the difference that is to be detected, the design of the study, and the sample size all affect the power of a study. The magnitude of this power will also depend on the choice of test used to analyse the data. Sample size is the easiest of these factors to modify. Thus, to avoid the disappointment of findings that one cannot draw conclusions from, sample size calculations must be performed at the design stage of any study.

In this chapter we are exploring sample size calculations in particular for survival type studies and their extensions to include more than two treatment arms as well as particular censoring situations. Parameters that underlie every one of those calculations are the power, the level of significance (Type I error rate), the underlying event rate and the size of the treatment effect sought [67]. Without taking account of the particular study setting, i.e. cohort, case-control, clinical trial, or the outcome measure used, we can represent the underlying structure of sample size calculations using a flow-chart as illustrated in Figure 2-1.

2.2 Sample size formulae and their development over the years

Since the 1960s many papers have been published on the subject of sample size calculations for clinical trials. The idea of achieving maximum power of tests with the minimum sample size possible has remained central over the years. Due to the huge variety of possible calculations, several books and papers, such as 'Sample size tables for clinical studies' by Machin & Campbell [81] have attempted to bring some of these together. This section mainly concentrates on sample size calculations for survival analysis, however, we will also refer to some of the other developments.

One of the earlier sample size tables was published by Halperin et al. [55] and is based on the sample size

$$N = \frac{4\{z_{1-\alpha}\surd[2\overline{p}(1-\overline{p})] + z_{1-\beta}\surd[p_E(1-p_E)+p_C(1-p_C)]\}^2}{(p_C-p_E)^2} \tag{2.1}$$

where p_C and p_E are the anticipated T-year cumulative event rates in the control and experimental group respectively, $\overline{p} = \frac{1}{2}(p_E + p_C)$ and $z_{1-\alpha}$ and z_β are normal deviates corresponding to a one-sided significance level α and power $1-\beta$. The assumptions are that i) there is no loss to follow up and ii) no non-event deaths occur. The event times in each treatment group follow an exponential model and the event rate for the control group is based on earlier studies.

George & Desu [50] consider a comparison of the number of patients required derived under an exact distribution of the test statistic and a normal approximation when the time-to-event is being studied. One of the main differences to later papers is that whilst survival times are assumed to be exponential, accrual is based on the Poisson distribution instead of the uniform. Again, no censoring is assumed to occur. Using simulations, the authors have found that sample sizes based on the normal approximation, which is based on the logarithmic transformation, is accurate. This is given by

$$N = \frac{4(z_{1-\alpha}+z_{1-\beta})^2}{\ln^2\Delta} \tag{2.2}$$

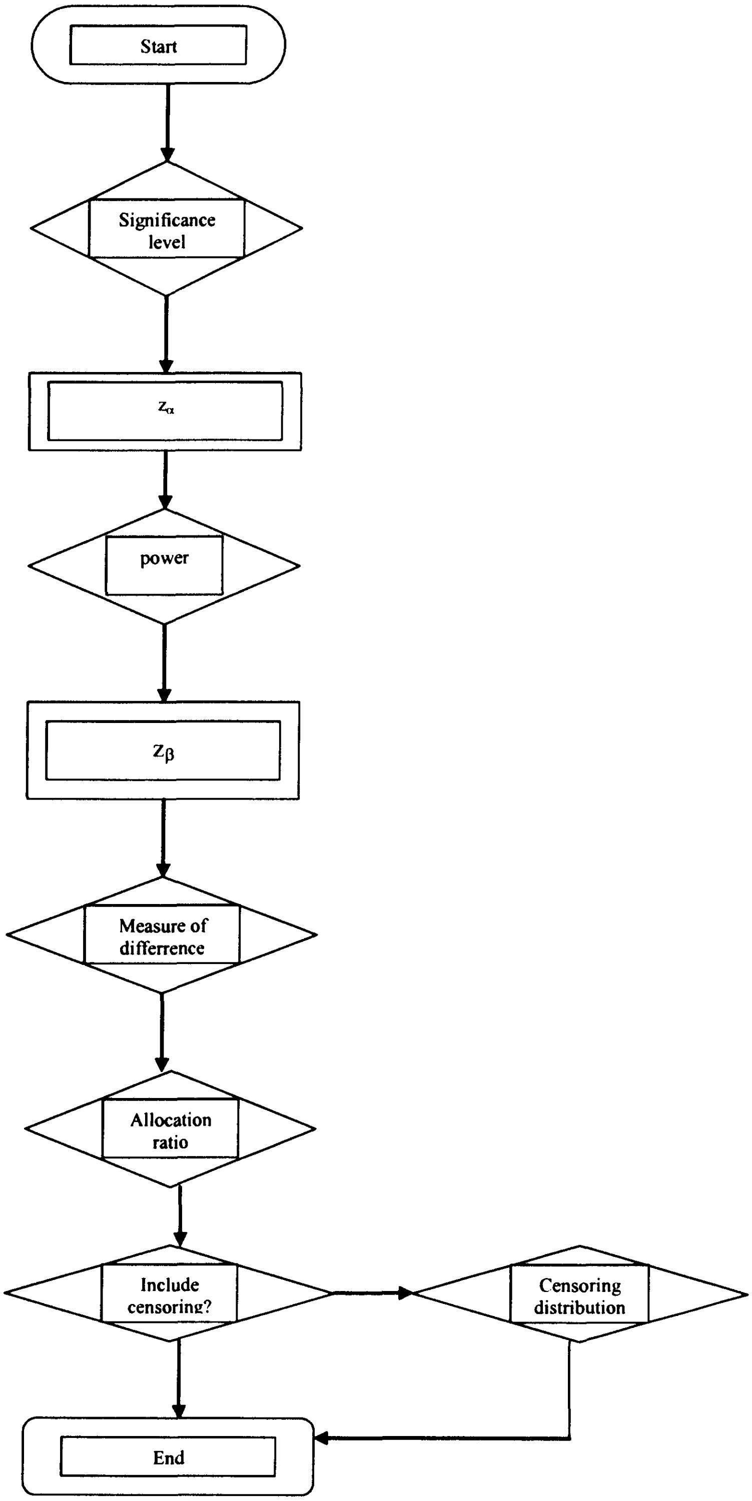


Figure 2-1: Flow chart to illustrate generic sample size structure

where $z_{1-\alpha}$ and z_β are defined as above and Δ denotes the hazard ratio in favour of the experimental group.

Schoenfeld [113] and Freedman [41] were among the first to propose sample size formulae for comparing two survival distributions using the logrank test while taking into account administrative censoring. Their formulae are based on the asymptotic expectation and variance of the logrank statistic. Between 1981 and 1983 Schoenfeld et al. [114] [113] [115] published three papers on sample size calculations and nomograms based on the logrank test. These are centered around the formula

$$N = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{(\ln^2 \Delta)\psi p(1-p)} \tag{2.3}$$

where $z_{1-\alpha}$, z_β and Δ are defined as above, ψ denotes the probability of not being censored by the end of the trial and p gives the proportion of patients allocated to the control treatment group. This formula assumes proportional hazards. The main difference between the two approaches by Freedman and Schoenfeld is that the formulation by Schoenfeld results in slightly lower estimates for the number of patients needed. In addition, Schoenfeld takes account of the presence of administrative censoring which occurs due to some patients not having experienced an event by the end of the study. Freedman encourages that modifications should be made allowing for withdrawal rates but does not consider the effect on sample sizes. He suggests to use

$$N = \frac{(1+\phi)(z_{1-\alpha} + z_{1-\beta})^2(1+\phi\Delta)^2}{\phi^2(1-\Delta)^2[(1-p_C)+(1-p_E)]} \tag{2.4}$$

where p_C and p_E are again the survival rates in the two groups and ϕ is the allocation ratio of patients to the control and experimental group.

Gail [45] considered the relative efficiency of using a test of proportions as opposed to the logrank test for sample size calculations with survival outcomes. Assumptions he made in this assessment were no withdrawals, local proportional hazards and uniform accrual rates. He found that in situations such as cardiovascular disease trials where treatment was administered relatively quickly in comparison to survival times a sample size calculation based on a test of proportions resulted in an efficiency close to one. However, in the case of cancer or other trials where the accrual period exceeds mean survival time, a proportions based sample size calculation led to a 39% larger sample size requirement compared to the logrank based requirements. At the same time the efficiency of the proportions test can drop to 72% or less in such a setting. In conclusion, power calculations specifically tailored to the logrank test should be used for studies with a total duration comparable to the mean survival time, if one intends to employ the logrank statistic for analysis purposes.

A comprehensive review of the sample size literature centering around testing the difference in proportions for two sample trial designs was published by Sahai & Khurshid [106]. They give formulae for various conditional and unconditional tests starting with Fisher's exact test. In addition to that they offer practical advice on how to use the formulae for clinical readers. Due to its volume readers are referred to their paper for further details of the sample size requirements.

The papers cited so far which have concentrated on survival outcomes have assumed an exponential survival distribution. Heo et al. [58], however, considered the Weibull model in their paper which would be more appropriate in the case of ageing research. This is due to the fact that the assumption of constant hazards breaks down when the follow-up time is long relatively to the life span of the study subjects and therefore the Weibull assumption would make the model more flexible. Calculations for the sample size are closely based on the Schoenfeld derivations [114] under the logrank test. The required sample size is then given by

$$N = \frac{(z_{1-\alpha} + z_{1-\beta})^2}{(k \ln \Delta)^2 \psi p(1-p)} \qquad (2.5)$$

for parameters defined as before and k denoting the shape parameter of the Weibull distribution. Hence in the case of k=1 we arrive at the exponential case and therefore Schoenfeld's formula 2.3.

2.3 Extensions to more than two treatment arms

Many trials today evaluate more than one experimental treatment group against standard therapy, as several promising treatment regimens become available at the same time and limited patient numbers are an issue. Makuch & Simon [82] have noted that the heuristic use of sample size formulae for two groups is inadequate in these cases. If we multiply the formula designed for two treatment arms by the required number of experimental arms we do not take into account multiple comparisons made in the analysis of the trial. A possible simple modification would be to take account of the multiple comparisons in the significance level used. Since the late 1980s a few authors have been suggesting different ways of approaching this problem.

Day & Graham [27] published a nomogram approach for trials in which the main method of analysis is ANOVA. They developed a new difference parameter, defined as the standard deviation of the expected treatment group means divided by the standard deviation of the measurements. This approach is quite simple to use since the nomograms are arranged as straight lines which makes reading off the required sample size much easier than in previous publications. Examples are given in the comparison of two, three and four groups and also in the case of a factorial design. In addition, reference is made to ordinal responses. However, a drawback of this method is that it only allows for equal group sizes. There are scenarios though when unequal allocation ratios may be beneficial. One example of this is the case where drugs are investigated for which only a small amount of prior information is available. In this case it is desireable to use unequal randomisation in order to find out more about the treatment. Another reason may be the very high cost of a particular treatment arm. If there is only a fixed sample size available for recruitment, unequal randomisation may confer large financial savings with limited impact on power.

For the case of survival endpoints both Ahnn & Anderson [1] and Liu & Dahlberg [77] have generalised an approach by Makuch & Simon. Both papers hence extend the logrank test sample size calculations derived by Schoenfeld [114] to the case of k treatment arms. Furthermore, Ahnn et al. consider the case of dose-response settings using Tarone's trend test and a stratified sample size calculation. Whilst Ahnn et al. base their work around the fact that the test statistic has a non-central chi-squared statistic, Liu & Dahlberg take the route of Fisher's least significant difference. The findings by Liu & Dahlberg suggest that their sample size derivation has the most power in the case where all treatment arms are at least as good or better than the control arm. If the hazards are more evenly spread between the treatment arms and the control group, then the power decreases with the sample size being inadequate to detect such differences. Simulations run by Ahnn & Anderson show that their sample size is fairly accurate for the case of three arms even if the alternatives are far from the null hypothesis. A further generalisation to unequal allocations was derived by Halabi & Singh [54].

2.3.1 Global comparisons of treatment arms

All of the above sample size calculations for more than two treatment arms as well as those provided by Barthel et al. [7], which are described in more detail in the next chapter, were derived for a global alternative hypothesis. This means that in a trial setting in which patients are randomised to one of K treatment groups, labelled $k = 1, 2, ..., K$, and $\lambda_k(t)$ is the hazard function in treatment group k $(k = 1, ..., K)$, the null hypothesis of equality of the K survival distributions can be expressed as $H_0 : \lambda_1(t) = \lambda_2(t) = ... = \lambda_K(t)$. The global alternative hypothesis $H_1 : \lambda_k(t) \neq \lambda_l(t)$ for at least one $k \neq l$ $(1 \leq k, l \leq K)$ states that for at least one pair of study arms the hazards are different at time t.

Such a global alternative hypothesis is used in a variety of trial settings. PACES [96] was a patient preference trial conducted in the USA which considered a comparison of placebo, acetaminophen (paracetamol) or celecoxib. Patients were randomised to each of the treatment arms and then crossed over to a different arm after a period of 6 weeks. At the end of treatment, patients were then queried about their preference between the two treatment periods. The trial also assessed efficacy using the WOMAC score. Sample size was calculated based on a global comparison of the treatment arms. Subsequently, pairwise comparisons were also conducted and reported due to the significant result in the global comparison. Thus the global analysis served as a trigger for any other comparisons which would only be conducted if the global result was positive. Wolmark et al. [144] also considered three treatment arms (in this case all active agents) for the treatment of Dukes' B and C Carcinoma of the Colon. In this case the trial was powered on pairwise comparisons of the treatment arms but a global comparison was reported. Both the pairwise and global comparisons were not significant. Sample size for a pairwise comparison of more than two treatment arms may be calculated using any of the methods in Section 2.3 above. We then need to account for the fact that more than one pairwise comparison is conducted by using a reduced nominal significance level through, for example, a Bonferroni adjustment on the heuristic approach of using the formula for the two treatment group situation and multiplying the required number of patients per treatment group by the number of groups to be compared. The resulting sample size will be more conservative than that for the global comparison.

Another setting in which a global comparison of more than two treatment arms is often used is that of a comparison of several doses of the same treatment. In a briefing paper the FDA reports efficacy and safety of three TNF blocking agents [38]. Efficacy for the different agents was assessed for different doses and results from a global comparison are reported. When doses are compared there is usually an intrinsic order in the treatments which can then be expressed in an ordered global alternative hypothesis, i.e. $H_1 : \lambda_1(t) \leq \lambda_2(t) \leq ... \leq \lambda_K(t)$. A modified ordered logrank test as well as sample size requirements are provided by Liu et al. [79]. Resulting sample size requirements will be less conservative than those given by Ahnn & Anderson [1]. Further discussion of the different testing strategies in multi-dose experiments is provided by Bauer et al. [9].

2.4 Treatment of more complex censoring situations

So far the sample size calculations cited have either not taken account of censoring at all or have included right censoring at the end of the follow-up period only. In practice, clinical trials pose far more complex censoring situations such as loss to follow-up, non-compliance and lag times. We define a patient as lost to follow-up if he/she does no longer provide trial data after randomisation due to circumstances such as moving to a different area. In contrast to this a patient is labelled as being non-compliant if he/she remains available for follow-up but no longer adheres to the treatment regimen he/she was randomised to at the beginning of the trial. In addition, in some trials the proportional hazards assumption may break down and different models have been suggested in order to take that into account in the sample size calculations.

Schork & Remington [116] suggested that loss to follow-up and non-adherence should be taken into account when determining the sample size. In their paper they consider a trial with a single treatment control comparison, a binary outcome variable and a relatively long period of observation. They examine the impact on sample size using subject shifting patterns between the treatment group and the control and demonstrate that the sample size can be expressed as a function of the frequency with which these patterns occur. For the case of loss to follow-up they suggest the estimation of the expected proportion at the beginning of the study which should then be added to the total sample size.

Lachin & Foulkes [71] extended an earlier sample size approach by Lachin [70] to non-uniform entry, loss to follow-up, failure to comply with treatment and stratified analyses. They furthermore suggested that whenever sample size calculations are employed, these should take account of the worst case scenario in terms of the hazard ratio and censoring. Non-uniform entry is based on a concave entry distribution (lower rate of intake than expected) such as the truncated exponential. They found that in this case a substantial increase in sample size is required to compensate for a small reduction in power. For the case of loss to follow-up they use exponentially distributed loss to follow-up hazard rates which are independent of those for mortality. Findings suggest that the effect on sample size is roughly proportional to the addition of these. Noncompliance was considered for the case where patients stop taking treatment as required but are not lost to follow-up. The assumption here is that patients who are non-compliant in one group will then be subject to the hazard ratio in the other group for the entire study. This leads to slightly conservative estimates.

Further extensions were provided by Yateman & Skene [146] who modelled patient entry as a piecewise linear function as an alternative to uniform entry. In addition they modelled survival and loss to follow-up distributions using piecewise exponential distributions.

The use of discrete Markov chains for modelling censoring was suggested by Lakatos [72] [73] in two papers. He proposes a method which takes account of the lag in the effectiveness of medication and one which takes account of non-proportionality of the hazards. Whilst the first paper only considers a binomial model, the second also provides extensions to the logrank test and the Tarone-Ware class of statistics. Markov Chains are modelled as follows: In order to assign probabilities to the transition matrices a step function with a jump at the end of each year is used. However, this can be modelled to include jumps at quarterly rates or the like. Accrual is modelled so that all patients are assumed to enter the trial at the beginning and are then administratively censored in accordance with accrual rates. When considering non-compliers, a decision needs to be made about how to treat these in the trial. Considerations to be taken into account are an analysis based on intention-to-treat, which would mean that non-compliers are not to be considered as censored, and whether one allows non-compliers to reenter treatment. Comparisons were also made between the proportional hazards and two types of lag models. A computer program based on these methods was suggested by Shih [121]. In addition she introduced prior distributions to express the uncertainties surrounding the parameters in the model.

Further extensions of this model to more than two treatment groups were made by Ahnn & Anderson [2] and a combination with their earlier approach for more than two treatment groups was sought. They show that this model can be especially useful where unexpected events occur during the course of a trial, such as an advanced stage cancer trial where noncompliance takes place due to unexpected toxicity.

2.5 Conclusions

Over the course of this chapter we have presented developments in sample size calculations in particular for studies with a survival type outcome. All of these fit into the framework set out in the flow-chart in Figure 2-1. In particular, as the next chapter will show, patients lost to follow-up or not adhering to the allocated treatment, e.g. crossing over to receive the regimen of the other treatment group, while still being analysed under intention-to-treat, can have a significant effect on the power of a trial and hence allowance for this scenario should be made in sample size calculations.

Chapters 3 and 4 will introduce the Stata program ART (Analysis of Resources for Trials) which incorporates all of the above sample size issues.

Chapter 3

Evaluation of sample size and power for multi-arm survival trials allowing for non-uniform accrual, non-proportional hazards, loss to follow-up and cross-over

3.1 Introduction

The logrank test is probably the most commonly used tool for designing and analysing clinical trials with a survival time outcome. The planning of such a trial must take into account not only the proposed method of analysis but also circumstances not usually encountered in other types of experiments. Patient accrual into a trial is staggered which means that patients enter the trial sequentially over an accrual period. Also, when complete, it is commonly followed by a fixed period during which patients are under observation for events but no new patients are entered. Further, administrative censoring occurs with some patients not experiencing an event by the time the trial ends. Usually, the statistical analysis of such a trial will consist of a test of the null hypothesis that there is no difference in survival between the treatments at a given significance level and power.

We present a general framework for sample size calculation in survival studies based on comparing two or more survival distributions using any one of a class of tests including the logrank test. Incorporated within this framework are the possible presence of non-uniform staggered patient entry, non-proportional hazards, loss to follow-up and treatment changes including cross-over between treatment arms. The framework is very general in nature and is based on using piecewise exponential distributions to model the survival distributions. We illustrate the use of the approach and explore its validity using simulation studies. These studies have shown that not adjusting for loss to follow-up, non-proportional hazards or cross-over can lead to significant alterations in power or equivalently, a marked effect on sample size. The approach has been implemented in the freely available program ART (for Stata). Our investigations suggest that ART is the first software to allow incorporation of all these elements. Further extensions to the methodology such as non-local alternatives for the logrank test are also considered.

In Section 3.2 we provide an outline of the underlying multi-arm sample size method used in this chapter. Section 3.3.1 illustrates our approach to staggered entry and loss to follow-up based on piecewise exponential distributions which also allows for non-proportional hazards. We propose an incorporation of treatment changes in Section 3.3.2. The performance of the method incorporating all these elements is demonstrated using simulation results in Section 3.4 and trial examples in Section 3.5. A discussion is provided in Section 3.6.

3.2 Multi-arm trials

Consider a trial setting in which a population of N patients are randomised to one of K treatment groups, labelled $k = 1, 2, ..., K$, and the K treatments are to be compared globally in terms of time to failure using a (weighted) logrank test. If $\lambda_k(t)$ is the hazard function in treatment group k ($k = 1, ..., K$), then the null hypothesis of equality of the K survival distributions can be expressed as $H_0 : \lambda_1(t) = \lambda_2(t) = ... = \lambda_K(t)$. The global alternative hypothesis $H_1 : \lambda_k(t) \neq \lambda_l(t)$ for at least one $k \neq l$ ($1 \leq k, l \leq K$) means that for at least one pair of study arms the hazards are different at time t. Let $\Delta_k(t)$ be the log hazard ratio in group k relative to group 1, that is $\Delta_k(t) = \log[\lambda_k(t)/\lambda_1(t)]$ ($k = 2, \ldots, K$). Further, let $\Delta(t) = (\Delta_2(t), ..., \Delta_k(t))'$. For the remainder of this section assume that $\Delta \equiv \Delta(t)$.

The logrank test is based on a comparison between the observed and expected numbers of events under H_0. Let $t_1 < ... < t_m$ be the distinct failure times, such as deaths or disease progressions, and assume no ties. Let O_k^j be the observed number of events in group k at time t_j ($j = 1, ..., m$). Let $r_k(t_j)$ be the number of patients at risk in group k at time t_j. The expected number of events $e_k(t_j; .)$ in group k at time t_j depends on the event history and on whether H_0 or the more general H_1 is assumed. Under H_0 we have simply

$$e_k(t_j; 0) = \frac{r_k(t_j)}{\sum_{l=1}^{K} r_l(t_j)}$$

[89] whereas under H_1,

$$e_k(t_j; \Delta) = \frac{r_k(t_j) \exp(\Delta_k(t_j))}{\sum_{l=1}^{K} r_l(t_j) \exp(\Delta_l(t_j))}$$

[24]. For comparing group k with group 1, the logrank test is based on the distribution under H_0 of the observed minus the expected number of events, that is on

$$U_k = \sum_{j=1}^{m} W(t_j)[O_k^j - e_k(t_j; 0)]$$

where $W(t_j)$ is a weight function [24]. The standard logrank test has $W(t_j) = 1$. Weights according to Tarone & Ware [130] and Harrington & Fleming [56] may be found in Appendix B.

The global logrank test statistic Q is based on the vector $U = (U_2, ..., U_K)'$ and is defined as the quadratic form

$$Q = U'V(0)^{-1}U$$

where $V(0)$ is the covariance matrix of U under H_0 (see Expression B.1 in Appendix B). Since U is asymptotically distributed as multivariate Normal $\mathrm{N}(0, V(0))$ under H_0 the distribution of Q is central χ^2 on $K - 1$ degrees of freedom [22].

To derive the sample size we consider a sequence of local alternatives to the null hypothesis, i.e. that $\Delta_k(t)$ is of the order $O(N^{-1/2})$ [113]. Thus a higher sample size will be required the closer $\Delta_k(t)$ is to 1. The resulting formula performs best under hazard ratios which are not too far from one, e.g. for hazard ratios around 0.6 - 1.67. Under local alternatives Q approximately follows a non-central chi-squared distribution on $K - 1$ degrees of freedom [22] with non-centrality parameter

$$\tau = NM'V(0)^{-1}M \tag{3.1}$$

where

$$M = (M_2(\Delta), ..., M_K(\Delta))$$

and

$$M_k(\Delta) = \frac{1}{\sqrt{N}} E(U_k|H_1)$$

Further details on the calculation of M and $E(U_k|H_1)$ may be obtained in Appendix B. The value of the non-centrality parameter τ needs to be obtained for a given power $1 - \beta$ and significance level α either from the chi-squared tables provided by Hayman et al. [57] or a statistical package. The required sample size is then obtained by solving Equation 3.1 for N replacing $V(0)$ and M by their asymptotic values (see Appendix B). For the simple case of the logrank test under proportional hazards (Δ_k independent of t) and no treatment changes N is given by

$$N = \frac{K\tau}{\psi[\frac{K-1}{K}\sum_{k=2}^{K}(\Delta_k)^2 - \frac{2}{K}\sum_{k=2}^{K}\sum_{q=2_{k<q}}^{K} \Delta_k\Delta_q]} \tag{3.2}$$

whereby ψ is defined as the probability of not being censored by the end of the trial [1]. When evaluating Expression 3.2 for two treatment groups only we arrive at Schoenfeld's formula 2.3.

A better approximation of the distribution of Q for more distant alternatives is given in Appendix C.

3.3 Implementation

The framework underlying the calculations incorporating staggered patient entry, loss to follow-up, cross-over and non-proportional hazards requires the total trial time to be split into several periods. For tractability these are taken to be of equal length. Hence we can examine the number of patients at risk and the occurrence of events in all groups separately for each period. The length of each period may depend on the amount of knowledge available about patient characteristics at the planning stage of the trial. In some instances, for example, we may have a lot of information about the survival distribution in the control group in which case one month long periods are advantageous. Furthermore modelling the survival distributions over each of the periods allows us to take non-proportional hazards into account. This is not only important for overall survival which may have non-constant relative hazards due to, for example, time delays in the effect of treatment but also cross-over which may vary over the course of the trial. For instance, patients may change their treatment or drop out towards the end of a long study, particularly if the frequency of follow-up visits declines.

3.3.1 Staggered patient entry and loss to follow-up

We define T as the total number of periods in the trial, i.e. T is the sum of the number of periods of accrual and follow-up. Each of these periods is of equal length. Patients are accrued over the periods 1 to R where $R \leq T$. Define $F^R(t)$ as the cumulative distribution function of recruitment time. For example, $F^R(t)$ may be represented as a piecewise truncated exponential of the form given by Cox & Oakes (p. 178) [24] or a uniform distribution, depending on which type of entry mechanism is deemed to be more appropriate. The number of patients N is then accrued using an exponential or uniform process. Furthermore $F^R(t)$ is allowed to have a point mass $F^R(0)$ at zero, allowing one to specify a certain proportion of patients randomised before the start of the first period of the trial. This proportion may vary between 0 and 100% of the total number of patients. Figure 3-1 illustrates the accrual pattern in a trial consisting of five periods where accrual takes place during the first four periods only. Additionally, a proportion of patients has been recruited before the start of the trial. The accrual pattern itself is uniform during each of the periods, however, it is not constant over the whole course of the accrual period.

Under the derivation of the probability ψ of not being censored given by Schoenfeld [114] patient entry occurs over the accrual period resulting in administrative censoring times after completion of planned follow-up, i.e. at $t = T$. However, in most trials some patients are lost to follow-up due to other reasons. This means that the observed survival time for each patient will be the minimum of the time to event, time to loss to follow-up or time to termination of the trial. It is assumed that time to loss to follow-up is independent of survival times. We define $S_k^L(t)$ as the survivor function of time to loss to follow-up and $S_k^E(t)$ as the survivor function of failure times where $k = 1, 2, ..., K$. Assume that $S_k^E(t)$ has been adjusted for cross-over, i.e. treatment changes (see Section 3.3.2). Both $S_k^E(t)$ and $S_k^L(t)$ can be approximated by piecewise exponential distributions with hazards ε_{ki} and μ_{ki} (treatment k, period i) respectively where $t \in [0, T]$. Denote the probability density functions associated with $S_k^E(t)$ and $S_k^L(t)$ by $f_k^E(t)$ and $f_k^L(t)$

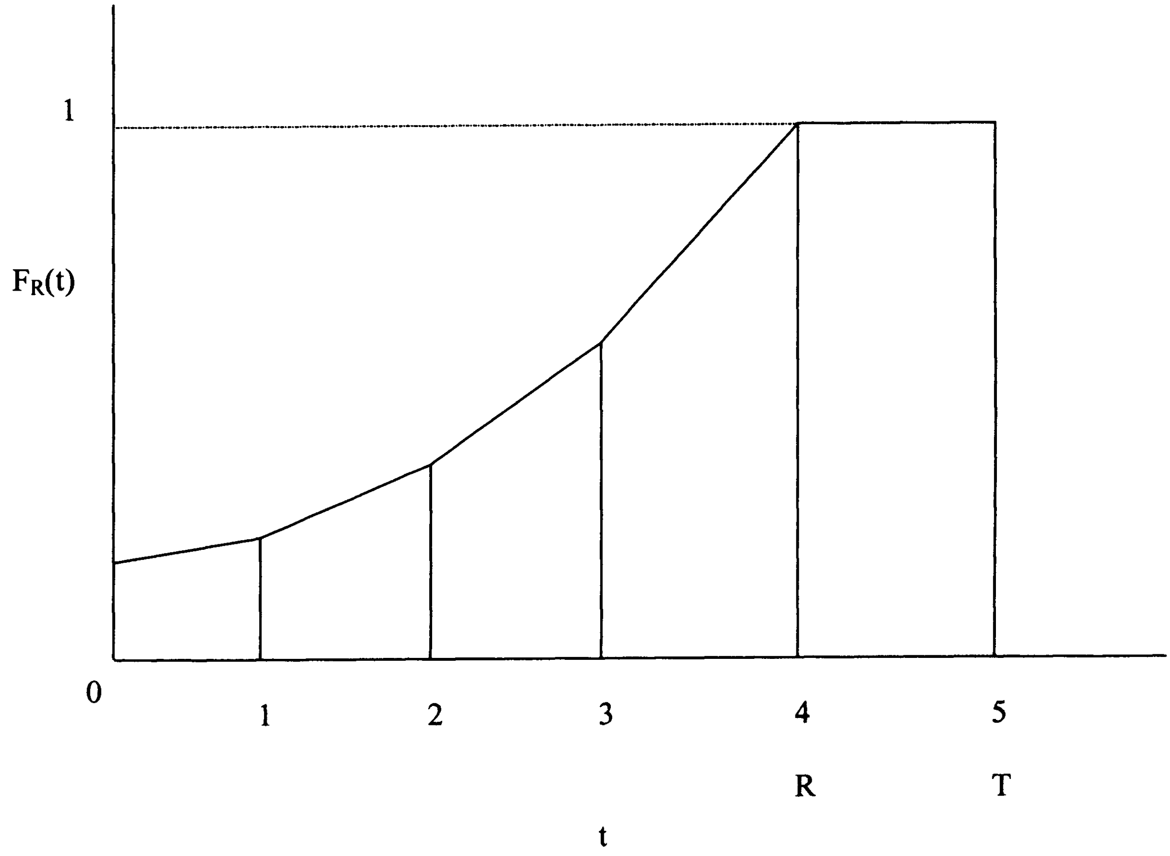


Figure 3-1: Cumulative distribution function of accrual over five periods of a trial with a point mass at zero

respectively. According to Yateman & Skene [146] the density for time to loss to follow-up is of the form

$$f_k^L(t) = \begin{cases} \mu_{k1} \exp\{-\mu_{k1} t\} & 0 < t \leq 1 \\ \mu_{ki} \exp\{\sum_{j=1}^{i-1} [j(\mu_{k,j+1} - \mu_{k,j})] - \mu_{k,i} t\} & i-1 < t \leq i, \; i = 2, ..., T \end{cases} \tag{3.3}$$

and the density for time to failure can be expressed in a similar manner.

Let s be the time at which a patient is accrued and $F^R(s)$ be the entry distribution function with properties as described above. We can express the distribution of time on the study or potential exposure time as $F^R(T-s)$ [71] [146] where $T - R < T - s < T$, i.e. $T - s$ is the administrative censoring time. In order to calculate the probability ψ of not being censored which is required to arrive at the sample size in Equation 3.2 we derive the proportion Π_k^E of patients experiencing an event in treatment group k over the duration of the trial. The probability that the event of a patient is observed is given by the integral of the probability that the event occurs at time t and the patient has not been lost prior to that time. These probabilities then need to be summed over all possible exposure times, that is

$$\Pi_k^E = \int_0^T F^R(T-s) S_k^L(s) f_k^E(s) ds \tag{3.4}$$

Hence the proportion of patients not censored by the end of the trial is given by $\psi = \sum_{k=1}^{K} p_k \Pi_k^E$ whereby p_k denotes the probability of being randomised to group k and $\sum_{k=1}^{K} p_k = 1$.

3.3.2 Cross-over

In our context we use cross-over to describe a patient who changes from the designated therapy regimen to that of another treatment group but remains available for follow-up. Analysis of the trial data under intention-to-treat is envisaged. This situation may arise in HIV or cancer trials where patients might, for example, change from a more intensive therapy to the therapy of the standard arm due to toxicity. Furthermore, we allow for patients changing to a treatment regimen not part of any of the treatment groups in the trial. In contrast to the method of Lakatos [73], as implemented by Shih in the SIZE program [121], patients crossing over from one treatment to another are not allowed to return to their original treatment in our derivation. This is a conservative assumption but it allows direct calculation of $S_k^E(t)$ adjusted for cross-over.

We calculate the distribution of time to failure adjusted for cross-over. This is necessary for the calculation of τ in Appendix B since there is no closed form for N under non-proportional hazards due to cross-over. Let t_E and t_C be times to failure and cross-over respectively, with corresponding survivor functions $S_k^E(t)$ and $S_k^C(t)$ and density functions $f_k^E(t)$ and $f_k^C(t)$ respectively. The hazard function of t_E if cross-over occurs at time t_C is

$$\begin{aligned} \lambda_{t_f}(t_E|t_C) &= \lambda_b(t) \ ; \ t < t_C \\ &= \lambda_a(t) \ ; \ t \geq t_C \end{aligned}$$

where $\lambda_a(t)$ and $\lambda_b(t)$ are the hazard functions for failure before and after cross-over respectively. Then,

$$\begin{aligned} S_k^E(t) &= \int_0^\infty P\{T \geq t\} f^C(t_C) dt_C \\ &= \int_0^\infty \exp[-\int_0^t \lambda_T(u|t_C)du] f^C(t_C)dt_C \\ &= S_0^E(t) S_k^C(t) + \int_0^t \exp[-\int_0^{t_C} \lambda_b(u)du - \int_{t_C}^t \lambda_a(u)du] f^C(t_C)dt_C \end{aligned}$$

where $S_0^E(t)$ is the survivor function in the absence of cross-over, i.e. $S_0^E(t) = P\{T \geq t | t_C = \infty\}$. The numerical evaluation of the above integrals is facilitated by the piecewise exponential assumption of the distributions of time to failure and cross-over.

3.4 Simulation results

To evaluate the performance of our method in terms of attaining pre-specified power, and in particular its implementation in our sample size program ART (Analysis of Resources for Trials) [101] [8] as described in more detail in Chapter 4, simulations were performed in Stata 8. Design specifications for all sets of simulations were two years of accrual, two years of follow-up, equal allocation to both treatment arms, uniform accrual, exponential survival and one year median survival in the control group. Furthermore, sample sizes were calculated for 90% power with a two-sided significance level $\alpha = 0.05$. In Tables 3.1

- 3.6 the simulated power is based on 5000 simulated trials which gives an approximate standard error of 0.4% and hence an approximate confidence interval around 90% power from 89.2% to 90.8%. All tables give simulation results for the adjusted and unadjusted sample size calculation. Hence they provide a comparison with the approach of Schoenfeld since the unadjusted sample size given is equivalent to sample size calculated using his Formula 2.3. Furthermore, simulations based on the sample sizes given by Shih's sample size program SIZE [121] were conducted and results from these are provided in each of the appropriate tables. Initial calculations for a trial without loss to follow-up, non-proportional hazards or cross-over show that sample sizes derived using SIZE are higher than sample sizes given by our method if the event rate is high, whereby our method gives power as desired. In a trial with a desired hazard ratio of 0.6 SIZE will give 2.5% higher sample size than ART for an event rate of 50% in the control arm whereas the difference between the methods will be only 0.5% if the control arm event rate is 10%.

3.4.1 More than two treatment arms

The results displayed in Table 3.1 illustrate simulation studies for trials with three treatment arms. Two experimental arms were simulated with a hazard ratio of $HR1$ and $HR2$ in comparison to the control arm respectively. All three arms were then analysed in a global logrank test. We can observe that power is maintained within the confidence bounds for all hazard ratio combinations.

3.4.2 Tied events

The derivation of sample size in Section 3.2 relies on the assumption of no tied events. We wanted to investigate how robust the calculations are to tied events occurring during the trial. In order to create tied events event times were rounded to two and three decimal places. This creates datasets with 37% and 2% of tied events respectively on average. The results are illustrated in Table 3.2. From these it is apparent that the calculations are robust to tied events.

3.4.3 Loss to follow-up

Table 3.3 outlines the simulations run for loss to follow-up. Time to loss to follow-up was simulated using an exponential distribution with a hazard calculated under a certain proportion of loss to follow-up by the end of the trial. Π_1^L and Π_2^L give the percentage of loss to follow-up assumed to have occurred in each treatment arm by the end of the trial.

From the power calculations under unadjusted sample size it is evident, at least under this model, that only high rates of loss to follow-up, i.e. 50% in both treatment arms, will lead to an important loss in power if they are not taken into account at the planning stage. This is because patients lost to follow-up over the course of the trial can still provide important and useful information if they are not lost at a very early stage. We need to be aware, though, that the estimate of the hazard ratio will only be unbiased if the reason for loss to follow-up is unrelated to the performance of the treatment regimen they are lost from. Nevertheless, under this assumption our approach performs well within the confidence interval around 90% power for all parameter combinations and generally slightly better than calculations according to SIZE. This comparison was not available for unequal proportions of loss to follow-up in the group since SIZE does not allow for that. Hence, while both methods allow for flexible calculations of loss to follow-up over the periods, SIZE does not allow for differing rates in each of the groups. Results for differing trial duration and allocation ratios were observed to be similar.

3.4.4 Non-proportional hazards

Simulation results under non-proportional hazards are displayed in Table 3.4. In this case the hazard in the experimental arm was changed for each patient after having survived two years in the trial which led to a change in the overall hazard ratio from HR1 to HR2. This was simulated by first assigning a probability to whether patients experienced an event before the time of changing hazard, i.e. at two years after a patient had entered a trial. If no event had been experienced, the exponential survival distribution was adapted to incorporate a change in hazards after this point causing a change in the hazard ratio from HR1 to HR2. These situations may occur when a treatment is very effective in the beginning but patients experience a levelling off of the treatment effect, which in turn brings the survival curves closer together over time or if, such as in a trial comparing surgery followed by chemotherapy with surgery alone, the two treatments have similar hazards in the beginning which then diverge over time. Unadjusted simulations were run by taking the first of the two hazard ratios ($HR1$) given in the table to calculate N. Another column of the table illustrates the impact on power and sample size by taking the arithmetic mean of the two hazard ratios when calculating the sample size required

Parameters		Analysis using global logrank test	
HR1	HR2	N	Power
0.6	0.9	344	90.7
0.7	0.8	714	89.9
0.8	0.7	714	90.4
0.9	0.6	344	90.1
0.8	0.8	1357	90.1

Table 3.1: Simulation results for three treatment groups
HR1 - hazard ratio in favour of first experimental group in comparison with control, HR2 - hazard ratio in favour of second experimental group in comparison with control, N - sample size calculated for 90% power, Power - power achieved through simulation with sample size N

Parameters	2 % tied events		37% tied events	
HR	N	Power	N	Power
0.6	206	90.0	206	90.1
0.7	408	90.1	408	90.1
0.8	1015	90.4	1015	90.4
0.9	4454	90.5	4454	90.5

Table 3.2: Simulation results for tied events
HR - hazard ratio in favour of experimental group, N - sample size calculated for 90% power, Power - power achieved through simulation with sample size N

Parameters			Adjusted for loss to follow-up		Unadjusted			SIZE		
HR	Π_1^L	Π_2^L	N	Power	N	Power	% diff N	N	Power	% diff N
0.7	0	0	408	90.0	408	90.0	0.0	415	91.1	+ 1.7
0.8	0	0	1029	89.5	1029	89.5	0.0	1029	89.5	0.0
0.7	5	20	424	89.9	408	88.9	- 3.9	n/a	n/a	n/a
0.7	20	5	424	90.3	408	88.9	- 3.9	n/a	n/a	n/a
0.7	5	5	414	89.6	408	89.1	- 1.5	423	90.9	+ 2.2
0.7	20	20	433	90.8	408	87.7	- 6.1	447	90.5	+ 3.2
0.7	30	30	448	89.7	408	87.4	- 9.8	466	90.6	+ 4.0
0.7	40	40	466	90.3	408	85.1	- 14.2	489	91.9	+ 4.9
0.7	50	50	487	89.7	408	83.3	- 19.4	516	91.2	+ 6.0
0.8	30	30	1112	90.3	1015	86.8	- 9.6	1154	91.3	+ 3.8
0.8	40	40	1155	90.0	1015	86.7	- 13.8	1209	91.0	+ 4.7
0.8	50	50	1206	89.6	1015	83.5	- 18.8	1276	91.0	+ 5.8

Table 3.3: Simulation results for loss to follow-up
HR - hazard ratio in favour of experimental group, Π_1^L - proportion lost to follow-up in control group by the end of the trial, Π_2^L - proportion lost to follow-up in experimental group by the end of the trial, N - sample size calculated for 90% power, Power - power achieved through simulation with sample size N, % diff N - change in sample size relative to adjusted use of ART for loss to follow-up in percent, i.e. % diff in N = (Adjusted / Unadjusted * 100) - 100, n/a - option not available in the program

Parameters		Adjusted for non-proportional hazards		Unadjusted			Adjusted using arithmetic mean			SIZE		
HR1	HR2	N	Power	N	Power	% diff N	N	Power	% diff N	N	Power	% diff N
0.6	0.9	274	89.9	206	80.9	- 33.0	619	99.8	+ 125.9	281	90.1	+ 2.6
0.6	0.8	249	90.1	206	85.3	- 20.9	408	98.6	+ 63.9	255	89.9	+ 2.4
0.6	0.7	227	90.1	206	87.0	- 10.2	285	95.8	+ 25.6	232	90.8	+ 2.2
0.7	0.8	458	90.2	408	85.6	- 12.3	619	96.7	+ 35.2	466	90.5	+ 1.8
0.8	0.7	869	89.3	1015	93.7	+ 16.8	619	78.2	- 29.8	882	90.5	+ 1.5
0.8	0.6	749	89.9	1015	96.9	+ 35.5	408	67.3	- 46.6	761	90.0	+ 1.6

Table 3.4: Simulation results for non-proportional hazards
HR1 - hazard ratio in favour of experimental group for first two years in trial, HR2 - hazard ratio after two years in trial, N - sample size calculated for 90% power, Power - power achieved through simulation with sample size N, % diff N - change in sample size relative to adjusted use of ART for non-proportional hazards in percent, i.e. % diff N = (Adjusted / Unadjusted * 100) - 100

for the trial.

We can observe from the simulation results that if sample size is calculated assuming proportional hazards in a situation where hazards vary over time, this can lead to significant over- or underestimation of sample size depending on the direction of the evolution of the hazard ratio over time. Furthermore, we can observe that already for a small change in the hazard ratio, from 0.7 to 0.8 for example, a loss in power of more than 4% occurs (compared to unadjusted calculations) if this change is not adjusted for.

3.4.5 Cross-over

In order to investigate the performance of our method when cross-over is expected to occur (Table 3.5), time to cross-over was simulated using the exponential distribution in the experimental group similar to the simulations looking at loss to follow-up. Thus patients were simulated to cross over to the other treatment group at a certain time in the trial if they had not experienced an event before that time. Following such an event, patients would then continue to follow the hazard of the treatment group they had crossed over to. Hence, cross-over occurred with probability Π_2^C, which is given as a percentage of patients in the tables.

The simulation results illustrate that adjusting for cross-over becomes particularly important as we approach 20% cross-over in one of the treatment arms (or 10% in both arms) if time to cross-over follows an exponential distribution. We have found that as cross-over from both arms increases, ART gives more conservative sample size estimates than SIZE. This may be due to the different assumptions underlying their calculations whereby patients are allowed to change treatment groups more than once over the course of the trial under SIZE.

3.4.6 Multiple adjustments

We furthermore evaluated the performance of the sample size approach under the presence of non-proportional hazards, loss to follow-up and cross-over in one trial as illustrated in Table 3.6. These scenarios were designed in the same way as the separate simulation studies for loss to follow-up, non-proportional hazards and cross-over. In this case, patients in the experimental treatment group were subjected to a change in hazards after two years if they had not had an event, been lost to follow-up or crossed over to the control arm before that point in time.

Apart from the assessment of performance in terms of power attained, a further objective was to assess whether the effect of these adjustments is additive in terms of power and sample size. The simulation results convey that the difference in terms of sample size between adjusting for loss to follow-up, non-proportional hazards and cross-over and not adjusting for any can be vast, in some cases as extreme as 63%. Similarly, actual power achieved may be nearly 20% less than the nominal power of 90%. In other situations, we can observe in the table that the presence of non-proportional hazards may offset the effect of cross-over in terms of power achieved. This situation arises if we designed the trial for a constant hazard ratio which was higher than the hazard ratio obtained by the end of the trial due to a decrease of the hazard in the experimental group over time.

Parameters			Adjusted for cross-over		Unadjusted			SIZE		
HR	Π_1^C	Π_2^C	N	Power	N	Power	% diff N	N	Power	% diff N
0.6	0	5	212	90.0	206	88.8	- 2.9	217	91.2	+ 2.4
0.6	0	10	218	89.9	206	88.2	- 5.8	224	90.8	+ 2.8
0.6	0	20	232	90.1	206	86.1	- 12.6	238	90.4	+ 2.6
0.6	0	30	248	90.7	206	84.2	- 20.4	256	90.7	+ 3.2
0.7	0	30	489	90.5	408	83.6	- 19.9	502	90.9	+ 2.7
0.8	0	30	1213	90.1	1015	84.2	- 19.5	1240	90.2	+ 2.2
0.9	0	30	5312	89.4	4454	84.0	- 19.3	5429	90.6	+ 2.2
0.7	10	10	458	90.0	408	87.0	- 12.3	467	90.5	+ 2.0
0.7	20	10	490	91.0	408	85.3	- 20.1	499	91.0	+ 1.8
0.7	30	10	527	91.7	408	83.3	- 29.2	537	92.0	+ 1.9
0.7	20	20	522	90.9	408	84.1	- 27.9	530	92.2	+ 1.5
0.7	30	30	606	91.7	408	78.1	- 48.5	609	91.4	+ 0.5

Table 3.5: Simulation results for cross-over
HR - hazard ratio in favour of experimental group, Π_1^C - proportion crossing over to different treatment regimen from control group by the end of the trial, N - sample size calculated for 90% power, Power - power achieved through simulation with sample size N, % diff N - change in sample size relative to adjusted use of ART for cross-over in percent, i.e. % diff in N = (Adjusted / Unadjusted * 100) - 100

Parameters						Adjusted		Unadjusted		
HR 1	HR 2	Π_1^L	Π_2^L	Π_1^C	Π_2^C	N	Power	N	Power	% diff N
0.6	0.7	30	30	0	20	274	90.8	206	80.7	- 33.0
0.6	0.7	30	30	0	30	291	89.2	206	78.2	- 41.3
0.6	0.8	30	30	0	20	296	89.5	206	76.4	- 43.7
0.6	0.8	30	30	0	30	313	90.0	206	74.8	- 51.9
0.6	0.9	30	30	0	20	319	89.9	206	75.4	- 54.9
0.6	0.9	30	30	0	30	337	89.6	206	71.9	- 63.4
0.8	0.6	30	30	0	20	964	90.1	1015	91.6	- 5.0
0.8	0.6	30	30	0	30	1036	89.8	1015	88.8	- 2.1
0.6	0.8	20	20	10	10	292	90.5	206	78.9	- 41.7
0.6	0.8	20	20	10	20	308	90.9	206	76.6	- 49.5
0.6	0.8	20	20	20	20	331	91.4	206	74.8	- 60.7

Table 3.6: Simulation results for loss to follow-up, non-proportional hazards and cross-over combined
HR 1 - hazard ratio in favour of experimental group for first two years in trial, HR 2 - hazard ratio after first two years, Π_1^L - proportion lost to follow-up in control treatment group by the end of the trial, Π_1^C - proportion crossing over to different treatment regimen from control group by the end of the trial, N - sample size calculated for 90% power, Power - power achieved through simulation with sample size N, % diff N - change in sample size relative to adjusted use of ART for loss to follow-up, non-proportional hazards and cross-over in percent, i.e. % diff in N = (Adjusted / Unadjusted * 100) - 100

3.5 Non-uniform accrual

We examined the accrual pattern of four cancer trials conducted by the MRC and compared these with a uniform pattern. The results are illustrated in Figure 3-2. For each trial recruitment per calendar month is plotted. The solid line illustrating uniform accrual was derived from the length of the accrual period as originally defined in each trial protocol. In TE08 (ISRCTN: 6475197) [18] actual and uniform accrual differed by four months, in BR11 (EORTC 26951) [21] by 19 months and in CR08 (ISRCTN: 79877428) [20] by seven months. Recruitment in CH03 (ISRCTN: 62576956) [19] was never completed. Fitting an exponential entry pattern with a scale parameter of -1 to the recruitment of TE08 mirrors its actual process more closely. ART allows the user to specify a certain fraction of patients to have been accrued before the start of the trial. BR11 is a trial where this facility was needed as nearly half of the patients required had already been entered by the EORTC (European Organisation for Research and Treatment of Cancer) before recruitment opened at the MRC. Hence in order for this trial to accurately calculate the projected end of recruitment one not only needs to take into account the actual accrual pattern but also the point mass at zero of the cumulative distribution function of recruitment time.

The impact of a concave exponential accrual pattern, i.e. an exponential entry pattern with a negative exponent, is further examined in Figure 3-3 and Table 3.7. Results in Table 3.7 were obtained by conducting 'what if' calculations in ART. Underlying all trial scenarios is a hazard ratio of 0.7, a median survival of five years and a two-sided 5% significance level. Comparisons were then made with a uniform recruitment pattern (Figure 3-3). Power was calculated for each trial for a sample size of 634 (the sample size needed under a uniform recruitment pattern) using the same trial length as in the uniform case, i.e. nine years, under exponential accrual with scale parameter γ. This illustrates that unless the recruitment pattern deviates substantially from the uniform distribution, the impact on power is modest. The most extreme departures from uniform accrual are given in the last three rows of Table 3.7 which represents a gradually increasing rate of accrual to a trial which starts recruiting very slowly. This shows that if we were to analyse the trial at the original time of analysis we would have significantly reduced power ($\geq$5%). One way of addressing this is by moving the planned time of analysis to the time point at which the planned number of events have been observed. The last column in the table illustrates the time at which these planned analyses can be performed while maintaining

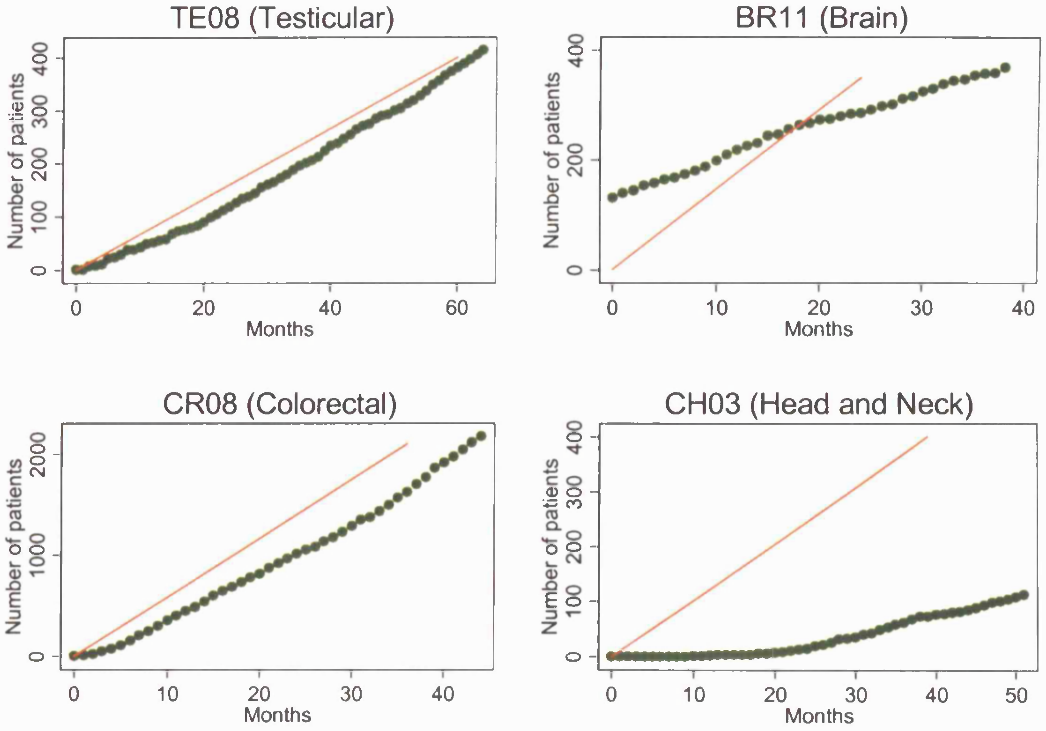


Figure 3-2: Accrual patterns in four MRC cancer trials Observed accrual patterns are given by the dotted lines. The solid lines illustrate a uniform accrual pattern based on accrual as anticipated in the trial protocols

the planned number of events and power for the trial. These results are in line with observations made by Lachin & Foulkes [71]. However, the impact of the accrual pattern on power will also depend on the shape of the survival distributions. If there is an early peak in the hazard, early events, and hence patients accrued early on in the trial, are more influential.

3.6 Discussion

Sample size calculations are necessary for all randomised controlled trials. They are particularly complex for trials with survival-type endpoints because they usually involve assessments and input of a number of parameters including: the control group survival distribution; the magnitude and form of the targeted difference to be detected; the rate of accrual of individuals to the study; the length of follow-up of individuals after accrual closure; and the potential for (time-related) dilution of any effect through, for example, loss to follow-up or cross-over. All of these parameters can have an important impact on the trial size needed. We have presented a general approach to sample size calculations for trials which allows for all these sources of variability. The methodology and associated software allow the user to specify, at the design stage, the use of a general family of logrank tests, including the Tarone & Ware and Harrington & Fleming families. Furthermore our Stata program allows for the specification of non-local alternatives by approximating the logrank test statistic Q using a scaled noncentral χ^2 on $k-1$ degrees of freedom. Simulations have, however, shown that this brings only minor improvements in accuracy since the method already performs well under local alternatives with more extreme designs (Figure 3-4).

We note a slight underlying difference between our approach and much of what has gone before. Sample size calculations based on Freedman [41] or Schoenfeld [114] as given in Equation 2.3 assume that the number at risk and hence the number of events are a constant for a given hazard ratio, power and significance level. However, the number of events given by ART will vary slightly even for small changes such as a different accrual rate in one of the periods since the number at risk is not treated as constant over the course of the trial but instead is calculated for each of the periods.

Simulation results show that our method works well in a variety of situations. These results also indicate that the adjustments particularly for non-proportional hazards, non-uniform accrual and cross-over may be substantial in terms of power and sample size.

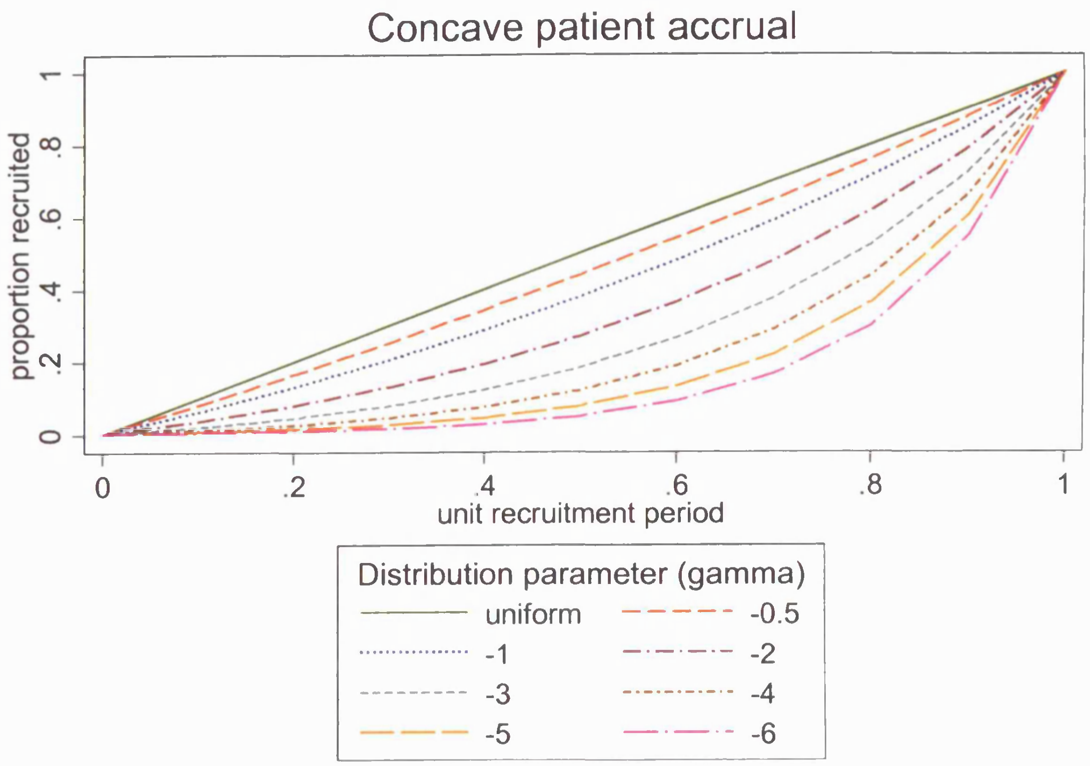


Figure 3-3: Concave exponential patient accrual patterns

Entry distribution	Power	Additional follow-up
γ	for R = 5, T = 9	required for 90% power
0	90.0 %	0
-0.5	89.4 %	0.25
-1	90.0 %	0.5
-2	87.4 %	0.75
-3	86.2 %	1
-4	85.6 %	1.25
-5	85.1 %	1.5
-6	84.6 %	1.75

Table 3.7: Impact on power and length of trial of concave recruitment pattern γ - exponential scale parameter, Power is that for N=634 at length of accrual (R) five years and follow-up (T-R) four years. All calculations use a hazard ratio of 0.7, median survival of five years and a two-sided 5 % significance level

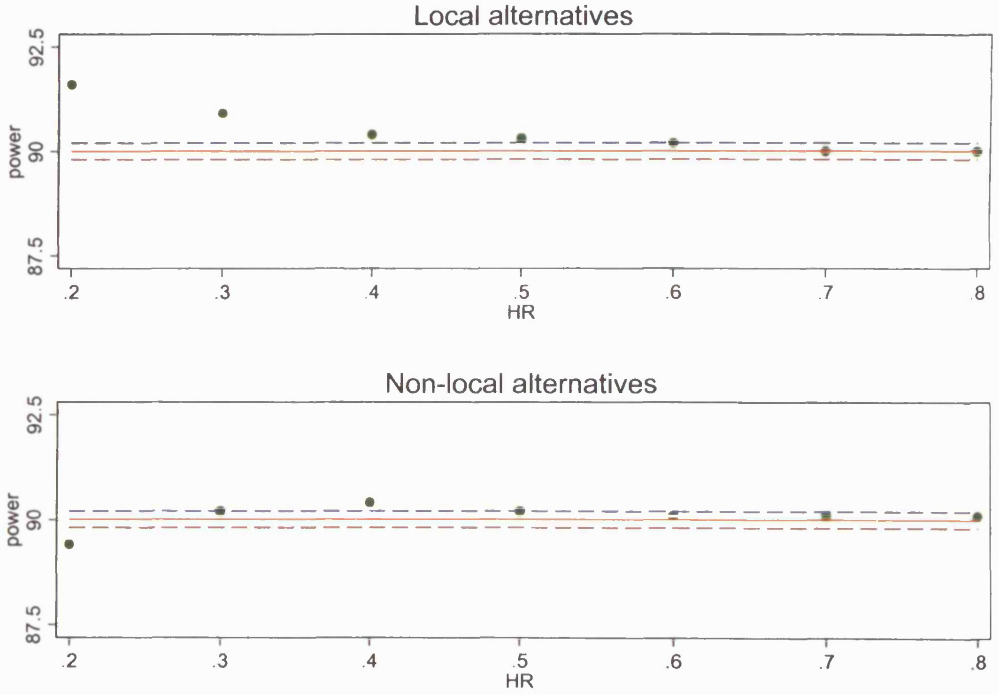


Figure 3-4: Simulation results for power based on sample size calculations using ART under local and non-local alternatives
Trial set-up: equal allocation to both treatment arms, accrual = two time periods, follow-up = two time periods. The dotted lines illustrate an approximate 95% CI around 90% power. Results based on 100,000 replications

Hence researchers should take particular care in specifying these parameters when designing a trial. Of course, trials in which a large percentage of patients were expected to be lost to follow-up would be unlikely to be successful for other reasons. One would have to cast doubt on the validity of the trial's results not only because of the loss of power but also because of the potential for bias if the missing outcome data due to loss to follow-up were associated with the outcome. The data in Table 3.3 are provided as a sensitivity analysis, and show that modest loss to follow-up may have only a minor effect on power and / or required sample size. However, cross-over, which occurs frequently particularly in longer term trials, does not generate missing outcome data since our methods assume that the treatment groups would be compared as randomised on an intention-to-treat basis. Loss of power is then the primary concern since cross-over will lead to a dilution of the difference in the treatment effect between the randomised groups.

It may be very difficult to specify all these variables with reasonable accuracy before the start of the trial. In this situation two approaches are very helpful. First, as one design stage, it is probably useful and prudent to perform sensitivity analyses varying these parameters to assess the impact of modest changes in them, to assess the robustness of the design under realistic departures from the design specified. Second, as the trial accumulates individuals and data, the design specifications can be checked against the real accumulating data. If there are important departures from these the impact on the trials operating characteristics (particularly the power) can be formally calculated and the trial can be potentially amended. For example, if during the course of the trial we find that cross-over from one treatment to another is greater than anticipated, then we may argue that a smaller difference than that originally specified in the alternative hypothesis should be targeted. In this case the sample size of the trial may be amended. We note that such 'administrative' amendments are perfectly acceptable during the course of the trial, as long as any decisions to change the trial size are made independently of, and preferably blind to, the estimate of the treatment difference currently being observed within the trial.

Furthermore our software may be used for the design of multi-arm trials where the primary question concerns a comparison of each experimental arm with the control. Consider a three-arm trial, i.e. two experimental arms and one control, with an overall type I error probability of 5%. After applying a Bonferroni adjustment allowing for multiple comparisons and a correlation of 0.5 between the two test statistics the sample size can then be calculated in ART using a two-arm design with a type I error of 3.5% [107]

and multiplying this by 1.5 to get the correct sample size for a three-arm trial. This calculation assumes a randomisation of 1:1:1.

Chapter 4

ART - Analysis of resources for trials

4.1 Introduction

Royston & Babiker [101] presented a menu-driven Stata program for the calculation of sample size or power for complex clinical trial designs under a survival time or binary outcome. This program allows for multi-arm trials with up to six treatment arms, an arbitrary time-to-event distribution, non-proportional hazards, unequal patient allocation, non-uniform rates of patient entry, loss to follow-up and cross-over of patients from their allocated treatment to an alternative treatment arm. In the present chapter, the program is updated to operate under the new Stata 8 dialog interface. Additionally, its name has been changed to `ART - Analysis of Resources for Trials`. We report here some further improvements to the software, such as allowing for the input of a one-sided significance level and the calculation of sample size for non-inferiority trials.

To recapitulate, for survival-time outcomes, the main assumption is that treatment groups will be compared using the logrank test. Computations are carried out according to the asymptotic distribution of the logrank test statistic Q. Here Q is defined as $U'V^{-1}U$, where U is the vector of the total observed minus the expected number of events in each of the k treatment groups in the design except for the control and V is the covariance matrix of U. A full report on the methodology and its performance in particular with respect to loss to follow-up, non-proportional hazards and cross-over is given in the previous Chapter 3 as well as in [7].

For binary outcomes a normal approximation to the binomial distribution is assumed. The program gives sample sizes which are slightly lower than those provided by the Stata command `sampsi` since it does not use a continuity correction.

4.2 New design of menu and dialogs

All features are available from the newly designed `ART` menu and associated dialogs. As before, on completion of the calculations the command line that generated the results will be displayed in the Review window. For reproducibility of the calculations we suggest that the user opens a log-file before executing the commands via the dialog which will hence save the command line. This log-file can then be edited to produce a do-file to repeat the calculations if desired.

When `artmenu` has been executed using `artmenu on`, a new item `ART` will appear on the system menu bar under `User`. This menu may be turned off by typing `artmenu off`. `ART` contains the following two items:

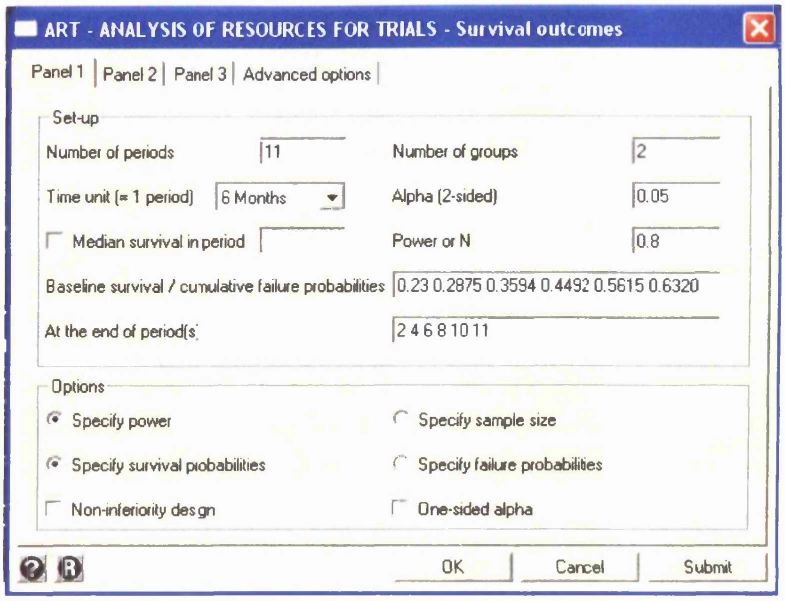
`Survival outcomes` Sets up all design parameters including advanced options such as loss to follow-up and cross-over for survival time trials

`Binary outcomes` Sets up design parameters for trials with a binary outcome under a simple design

Since no considerable changes have been made to the `Binary outcomes` facility this chapter will concentrate on the changes made to `Survival outcomes` and readers are referred to the original article by Royston & Babiker [101] for further information on trials with binary outcomes. At any stage the user may obtain further information on the use of the menu by clicking on the `?` button.

4.2.1 Survival outcomes - panel 1

Figure 4-1 illustrates the new dialog window for `Survival outcomes`. `Panel 1` requires the input of the basic trial set-up. The main change from the old dialog is the input of the survival/failure probabilities. These can now be input by either specifying median survival in a particular period or by filling in the cumulative probabilities at the end of periods as illustrated in Figure 4-1. Furthermore, the actual time units of periods may be specified, such as years, 6 months, quarters, months etc.. The choice of these does not have any impact on the sample size calculations themselves but is displayed in the final output to remind the user of the timescale assumed.



Set-up

Number of periods: 11

Number of groups: 2

Time unit (= 1 period): 6 Months

Alpha (2-sided): 0.05

□ Median survival in period

Power or N: 0.8

Baseline survival / cumulative failure probabilities: 0.23 0.2875 0.3594 0.4492 0.5615 0.6320

At the end of period(s): 2 4 6 8 10 11

Options

● Specify power

○ Specify sample size

● Specify survival probabilities

○ Specify failure probabilities

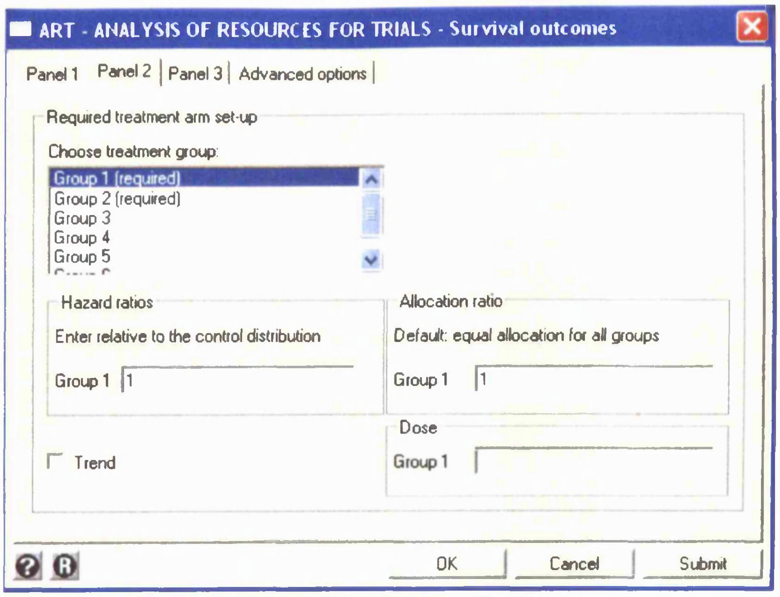
□ Non-inferiority design

□ One-sided alpha

Figure 4-1: A completed Panel 1 screen for survival outcomes

4.2.2 Survival outcomes - panel 2

Hazard ratios for each treatment group relative to group 1 as well as allocation ratios may be entered on `Panel 2`, as illustrated in Figure 4-2. This needs to be done for the number of groups specified on `Panel 1`. Only one value per treatment group needs to be entered for the hazard ratio if these are assumed constant over time. In the case of non-proportional hazards, one value may be entered for each period of the trial. For example, if the number of periods has been set at 11 in `Panel 1`, 11 hazard ratios may be entered in `Panel 2` for each of the groups. If for a given group fewer hazard ratios are entered than the number of periods, the remaining hazard ratios are taken to have the same value as the last specified hazard ratio. In addition, if no hazard ratio is specified for a particular group, its value in a given period is taken to be the geometric mean of the hazard ratios specified for the same period across all the groups for which a value has been entered. When a test for trend is chosen, the dose may be input for each treatment group.



Required treatment arm set-up

Choose treatment group:

- Group 1 (required)
- Group 2 (required)
- Group 3
- Group 4
- Group 5

Hazard ratios

Enter relative to the control distribution

Group 1: 1

Allocation ratio

Default: equal allocation for all groups

Group 1: 1

Dose

Group 1:

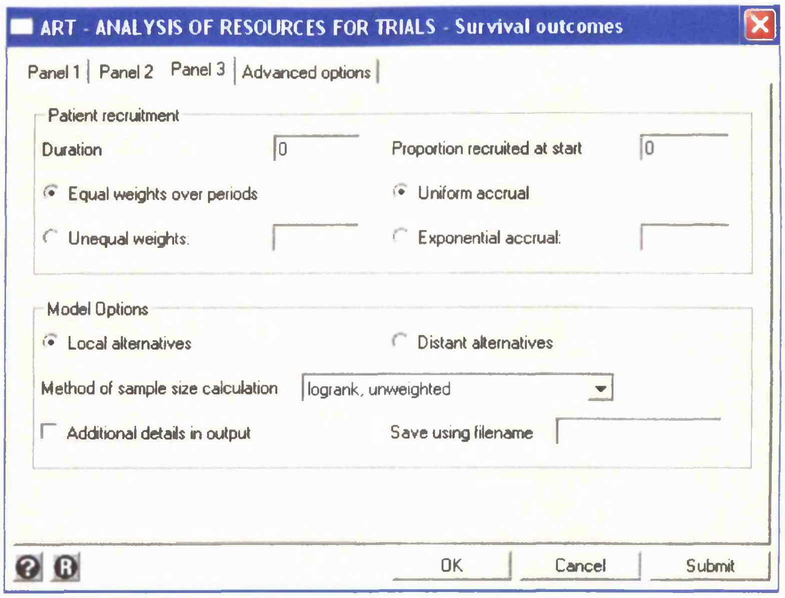
☐ Trend

Figure 4-2: Panel 2 screen completed for a two arm trial

4.2.3 Survival outcomes - panel 3

Panel 3, which is illustrated in Figure 4-3, requires the input of patient recruitment options and the selection of the analysis method from the dropdown list. The inputs are similar to those of the original dialog. By default, calculations will be run using equal weights over the periods. If this is not the case, unequal weights may be entered, e.g. 1 2 2, for each of the periods over which recruitment takes place. As before, steady recruitment using the uniform distribution is assumed as a default. If exponential accrual is chosen instead, the rate needs to be entered in the `Exponential accrual` box.

The default method of computation is the unweighted logrank test under local alternatives. This implies that sample sizes are derived under the assumption that hazard ratios between treatment groups are not far from one. However, simulations provided in Chapter 3 have shown that the improvements in terms of accuracy gained by computing sample size under distant alternatives are minimal. Sample sizes derived through computations under local alternatives will be slightly conservative for hazard ratios $<$ 0.5.



Patient recruitment

Duration 0

Proportion recruited at start 0

● Equal weights over periods

○ Unequal weights:

● Uniform accrual

○ Exponential accrual:

Model Options

● Local alternatives

○ Distant alternatives

Method of sample size calculation logrank, unweighted

□ Additional details in output

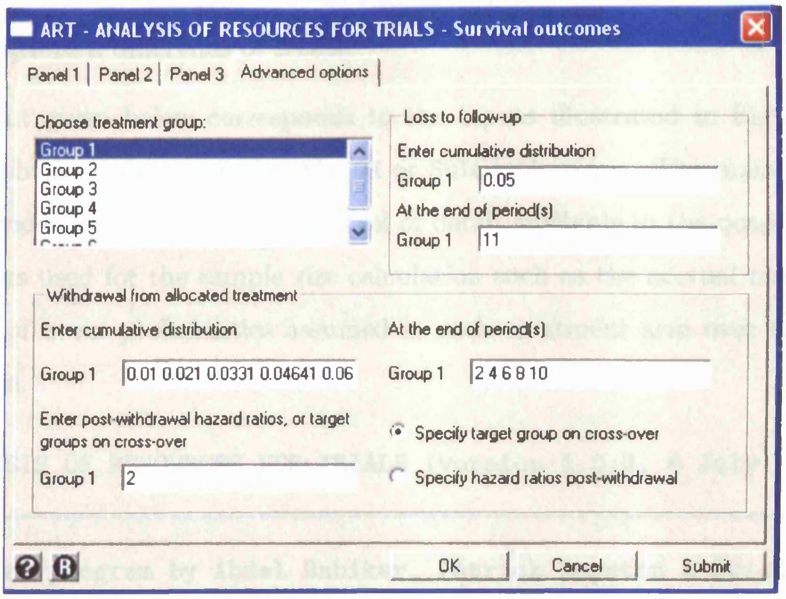
Save using filename

Figure 4-3: A completed Panel 3 screen illustrating input of recruitment options

4.2.4 Survival outcomes - advanced options

The last part of the dialog window for `ART` shown in Figure 4-4 allows the input of loss to follow-up and cross-over for each of the treatment groups in the trial as specified in `Panel 1` in a similar manner to the input of survival probabilities and hazard ratios in Figures 4-1 and 4-2.

Both loss to follow-up and cross-over need to be entered as a cumulative distribution. The user may then choose to `Specify target group on cross-over` or to `Specify hazard ratios post-withdrawal`. The first option assumes that patients withdrawing from treatment of a particular group will receive the treatment regimen of the target group and hence take on that hazard after crossing over. If the second option is chosen a post-withdrawal hazard ratio function relative to the hazard of the control arm failure time distribution needs to be entered for each arm that is subject to cross-over. Similar to the hazard ratios between groups entered in `Panel 2` (see Figure 4-2) as many values as there are periods may be entered. If the number of values entered is less than the number of periods, then the last hazard ratio value applies to the remaining periods. This option is favourable over the first if patients withdrawing from allocated treatment over the course of the trial are expected to do much worse than either treatment group for example.



Choose treatment group: Group 1

Loss to follow-up

Enter cumulative distribution

Group 1: 0.05

At the end of period(s)

Group 1: 11

Withdrawal from allocated treatment

Enter cumulative distribution

Group 1: 0.01 0.021 0.0331 0.04641 0.06

At the end of period(s)

Group 1: 2 4 6 8 10

Enter post-withdrawal hazard ratios, or target groups on cross-over

Group 1: 2

● Specify target group on cross-over

○ Specify hazard ratios post-withdrawal

Figure 4-4: Advanced options for survival outcomes

4.3 Optima

Optima is a clinical trial currently running in the UK, Canada and the US which is designed to determine the optimal management of patients with HIV infection for whom first and second line highly active antiretroviral therapy (ART) has failed. Patients are randomised equally between standard ($\leq$ 4 drugs) - and mega ($>$ 4 drugs) - ART. The assumptions for sample size calculation, based on earlier data on similar patients, were as follows: The standard-ART cumulative event rate in year 1 is 23% with a 25% annual increase thereafter until the end of the study and cross-over from mega- to standard-ART is 5% in year 1 and decreases by 50% every year thereafter. The hazard ratio is 0.7 and loss to follow-up at 5.5 years is 5% with drop-in from standard- to mega-ART at 1% in year 1 (increasing by 10% every year thereafter). Furthermore a significance level of 5% with 4.5 years accrual and one year minimum follow-up under a power of 80% were assumed. Under these assumptions our program predicts that a sample size of 490 with 318 expected events will be sufficient to detect a clinically relevant difference between the treatment groups. In comparison, if loss to follow-up and cross-over are not adjusted for we arrive at a sample size of 379 with 248 expected events.

If this trial had been designed with 90% power, as is quite frequently done in practice, we would find a difference of 29.1% in sample size between adjusted and unadjusted calculations, i.e. an increase in sample size due to adjustment for loss to follow-up and cross-over from 508 to 656 patients, and an increase from 332 to 425 required events which translates to a difference of 28.0%.

The output given below corresponds to the inputs illustrated in Figures 4-1 - 4-4 and may be obtained upon pressing the OK or Submit buttons. The main improvement from the previous version concerns the level of detail available in the output in terms of the parameters used for the sample size calculation such as the accrual method and the development of event probabilities assumed in each treatment arm over the number of periods chosen.

```
ART - ANALYSIS OF RESOURCES FOR TRIALS (version 1.0.5, 6 July 2005)
------------------------------------------------------------------------------
A sample size program by Abdel Babiker, Patrick Royston & Friederike Barthel,
MRC Clinical Trials Unit, London NW1 2DA, UK.
------------------------------------------------------------------------------
Type of trial                           Superiority - time-to-event outcome
Statistical test assumed                Unweighted logrank test (local)
Number of groups                        2
Allocation ratio                        Equal group sizes

Total number of periods                 11
Length of each period                   6 months
Cum. event probs per period (group 1)   0.123 0.230 0.259 0.287 0.324 0.359
                                         0.406 0.449 0.509 0.561 0.632
Cum. event probs per period (group 2)   0.087 0.167 0.190 0.211 0.240 0.268
                                         0.306 0.341 0.392 0.438 0.503
Number of recruitment periods           9
Number of follow-up periods             2
Method of accrual                       Uniform
Recruitment period-weights              1 1 1 1 1 1 1 1 1 0 0

Hazard ratios as entered (groups 1,2)   1, 0.7
Hazard ratios per period (group 1)      1.000 1.000 1.000 1.000 1.000 1.000
                                         1.000 1.000 1.000 1.000 1.000
Hazard ratios per period (group 2)      0.700 0.700 0.700 0.700 0.700 0.700
                                         0.700 0.700 0.700 0.700 0.700
```

```
Alpha                                 0.050 (two-sided)
Power (designed)                      0.800

Total sample size (calculated)        825
Expected total number of events       287
---------------------------------------------------------------------------
Values given below apply to each group at the end of the trial
---------------------------------------------------------------------------
Unadjusted event probs (groups 1,2)   0.632, 0.503
Unadjusted loss to follow-up probs    0.050, 0.050
Unadjusted cross-over probabilities   0.068, 0.098

Expected proportions of event         0.392, 0.303
Expected proportions lost to follow-up 0.022, 0.024
Expected proportions of cross-over    0.026, 0.072
---------------------------------------------------------------------------
```

The first part of the output gives an overview of the trial parameters chosen by the user at the time of filling in the dialog menu. A detailed display of the cumulative event probabilities in the treatment groups and the hazard ratios over each of the periods in the trial allow the user to check that the trial design was input correctly. Sample size and number of events needed for the trial design are given towards the end of that ouput.

The second part of the output appears only if the `Additional details in output` option is checked. It provides further information regarding the expected performance in all treatment groups by the end of trial, in particular with regards to loss to follow-up and cross-over proportions in all arms.

Furthermore, the user may save probabilities and hazard ratios used in the calculations to a new file by filling in the `Save using filename` box.

4.4 Comparison with other available software

Several sample size programs are currently available which provide calculations for those trials with survival-type data which are to be designed and analysed using the logrank

	Cost	Statistical package needed	logrank sample size	flexible accrual	non - uniform accrual	non - prop. hazards	loss to follow-up	flexible loss to follow-up
ART	Free	Stata	✓	✓	✓	✓	✓	✓
cpower	Free	R	✓	✓				
Clinical Trials Design Program (v. 1)	$299	none	✓	✓			✓	
EGRET SIZ (v. 1)	$465	none		✓	✓		✓	✓ not by group
Ex-Sample (v. 3.0)	$125	none	✓	✓				
NCSS PASS (2004)	$899.95	none	✓	✓	✓		✓	✓
NQuery advisor (v. 5.0)	$995	none	✓	✓			✓	
Nsurv (v. 2.2)	$140	none	✓	✓			✓	✓ not by group
POWER	$10	none	✓	✓				
Power & Precision (v. 2)	$995	none	✓	✓	✓	✓	✓	✓
PS Power	Free	none	✓	✓				
Schoenfeld	Free	none	✓	✓				
SIZE	Free	SAS	✓	✓		✓	✓	✓ not by group
Statistica Power Analysis (v. 6)	€619	Statistica	✓	✓				
UnifyPow (v. 2002.08.17a)	Free	SAS	✓	✓				

	cross - over designs	multi - arm designs	Results			Methodology authors
			sample size	number of events	power	
ART	✓	✓	✓	✓	✓	Barthel et al. (2005)
cpower (10/03/2004)	✓		✓			Lachin & Foulkes (1986) [71] Schoenfeld (1983) [114]
Clinical Trials Design Program (v. 1)			✓	✓		Freedman (1982) [41] Rubinstein et al. (1981) [104]
EGRET SIZ (v. 1)		✓	✓		✓	Self et al. (1992) [120]
Ex-Sample (v. 3.0)			✓			Schoenfeld & Richter (1982) [115]
NCSS PASS (2004)			✓		✓	Lachin & Foulkes (1986) [71]
NQuery advisor (v. 5.0)			✓	✓	✓	Lakatos & Lan (1992) [72]
Nsurv (v. 2.2)			✓			Lachin & Foulkes (1986) [71]
POWER (v. 1.4)			✓		✓	Schoenfeld & Richter (1982) [115]
Power & Precision (v. 2)			✓		✓	Schoenfeld (1983) [114] Lakatos (1988) [73]
PS Power (v. 2.1.30)			✓			Schoenfeld & Richter (1982) [115]
Schoenfeld (03/05/2001)			✓		✓	Schoenfeld (1983) [114]
SIZE (26/06/1996)	✓		✓	✓	✓	Lakatos (1988) [73]
Statistica Power Analysis (v. 6)			✓			Schoenfeld (1983) [114]
UnifyPow (v. 2002.08.17a)			✓		✓	Self et al. (1992) [120]

Table 4.1: Properties of available sample size programs
Disclaimer: the features and costs of programs mentioned in this table were, as far as the author is aware, correct at the time of writing. The author is happy to change any information on the programs as necessary

test. Table 4.1 displays a large selection of these. Those which are identified as 'Free' in Table 4.1 are available for download over the internet free of charge and the rest are commercially available. Most of the programs, as is illustrated in Table 4.1, provide calculations under the logrank test and allow for the incorporation of accrual and follow-up times. Many also allow 'loss to follow-up' expressed as a proportion of patients lost by the end of the trial. However, most methods do not provide adjustments for non-uniform accrual of individuals into the trial, non-proportional hazards, cross-over (from one treatment to the other) and multi-arm trials. The program SIZE perhaps comes closest to ART in achieving all these aims. However, SIZE does not allow for non-uniform accrual into the trial nor does it allow for more than two arms in a trial. Furthermore, at least in the simulations we have performed, the adjustment provided for most of these parameters in SIZE can lead to slightly overpowered designs. In terms of software needs ART requires an installation of Stata while SIZE requires SAS.

4.5 Conclusions

The new design of the dialog menu exploiting features introduced in Stata 8 and more detailed output are the main improvements to `ART`. In addition, the sample size calculations may now be performed for non-inferiority designs. This option may be specified on `Panel 1` (see Figure 4-1) while all other parameters are input in the same way as described above. Furthermore, the program now allows for the choice of a one-sided alpha which may also be specified on `Panel 1`. Finally, the help files have been updated. In some instances the user may want to run several calculations with similar parameters and in this case does not require the header given in the output for each of the calculations. To suppress this output the option `nohead` may be added at the end of the command line. Our approach and the associated ART software also provides sample size calculations in the context of trend tests on dose/response studies.

In summary, users should find the new version easier to use and more informative than the first release. The validity of the calculations has been checked via extensive simulation studies of which some details are provided in the previous Chapter 3 and in Barthel et al. [7].

Further work includes the extension of the methods described in Sections 3.2 and 3.3

by allowing periods of different lengths to the requirements of multi-stage trial designs, for example as described by Royston et al. [103] and in Chapters 5 - 7. In addition, we may consider to relax the assumption of tied events.

Chapter 5

Surrogate markers and multi-stage trials - a review

5.1 Introduction

With the new advances in molecular biology and the ever increasing identification of new molecular targets for therapy, potential cancer agents are increasingly becoming available. However, inevitably for a variety of practical reasons only a limited number of patients can be entered into clinical trials in order to establish efficacy. An increasing desire in a variety of disease areas for new and promising drugs to be approved for marketing as soon as possible has led to approval being based on intermediate outcome measures, such as biomarkers, rather than on long-term clinical outcome measures. In this context several authors including Ellenberg & Fleming [31] have explored the use of surrogate outcomes in order to reduce the length of trials as well as the possibility of multi-stage designs [110] [141] which allow the testing of several agents in one trial.

During the course of this chapter we will first give an overview of the statistical debate surrounding the validation of surrogate markers. Following that, a short introduction to the medical aspects of the debate and the practical use of surrogate markers, especially in the case of cancer, HIV and cardiovascular disease trials, is given. The second part of the chapter considers sequential methods and two-stage selection designs in particular. We conclude the chapter on a combination of both methods, i.e. the integration of surrogate markers into multi-stage selection designs.

5.2 Surrogate Markers

5.2.1 Introduction

According to Ellenberg [31] *'investigators use surrogate endpoints when the endpoint of interest is too difficult and/or expensive to measure routinely and when they can define some other, more readily measurable, endpoint, which is sufficiently well correlated with the first to justify its use as a substitute'.* Surrogate markers or intermediate endpoints have received ever increasing attention during the past 20 years and their use in clinical trials has been the subject of much debate. First and foremost this debate centres around the question of how to define and validate such markers [91] [29] [111] [136] [59] [23].

ICH Guidelines on Statistical Principles for Clinical Trials state that *'In practice, the strength of the evidence for surrogacy depends upon 1) the biological plausibility of the relationship, 2) the demonstration in epidemiological studies of the prognostic value of the surrogate for the clinical outcome and 3) evidence from clinical trials that treatment effects on the surrogate correspond to effects on the clinical outcome'* [87].

5.2.2 The Prentice criterion

The most often cited definition of a surrogate marker was given by Prentice [99] in 1989 and is known as the Prentice criterion. This defines a surrogate marker as *'a response variable for which a test of the null hypothesis of no relationship to the treatment groups under comparison is also a valid test of the corresponding null hypothesis based on the true endpoint'.* Thus the surrogate variable is required to capture any relationship between the treatment under consideration and the true endpoint employed. Mathematically this can be expressed using some function f_D

$$f_D\{t; S(t), A\} \equiv f_D\{t; S(t)\} \tag{5.1}$$

where for a conditional probability distribution $S(t)$ a surrogate for the primary endpoint D should be able to capture the dependence of D on treatment A. Hence the surrogate variable is required to be fully sensitive to any treatment difference in true endpoint rates and the treatment under consideration should not be allowed to influence the endpoint of interest via a mechanism unrelated to the surrogate. His operational criteria require that

- the treatment has a significant impact on the surrogate endpoint
- the treatment has a significant impact on the primary endpoint
- the surrogate has a significant impact on the primary endpoint
- the full effect of treatment upon the primary endpoint is captured by the surrogate

Thus an important drawback of this method is that evidence from trials with nonsignificant treatment effects may not be used, even though these trials may be consistent with a desirable relationship between both endpoints [84].

Prentice considers papers by Ellenberg & Hamilton [31] and Wittes et al. [143] in order to see how their choice of surrogate variables compare with his criterion. Ellenberg & Hamilton have suggested progression free survival as a possible surrogate to survival for cancer trials. Hence an event will in this case be defined as either disease progression or death. This marker may be difficult to validate in terms of the Prentice criterion since this would involve a comparison of death rates among these patients with corresponding rates for a comparable group of patients without a prior cancer diagnosis. Wittes et al. have proposed the use of markers such as blood cholesterol in trials of cholesterol lowering drugs. These, however, do not fulfill the criterion if a new intervention reduces the risk rate for the surrogate by some pathway unrelated to the development of a fatal event. One apparent problem with the Prentice approach is that it is very restrictive and thus rarely applicable in practice.

Nevertheless, many authors since then have employed the Prentice criterion in an attempt to validate the choice of surrogate endpoints. Freedman & Graubard [42] employed the Prentice criterion in order to validate surrogate markers in the context of chronic diseases. They reproduced Prentice's mathematical expression, however, in this case for a binary endpoint A, i.e. $A = 0, 1$. Furthermore they suggested the following procedure which may be used to authenticate a surrogate endpoint:

'Step A Test for interaction between intermediate endpoint and treatment. If a significant interaction is found there is strong evidence against [the] criterion (...) and the procedure may stop.

Step B If there is no significant interaction, adopt a no interaction model and test for a treatment effect. If there is a significant treatment effect there is strong evidence *against [the] criterion (...).'*

If a significant interaction term can be found in Step A, then that means that there is strong evidence against the Prentice criterion and we would therefore stop at this step. In a linear logistic regression setting, Step B would mean that the linear logistic model will be adopted

$$\log[\frac{p(D=1|S=s_i,A=j)}{1-p(D=1|S=s_j,A=j)}] = \mu + \sigma_j + \tau_j \tag{5.2}$$

where σ_j is some parameter in the model and τ_j is taken to represent the jth treatment effect. One problem with this validation is that in the case of a statistically significant result, the Prentice criterion will be rejected. However, if the result is not statistically significant, one cannot assume that the criterion is fulfilled.

Begg & Leung [10] also criticize the Prentice approach. Central to their argument are the standardised mean treatment effects on D and S respectively, and the correlation ρ between them. They point out that under the Prentice criterion, *'the standardized effect of the treatment on the surrogate end point must be greater than the standardized effect of the treatment on the true end point by a factor that is proportional to the inverse of the correlation coefficient'* ρ. They propose a new conceptual framework which centres around two principles. The first generally states that the gold standard is represented by the analysis based on the true endpoints. The second principle states that the validity of a surrogate end point should be measured using the probability that trial results which are stipulated from the surrogate marker are 'concordant' with results which would have been obtained had the true primary endpoint been used. The criterion for concordance is arbitrary. One possible definition of concordance would be that both results based upon S and D are significant or not at the 5% level.

Fleming et al. [37] acknowledge in their paper that the Prentice criterion is often of little practical use. However, instead of providing a new measurement they suggest that instead of using surrogate endpoints, one should use auxiliary ones. The auxiliary variables do have a relationship with the treatment and the endpoint of interest but are not used as supplements. Instead, they are used to provide information on missing data from the endpoint of interest. One example is the use of biological marker data such as performance status, immune function and weight change which may provide small improvements in the efficiency of unbiased treatment effect estimates on the primary endpoint. Two approaches to their use employing ideas based on augmented score and augmented likelihood methods are outlined. Nevertheless, from the research they have conducted they conclude that only very modest gains can be made from the use of such 'auxiliary' variables.

5.2.3 Proportion Explained and Relative Effect

As the Prentice criterion is somewhat idealistic, Freedman et al. [42] suggest to measure the proportion of treatment effect explained (PE). A way of estimating this proportion is given by

$$PE = 1 - \frac{\beta_S}{\beta} \tag{5.3}$$

where β_S and β are the estimates of the treatment effect on the final endpoint with and without an adjustment for the surrogate variable calculated from a logistic regression. This approach is also employed by Lin et al. [76]. For a variable satisfying the Prentice criterion this proportion would then be expected to be equal to one.

Methods to simplify the calculations associated with the derivation of the confidence intervals around the PE are described in Chen et al. [15]. Their procedure allows the calculation of treatment effects before and after adjustment for the surrogate simultaneously from a single model. The original methods as proposed by Lin et al. [76] require the estimation of the PE from two separate models which is computationally demanding. Estimates derived using Chen et al. are numerically comparable to the conventional ones. In addition and more importantly the new procedure may also be applied in the setting of multiple-covariate models for the decomposition of the overall treatment effect. This allows the comparison of PE among several surrogate markers.

A significant problem with this approach is that the confidence intervals tend to be very wide. Two other problems are that the proportion explained is not well calibrated as a measure of a proportion and that the measurement is not unique. We could therefore use an alternative measurement

$$PA = 1 - [\frac{\exp(2\beta_S) - 1}{\exp(2\beta) - 1}] \tag{5.4}$$

where 2β is defined as a measurement of the log odds ratio of disease given exposure. The differences in these two measurements range from 0% to 23.69% for varying values of β_S [74]. Lin et al. point out that the employment of this variance formula and the extension they developed requires much larger trials or meta-analyses since precise estimation requires a large value of the ratio of the treatment effect relative to its standard error. Daniels & Hughes [25] further criticize this idea. They highlight another major problem: Should competing mechanisms of action be in operation, the proportion of treatment effect is an erroneous concept since it can take values outside the range from zero to one. Furthermore, PE will tend to be unstable when β is close to zero, a situation that could occur in practice [84].

Buyse & Molenberghs [13] extend the criticism of the proportion explained and go on to develop a new approach which centres around the relative effect RE. If A is defined to be the treatment, RE is the effect of A on D relative to that of A on S, and γ_Z which is the association between S and D after adjustment for A. An intuitive approach for RE is given by

$$RE(T, A, S) = \frac{\beta}{\alpha} \tag{5.5}$$

where α, β and γ are given by the logistic models

$$\ln(\frac{P(S_i = 1|A_i)}{P(S_i = 0|A_i)}) = \mu_{AD} + \alpha A_i$$

and

$$\ln(\frac{P(D_i = 1|A_i)}{P(D_i = 0|A_i)}) = \mu_{AD} + \beta A_i$$

and

$$\ln(\frac{P(D_i = 1|S_i)}{P(D_i = 0|S_i)}) = \mu_{SD} + \gamma S_i$$

respectively. RE can then be interpreted as linking the surrogate and true end point on the population averaged level and γ_Z as describing the subject-specific association between them. As in the case of PE however, the number of observations should be large for RE to be of practical value and hence a meta-analysis is often needed. Two problems associated with RE are the width of its confidence intervals and the fact that it might change with the strength of the association between A and the outcomes itself. Another drawback is that RE is model dependent in its definition. Buyse & Molenberghs have illustrated the use of the RE in their paper employing a study by the Pharmacological

Therapy for Macular Degeneration Study Group run in 1997. Here the effect on vision was tested in two groups of patients, one receiving interferon α and the other a placebo. Buyse & Molenberghs compare the primary endpoint, loss of at least three lines of vision after one year, to the surrogate endpoint of loss of at least two lines of vision after six months. Since their initial analysis fails to provide evidence that the full effect of A on D is mediated through S they then look to compare the proportion explained and relative effect.

	PE	**RE**
Estimate	0.45	0.94
95% CI$_L$	-0.30	0.20
95% CI$_U$	4.35	3.15

Table 5.1: PE and RE with respective confidence intervals for macular degeneration study [74]

Table 5.1 illustrates the width of the confidence intervals around the estimates in this context. We can see that the confidence intervals are very wide for both these approaches. In addition, the confidence interval for PE includes negative values which by definition should not be the case.

5.2.4 Meta-analysis and the degree of correlation between surrogate marker and primary endpoint

As a single trial provides a single estimate of effect sizes on the primary endpoint and a surrogate outcome measure, much attention has been paid to the use of meta-analysis and the degree of correlation between the treatment effect on the surrogate marker and the primary endpoint of interest [143] [64]. Daniels & Hughes [25] use a meta-analytical approach based on Bayesian methods and using bootstrap analysis. They illustrate this method utilizing data from 15 trials in order to explore the association between treatment differences on the development of AIDS or death and the CD4 count. A non-parametric bootstrap is then employed to estimate the correlation between the estimators of the treatment difference on the log hazard ratio for survival and on change in CD4 count. Non-informative priors were placed on the fixed effects and regression coefficients. Three different priors, DuMouchel, shrinkage and a flat prior, were used for the between-study variance. Results from all three of these were similar and showed that CD4 count does not seem to be a good surrogate marker.

Gail et al. [43] discuss the strengths and weaknesses of the meta analytical approach as employed by Daniels & Hughes and extend it. As had already been pointed out by Daniels & Hughes, it may be difficult to specify a joint distribution for D and S and hence Gail et al. introduce separate marginal models for both. However, there are still a number of difficulties with this approach. Firstly, it can be difficult to define the category of drug studies to which a particular study belongs. The parameters of these studies may not only differ due to disparities in drugs used but also because the populations are different. Secondly, it may be the case that there are too few studies with enough reliable information on S and A. Thirdly, individual level data is needed to estimate the distributions which may not be possible to obtain. Fourthly, the precision of estimated treatment effects is limited in a meta-analytical setting. Lastly, the approach by Gail et al. does not include survival analysis.

Lewis has explored this issue further in his thesis [74]. In order to capture the association and dependence of surrogate marker and primary endpoint he uses measures for subject and trial level correlations. Subject level correlation is defined as *'the product moment correlation between the treatment effect on the surrogate variable and the treatment effect on the final endpoint estimated from individual patient data, within a specific randomised trial.'* Trial level correlation (TLC) on the other hand is referred to as *'the correlation between the treatment effect on the surrogate variable and the treatment effect on the final endpoint at the trial level estimated from individual summary parameters from a number of randomised trials.'* Mathematically this relationship can be expressed as

$$\rho^{TLC} = \frac{\sum_{i=1}^{N} [\frac{\widehat{\beta^i_D} - \overline{\beta_D}}{\sigma} * \frac{\widehat{\beta^i}_S - \overline{\beta_S}}{\delta}]}{(N-1)} \tag{5.6}$$

where β^i_D and β^i_S represent the true treatment effect on the final and surrogate endpoint respectively and σ and δ are the variances representing the sample variation of the treatment effect on the surrogate variable and the treatment effect on the final endpoint. Total sample size is given by N. The means $\overline{\beta_D}$ and $\overline{\beta_S}$ are given by

$$\overline{\beta_D} = \frac{1}{N}\sum_{i=1}^{N}\widehat{\beta^i_D}$$

and

$$\overline{\beta_S} = \frac{1}{N}\sum_{i=1}^{N}\widehat{\beta^i_S}.$$

Lewis' approach does not aim to attain the Prentice criterion but instead is designed to be easily interpretable. He illustrates this approach in a binary setting and goes on to extend it to the meta-analytical approach.

Downsides of the meta-analytic framework are described in Molenberghs et al. [84]. They point out that the modelling exercise increases in complexity as the need arises for a joint, hierarchical model for the surrogate and true endpoints. Furthermore, a different model is needed depending on the type of outcome. As a consequence, they introduce a unified theory which avoids the different specifications of trial level surrogacy and individual level surrogacy.

5.2.5 The use of surrogate markers

There are varying views as to what extent surrogate markers should be employed. Ellenberg & Hamilton [31] note that whilst they feel that promising surrogate markers exist in the context of cancer clinical trials, use of these should not preclude long-term survival follow-up. While some randomized studies have demonstrated differences in response rates without any apparent differences in survival time tumour response is often used as a surrogate measure in conjunction with survival as the primary endpoint. The choice of surrogate marker also depends upon the patient population in the study. In a population in which a full recovery is possible surrogates such as disease free survival may be employed. Tumour response is only feasible if all patients have measurable tumour size. Another problem that requires attention regarding the choice of surrogate measures is that whilst a certain treatment may seem beneficial with regard to response in the short run the benefits may be outweighed by adverse long-term effects such as toxicity. Wittes et al. [143] state that whilst a primary endpoint measures the clinical benefit, a surrogate really measures the disease process.

Four potential problems with surrogate markers that are correlated with the endpoint of interest were identified by Fleming [36]. Firstly a surrogate end point may not involve exactly the same pathophysiologic process that results in the clinical endpoint. Secondly, the treatment may only affect the pathway mediated through the surrogate endpoint or thirdly, pathways which are independent of the surrogate. Lastly, the treatment may also affect the true clinical endpoint by unintended mechanisms of action which are independent of the disease process.

Koopmans [68] controversially proposed that surrogate endpoints and biomarkers should be used as support for proof of effectiveness and that clinical endpoints such as survival can be investigated after the drug's introduction. On the other hand, he does, however, point out that often surrogate markers such as response rate are not scientifically substantiated. He also proposes that quality of life should be used as a surrogate marker. The rationale behind this is that often, especially in the case of cancer, patients will not recover fully and drugs should therefore aim to improve their remaining life. This is a difficult issue though since it is not very clear whether patients really do prefer symptom relief to prolongation of survival.

Ellenberg [30] highlights another important advantage of surrogate endpoints. She believes that since trials using surrogates are conducted quicker, they are less likely to be affected by extraneous factors such as dropouts or other forms of non-compliance, and competing risks. Ellenberg raises three issues, which may play a role in deciding on whether to use a surrogate marker. The first one is that the magnitude of the treatment effect on the surrogate should be regarded as being important. Hence, the potential of a surrogate marker is much greater if the effect on it is substantial. Secondly, consideration should be paid to the duration of the effect. Lastly, assessments should be made depending on the severity of the disease as this gives an indication of how quickly a trial needs to be conducted. According to Ellenberg the main problem with the use of surrogates arises when a treatment is not being compared with a placebo but instead when two active treatments are being compared since in this case the biological activity is not the main interest of the study.

Fleming et al. [37] have adopted the theory of the two-stage carcinogenesis model advocated by Moolgavkar & Knudson [85]. Based on this approach, they believe that disease promoter endpoints might prove to be good surrogate endpoints. An example of this would be HIV-specific humoral and cellular immune responses in the development of HIV vaccine trials. For cancer trials they suggest that the use of surrogates such as performance status, weight change, immune function, and toxicity data might be beneficial. However, they ask for caution and illustrate, using the example of cardiac arrhythmia and chronic granulomatous disease, that highly misleading conclusions can be obtained using biological markers as replacement endpoints.

Cancer

Kelloff et al. [62] aim to provide a possible strategy for the application of surrogate markers in the area of cancer chemoprevention development. This strategy involves the identification, validation and use of phenotypic biomarkers and genotypic biomarkers as surrogate markers for cancer incidence. The surrogate end points they propose to use are biomarkers such as proliferation and differentiation indices, gene and chromosome damage and serum biomarkers. Much attention of this paper is paid to the clinical problems in identifying and monitoring these markers, whereby they define the gold standard of surrogate marker validation to be a comparison with cancer incidence reduction. Kelloff et al. believe that through the use of surrogate endpoints the lengths of Phase II and Phase III studies can be reduced to less than 3 years. Kelly [63] also proposes the use of biomarkers such as the PSA (prostate-specific antigen) level but he warns that these need to be used cautiously as some agents have shown to be affecting the PSA level independently of affecting cell growth. He therefore proposes a model in which several agents are tested for PSA level effects and the best is then selected to undergo further testing (see Section 5.3.3).

Day & Duffy [26] have illustrated the use of surrogate endpoints in screening for breast cancer. They come to the conclusion that the use of surrogates in this case leads to a threefold decrease of the variance of the hazard and the availability of results 10 years earlier than through the use of the true endpoint mortality. The surrogate used here is that of predicted mortality which is validated in the paper using the Prentice criterion. Another benefit of the use of predicted mortality perceived by the authors is that of greater expected information contributed by each patient. This is because predicted mortality provides information on the continuous probability of death of each patient whereas mortality is only a binary outcome.

Kelsen [64] concentrates on a meta-analysis run by Buyse & Molenberghs investigating the assessment of colorectal cancer drugs using surrogates. The critical issue at hand here is whether an objective response to treatment is merely associated with better survival, or whether tumour regression (partial or complete) itself lengthens survival. This question has to date not been answered satisfactorily. This means that if we rely on response rate as a surrogate marker, we can reduce drug assessment times substantially but such results need to be treated with caution. It highlights the fact that surrogate markers do not only need to satisfy statistical criteria but also need to be assessed for their biological validity.

A paper by Fleming [35] discusses surrogate markers currently in use in cancer and

AIDS trials. First of all he distinguishes between surrogates used in Phase II and III of trials, since in the first case the primary objective is to establish biological activity whereas in the latter case emphasis is put on evaluating the role a treatment should have in clinical practice. In cancer trials, emphasis has been placed on tumour response as a surrogate for survival. Fleming illustrates that this approach is very unsatisfactory in the colorectal cancer setting. However, despite the limitations of tumour response and biological markers as surrogate endpoints, Fleming does acknowledge their value in providing information. He also believes that the reliability of such surrogate markers will improve as we learn more about the disease process.

HIV / Aids

Fleming [35] describes the case of CD4 counts, which have been widely used as a surrogate for the onset of AIDS or death, but for which it has lately been established that they are not reliable enough. This was found out during the conduct of a comprehensive collection of trials by the June 1993 National Institute of Health sponsored SOTA conference. Here the effect of treatment on primary endpoint and CD4 count was compared and the relationship between CD4 count and survival found to be very unsatisfactory [35]. Fleming has looked at 13 different trials where data on both CD4 count and survival has been collected. He found that there was a very high false positive rate of treatment effect on CD4 count when comparing it to the treatment effect on survival. Furthermore he deduces that if the treatment difference in terms of CD4 count is indeed large, then a prediction on outcome in survival terms is more accurate.

Similar observations concerning the use of CD4 count as a surrogate are made by Lewis [74] in his thesis. He used data from a meta-analysis originally conducted by Daniels & Hughes [25] and found that the CD4 count is not particularly strongly associated with the treatment effect on the event of AIDS or death for individual trials at the individual subject level. In fact, his results indicate that the intervention effect on the onset of AIDS or death is only approximately $\frac{1}{67}$ of the intervention effect on CD4 count. Out of 20 studies, only five predicted a significant intervention effect on the onset of AIDS or death over the course of two years. In two cases out of those the prediction intervals were, however, too wide.

Cardiovascular diseases

O'Connor et al. [86] define four types of endpoints in relation to cardiovascular disease trials. The main endpoint is death; a nonfatal event endpoint is one that a patient wants to avoid such as myocardial infarction or stroke, a true clinical endpoint describes a specific symptom which the patient can feel or which influences his quality of life and a surrogate endpoint is one which the patient cannot feel but which is correlated with death. To date it has not been possible to find a surrogate that correlated perfectly with the main endpoint. O'Connor et al. propose to combine nonfatal or surrogate endpoints with number of deaths in the analysis, thus achieving a higher number of events. However, they also point out that this can be problematic since if the event rate for the surrogate is substantially higher than the event rate for death, any true effect on death may be camouflaged. One example of this is the Dilated Cardiomyopathy trial. Here, death and the need for transplantation were combined. Since transplantation was the largest contributor to the reduction in death, the trial results would have been seriously flawed had those two end points not been analysed separately in the end.

Possible surrogate markers that have been proposed in the past for various studies are hospitalisation, left ventricular mass, ejection fraction, ventricular volumes and maximal oxygen consumption. Still, results on correlation between these surrogate endpoints and the main endpoint have been inconsistent as has been shown in the Cooperative North Scandinavian Enalapril Survival Study and the Veterans Administration Vasodilator-Heart Failure Trial.

5.3 Multiple stage trial designs

This section is concerned with multi-stage trial designs. In general, in this type of design patients are accrued over a period of time after which they are analysed. At this point a stopping rule is applied which decides whether the trial will continue accruing patients and move to the next stage or terminate early. The main aim is to reduce the number of patients required for the trial as well as to reach a conclusion earlier than in a standard parallel group design. This is facilitated by the scope for stopping a trial early for inferiority and / or superiority of an experimental treatment over the control regimen.

Depending on the type of design, information on the effectiveness of treatments is accumulated over all stages in the trial or separately for each stage. However, for trials with a survival-type endpoint, the first method is generally employed. This leads to a level of correlation between the test statistics after each of the stages.

Early on designs were based on sequential methods which are described in Section 5.3.1. These evolved to two-stage and selection designs as outlined in Sections 5.3.2 and 5.3.3. Other adaptations such as the change of the originally envisaged endpoint or recalculation of the sample size during the course of the trial have also been examined. Issues with these types of adaptive designs are dealt with in Posch et al. [97]. However, these are not relevant to this thesis and will therefore not be examined at this point.

5.3.1 Sequential methods

When conducting a clinical trial, the ethical approach is to involve the smallest number of patients possible and use data from these to conduct a valid analysis. Data from many clinical trials is often collected over a comparatively long period of time which gives us scope to stop the trial early in case there are strong indications for or against a certain arm of the trial. This requires the sequential design of a clinical trial since the data need to be analysed at time points during the planned course of the trial. Such methods can also be very flexible in how they are employed as they do not require the same number of subjects in each successive analysis step nor the same number of subjects in each arm. The earliest published account of a sequential clinical trial appeared in 1954 by Kilpatrick & Oldham [66] and was designed for a comparison of bronchal dilators. During the 1950s, 60s and 70s there were regular but few accounts of such trials.

It needs to be emphasized that the interim analyses conducted determine only whether stopping should take place but they do not provide a complete interpretation of the data. In a sequential design setting the reasons due to which the trial may be discontinued include the following:

1) The experimental treatment is obviously worse than the control

2) The experimental treatment is obviously better than the control

3) There is little chance that the experimental treatment is better

Reasons for continuation of the trial may include:

1) A moderate advantage of the experimental treatment appears to be possible. Such an advantage may be clinically worthwhile and thus it is important to estimate its magnitude

as well as we can

2) The event rate observed is low and thus more patients are needed to achieve the desired power [135]

There are two main types of sequential procedures which can be identified. The first one is known as the 'boundaries approach' which includes the sequential probability ratio test and the triangular test and the second derives from a 'repeated significance testing' approach. Sebille and Bellissant [119] have conducted a comparison of these methods using simulation studies. They come to the conclusion that all methods satisfactorily maintain type I and II error rates whilst the triangular test approach seems to be the most satisfactory with regard to substantial reductions in sample size required. Two examples of the conduct of a trial using a triangular test are given in Whitehead [140]. One is an immunosuppression trial conducted at the Fred Hutchinson Cancer Research Center in Seattle and the other a survival study of inoperable lung cancer conducted at the Queen Elizabeth Hospital in Birmingham. In both cases the trial was stopped early due to inferiority of the experimental treatment and thus resulted in a significant reduction in sample size. Figure 5-1 illustrates the sample path from each interim analysis with each circle relating to one analysis point. We can see that at the last interim analysis the sample path has crossed the lower boundary of the christmas tree correction (inner dotted lines) and hence it can be concluded that the experimental treatment is inferior to the control.

5.3.2 Two-stage design based on Ellenberg and Eisenberger

Wieand & Therneau [141] base their two-stage designs upon a discussion by Ellenberg & Eisenberger in 1984. Ellenberg & Eisenberger had presented a two-stage plan for dichotomous or survival outcomes where the time-point of interest was short relative to the accrual period. The main benefit of this plan was the reduction in sample size with only a negligible loss of power. However, they did not give details at the meeting as to how the loss in power may be determined. Their design and the following designs based on it use the same outcome for the first and second stage of the trial. Wieand & Therneau propose a design saying that *'the two-stage rule is to observe n patients on each treatment and stop if at that point the response rate for the test treatment is the same or worse that that for the control treatment.'* If this is not the case then the

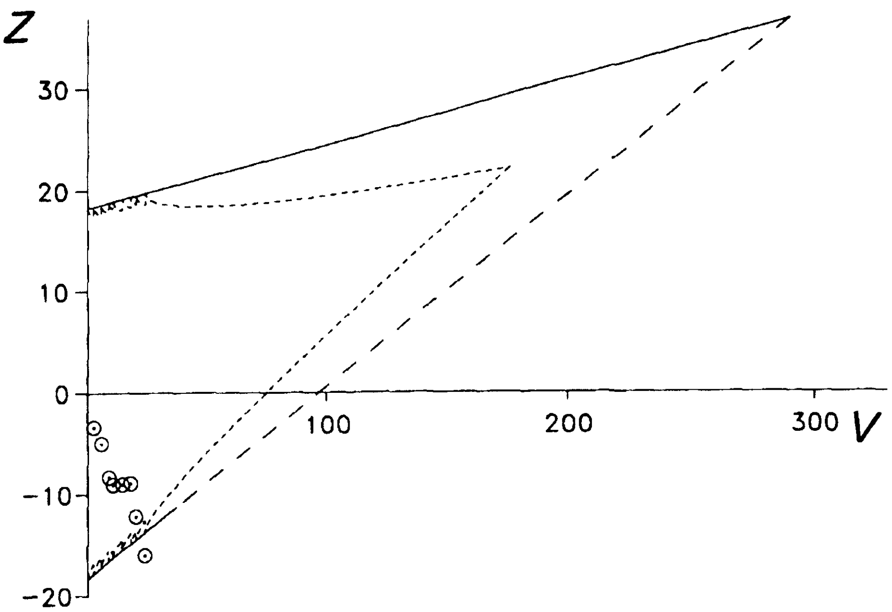


Figure 5-1: Triangular test for inoperable lung cancer study [140]

trial will be continued and a traditional analysis conducted. Underlying their power calculations is the binomial distribution from which they deduce that there is a minimal effect on the power of the study under their two-stage plan, however, the loss of power does increase as the fixed sample power is increased from 0.8 to 0.9. Whilst Ellenberg & Eisenberger computed their sample sizes under the assumption of the Fisher-Irwin exact test, Wieand & Therneau have computed these using an unconditional statistic. Taking p_C to represent the response rate of a control treatment and p_E the response rate of a test treatment, they use

$$N = \frac{0.5(z_\alpha + z_{1-\beta})^2}{[B(p_C) - B(p_E)]^2} \tag{5.7}$$

to calculate the sample size, where $2N$ is the total sample size needed, $z_{1-\alpha}$ and z_β are normal deviates corresponding to a one-sided significance level α and power $1-\beta$ and the angular transformation $B(x) = \arcsin\sqrt{X}$. Several modifications of this formula have also been suggested depending on the type of stopping rule used. These formulations lead on average to sample sizes at around $\frac{3}{4}$ of those for fixed sample sizes. Wieand & Therneau propose to use this design in the case where tumour response or some other binary outcome is of interest.

Thall et al. [134] propose a two-stage design for use in a randomised clinical trial with dichotomous outcomes which is also based on the design suggested by Ellenberg & Eisenberger in 1984. They have aimed to minimize sample sizes after placing constraints on the type I and type II errors. The design is as follows: In Stage 1 $2N_1$ patients are equally allocated to E, experimental treatment, and C, control. X_{E_i} and X_{C_i} denote the binomial success counts and δ_i denotes the difference in the sample proportions for $i = 1, 2$.

$$\widehat{p_{.1}} = \frac{(X_{E1} + X_{C1})}{2N_1} \tag{5.8}$$

and

$$\widehat{p_{..}} = \frac{(X_{E1} + X_{C1} + X_{E2} + X_{C2})}{2N} \tag{5.9}$$

where

$N = N_1 + N_2$ and $q = 1 - p$

in general. The cut-offs y_1 and y_2 are chosen so as to maintain the pre-specified type I and type II error rates. We continue to Stage 2 iff

$$Z_1 = \frac{\delta_1}{(2\widehat{p_{.1}}\widehat{q_{.1}}/N_1)^{\frac{1}{2}}} > y_1 \tag{5.10}$$

otherwise H_0 is accepted and the trial terminated. In Stage 2, an additional $2N_2$ patients are randomised equally to E and C. π denotes the Stage 1 sample proportion. If

$$Z_2 = \frac{\{\pi\delta_1 + (1 - \pi)\delta_2\}}{(2\widehat{p_{..}}\widehat{q_{..}}/N)^{\frac{1}{2}}} > y_2 \tag{5.11}$$

H_0 is rejected, otherwise, accept H_0. Thall et al. then optimize the procedure to obtain minimal sample sizes whilst employing normal approximations to the binomial and numerical approximations. They deduce that there are substantial savings in sample size if the trial is terminated early, i.e. after Stage 1, compared with the fixed sample size approach. The main difference to the earlier design by Ellenberg & Eisenberger is that size and power are pre-defined for a given alternative, and the sample size is minimized under the constraints.

Simon [124] presented a similar design, however, this time for a Phase II clinical trial. His design is optimal in the sense that it achieves the lowest possible sample size in the case where the trial is terminated early. Early termination of the trial can only occur if the experimental drug has activity below a certain cut-off point p_0. In this case the null hypothesis is accepted. Acceptance of H_1 after the first stage is not permitted. The design illustrated is based upon the cumulative binomial distribution. Simon defines the expected sample size to be

$$E(N) = N_1 + (1 - PET)N_2 \tag{5.12}$$

where $N_1$and N_2 denote the sample size at Stages 1 and 2 respectively and

$$PET = B(r_1; p, N_1)$$

denotes the probability that r_1 or fewer responses are observed in Stage 1 and hence the trial is terminated then. He goes on to determine optimal designs for pre-specified error probabilities and concurrs that the optimal two-stage design does not necessarily minimize the maximum sample size N subject to the error probability constraints. A comparison of possible optimal and minimax designs can be found in his paper. The minimax design seems to be more attractive in the case where the expected sample size is small and accrual rates are low. A reason for that is that under the optimal design this will coincide with a very small first stage. However, in the case of heterogeneous populations this may not be desirable. He continues with a comparison of his designs to those of other authors such as Fleming, coming to the conclusion that his design achieves lowest expected sample sizes for several error combinations, but points out that a major problem with such comparisons is that two designs are often not equivalent with regard to the error probabilities. An extension of this method to a Bayesian decision-theoretic setting is provided by Jung et al. [61].

Following on from Simon, Chen & NG [16] use his optimal and minimax designs and apply them to a flexible setting. They define the expected sample size

$$E(N) = N_1 + (1 - PET)(N_2 - N_1) \tag{5.13}$$

the average probability of early termination (APET), the average total probability of rejecting treatment (ATPRT) and the Average expected sample size (AEN). The flexible design allows the ATPRT to be between $1-\alpha$ and β. Therefore, when applying this to the design of a head and neck cancer trial the expected sample size is reduced compared to the one under Simon's fixed design for the minimax option. Under the optimal design option, both calculations give nearly the same answers. One disadvantage of this design is that it does not allow for early termination of the trial if there is a long run of failures at the start. To combat this problem, a range of three-stage designs have been suggested, for example Ensign et al. (1994) [32]. Here the sample size is very closely linked to the power $1-\beta$ of the study, i.e.

$$E(N)(p) = N_1 + N_2\{1-\beta_1(p)\} + N_3\{1-\beta_1(p)-\beta_2(p)\} \tag{5.14}$$

However, since all of these designs are for Phase II studies, they deal with comparatively small sample sizes and therefore are not practical for our Phase III trials.

5.3.3 Two-stage selection designs

So far only designs to compare one experimental treatment to a control have been introduced. Nevertheless, two-stage designs can be used to select a promising treatment from a number of different treatments. At the same time they retain the advantage of reduced sample size identified in sequential designs, especially when most of the agents are observed to have little or no activity. Central to all of these designs is that we start with several experimental treatments out of which, in comparison with the control, the most promising is chosen [105] [124] [34]. *'Multiple stage plans are specified by the number of units examined at each stage, the number of stages, and the acceptance points and the rejection points associated with each stage'* [117].

One such design which can be used to decide between several experimental treatments of interest was proposed by Thall et al. [133]. A year before that Thall et al. published a paper employing a two-stage design for pilot studies [132]. The central idea to both papers is that the highest success rate among the experimental treatments is identified. This is advantageous from an ethical viewpoint since exposure to ineffective therapies is minimized while resources may be allocated to test more treatments compared to the standard parallel two-arm designs. Methodology of the second paper is now described. If the success rate of the best experimental treatment falls below a certain cut-off value than the trial is terminated. If it is above this value, the trial proceeds to Stage 2, where the 'best' experimental treatment is compared to a control. The advantages of this design are that it has both high power and a high probability of termination should no experimental treatment be superior to the control. The type of trials considered here are based in a binomial setting with either success or failure as a possible outcome. Expected sample size depends in this case on the success probabilities and has been identified as

$$E(N) = kN_1 + \pi 2N_2 \tag{5.15}$$

where π is the probability of continuing to Stage 2. In addition to the overall power, Stage 2 power may be specified. Whilst the binary assumption may be relaxed, a major problem with this approach could stem from the fact that decisions need to be made on observations relatively soon after treatment commences. Care also needs to be taken in determining the cut-off for Stage 1 since if this is too high possible improvements on the control may be missed. It is therefore most favourable in the case where at least one treatment is expected to display a significant advantage. This design may be adapted by including the control in the first stage.

A similar design was suggested by Schaid et al. [110]. The main differences to the design described above are that it allows for more than one experimental treatment to be taken forward to Stage 2 and that in the case of a substantial survival advantage of one of the experimental treatments over the control the trial may also be terminated early. Hence two boundaries y_1 and y_2 are identified before the start of the trial with y_2 being the upper boundary identifying a substantial survival advantage. In the case of this design y_1 is based on clinical judgement rather than optimization. Each of the experimental treatments are being compared to the control which calls for the definition of α to be the pairwise alpha-error for each comparison. Schaid et al. have identified the expected total sample size as being

$$E(N) = (k+1)N_1 p_0 + \sum_{j=1}^{k} \{N_2(j+1) + N_1(k-j)\} p_j \tag{5.16}$$

where N_1 is the sample size in Stage 1, N_2 the sample size in Stage 2, p_0 the probability of stopping accrual at the first stage and p_j the probability that accrual will continue for the standard treatment and j of the experimental treatments. The design is then termed optimal if it achieves the lowest expected total sample size when the null hypothesis is true given α and $(1-\beta)$. During the course of Monte-Carlo simulations using a FORTRAN program the authors found that *'the rule offers the largest reduction in $E(N)$ when the deaths are occurring quickly relative to the accrual rate and when there are several experimental treatments'.*

Liu et al. [78] have criticized the above approach in that they do not believe that it fits into a cancer trial setting. According to the authors treatments for cancer are usually first tested on advanced tumour patients and once a promising treatment has been identified this is then compared to a control using patients which are relatively early in their disease stage. Hence they believe that a design which allows for the progression of more than one experimental treatment into the second stage is at odds with this approach. Instead they propose a design which has a fixed sample size, giving no possibility of an excessive number of patients and only allowing the progression of at most one experimental treatment into the second stage. This approach is based upon the Cox model and they advocate the usage of it in Pilot studies.

To circumvent the problem that Liu has pointed out above, Simon et al. [126] have chosen to only use patients from the second stage in their analysis after Stage 2. They have studied two possible types of design, one that includes the control in the first stage and one that does not [132] [133]. The authors believe that these types of designs are most applicable when it is very unlikely that there will be more than one treatment that is better than the control and when the patient numbers available are too small to evaluate more than one experimental regimen.

The above approaches by Thall et al. [133] and Schaid et al. [110] are generalised by Stallard & Todd [128] in two ways. Firstly, through the use of the efficient score as a test statistic the method becomes applicable to binary, normally distributed or failure time responses and furthermore allows the incorporation of covariate information at both the interim and final analyses. Secondly, they consider a sequential trial setting in which a number of interim analyses comparing the selected and control treatments are performed. However, it is required that at most one experimental treatment is selected at the first interim analysis. Thus if there is a group of treatments which are superior to the control and one wishes to select the best out of those, this method is not applicable since it would be desirable to only drop ineffective treatments at an early stage.

5.4 Surrogate markers and two-stage designs combined

In this thesis we propose the combination of both surrogate markers and a multi-stage design. Suggestions of such a combination have been made by a few authors. Kelly [63] emphasizes the need for this combination when he says that *'Only with the appropriate selection of disease state, trial design, and endpoints will we be able to select the most promising regimens to move forward'.*

Flandre & O'Quigley [34] have considered this type of design. Their definition of surrogacy is *'a response variable of prognostic value obtained during follow up, which indicates an objective progression of disease'.* Using this definition, they closely follow the Prentice criterion. Their design is as follows:

Stage 1: all patients are followed to the primary endpoint and information on the surrogate is collected in order to evaluate the strength of the relationship between surrogate endpoint and survival

Stage 2: follow up is terminated when patients reach the surrogate event.

The validity of the surrogate variable is tested using a standard likelihood ratio test. Information collected during the first stage consists of either survival time and the surrogate variable or just survival time, depending on whether the surrogate event occurs before the death of a patient or not. Because of the way in which this trial is designed, Stage 1 could either be part of the trial or an earlier trial could be used. The survivorship model presented is based upon an earlier model developed by Slud & Rubinstein [127]. The authors give two examples of trials for resected lung cancer from which the sample sizes N_1 and N_2 of Stages 1 and 2 were drawn a posteriori. Relapse has been considered as a time-dependent surrogate endpoint. The main problem with the approach is that it is based upon the Prentice criterion which, as described above, is very rarely satisfied in practice. What we also find problematic is the use of the surrogate endpoint in the second stage. We therefore propose a design as illustrated in the next chapter.

5.5 Summary

A growing number of trials today employ surrogate markers, either to complement the information available or to replace the primary endpoint at one stage in the trial. Recent trials have, however, shown that this can be a dangerous practice as often the relationship between treatment, surrogate marker and primary endpoint has not been well established.

A very prominent example of this is the use of CD4 counts as a surrogate marker for death in the case of HIV clinical trials. Methods such as the Prentice criterion, the proportion explained or the relative effect, and measures of correlation founded in meta-analysis aim to provide a basis for the establishment of such a relationship, if it exists. Today many authors have recognized that the Prentice criterion is very rarely attainable in practice. Nevertheless, this does not make it obsolete but instead provides an ideal situation which every surrogate/primary endpoint relationship should be compared to. A good surrogate should satisfy two properties. Firstly, the surrogate endpoint must predict the primary endpoint on an individual patient level. Secondly, the effect of a treatment on a surrogate endpoint must predict the effect of that treatment on the primary endpoint. Unfortunately, we are rarely in this position for most of our common diseases. However, whenever employing surrogate endpoints we need to be cautious since any evidence of a relationship between the surrogate and primary endpoint will have been derived from earlier trials. In some cases the strength of this relationship may change when a new therapy regimen is used.

With the arrival of new advances in molecular biology and the ever increasing knowledge about our organism, potential agents which may improve patient outcomes are increasingly becoming available. Methods such as two-stage selection designs are therefore necessary in order to get the new drugs to the patients as soon and as safely as possible. Once a relationship has been established between the effect of treatment on the primary endpoint and on the surrogate, surrogate markers incorporated into the two-stage design could, in principle, provide a significant reduction in both trial time and sample size.

Chapter 6

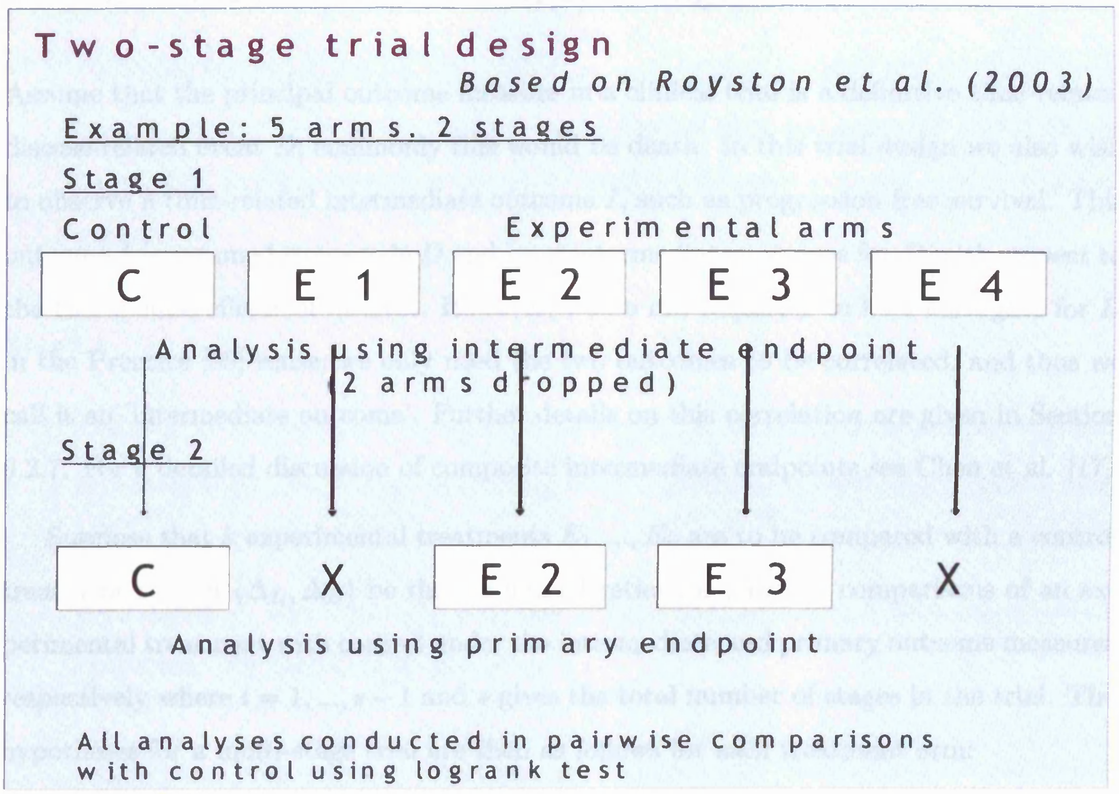
A multi-stage design

6.1 Introduction

Royston et al. [103] proposed a design employing an intermediate outcome in the first stage of a two-stage trial with multiple research arms. Such an intermediate outcome is not required to be a perfect surrogate for the final outcome in the Prentice sense but rather it is essential that the effect sizes of the new treatment on the intermediate and final outcome measures are related. The main aims are to reject as quickly and reliably as possibly any new therapies that are unlikely to show a worthwhile effect in terms of the primary outcome measure and to continue testing those therapies which are likely to show such an effect.

The design itself is based on eliminating inferior treatments at an early stage, and hence allowing through to the second stage only those treatments which show a predefined degree of advantage against the control treatment. In the first stage, the experimental arms are compared pairwise with the control according to the intermediate outcome measure. Treatment arms that survive this comparison then enter a second stage of patient accrual which culminates in pairwise comparisons against the control based on the primary endpoint. An example of such a trial with four experimental arms and one control over two stages is given in Figure 6-1.

The overall operating characteristics in this design are computed from the Stage 1 and 2 type I and II error rates as well as the correlation between treatment effects on the intermediate and primary outcome measures. An important assumption is that the log hazard ratios on the intermediate and primary outcome follow a bivariate Normal



Two-stage trial design

Based on Royston et al. (2003)

Example: 5 arms, 2 stages

Stage 1

Control

Experimental arms

C E1 E2 E3 E4

Analysis using intermediate endpoint (2 arms dropped)

Stage 2

C X E2 E3 X

Analysis using primary endpoint

All analyses conducted in pairwise comparisons with control using logrank test

Figure 6-1: Two-stage design based on Royston et al.

distribution. We may estimate this correlation from previous trials.

Recently ICON5, a trial comparing several ovarian cancer treatments, which employs this methodology, has been conducted at the Medical Research Council, London, together with collaborators in the USA, Italy and Australia. Furthermore, a number of trials in a variety of cancer sites are currently in the planning stages. However, two of these, STAMPEDE and ICON6, require more than one stage using the intermediate endpoint which has led to the work presented in this chapter. More information on these two trials is provided during the course of this chapter. This extension is important especially when dealing with new agents in cancer trials because very little is known about the effect of these drugs, both on their own and in conjunction with chemotherapy agents for example. Thus we want to allow for very early looks at which we can reject agents which show either no promising or even an adverse effect.

Thus, in this chapter the 2003 design is extended to allow for more than two stages in the trial. Mathematical details for the calculation of sample size in the two-stage as well as the extension to the multi-stage setting are provided in Section 6.2. An analysis of some of the assumptions underlying the calculations is provided in Section 6.3.

6.2 Extension to more than two stages

Assume that the principal outcome measure in a clinical trial is a definitive time-related disease-related event D; commonly this would be death. In this trial design we also wish to observe a time-related intermediate outcome I, such as progression free survival. This outcome I is assumed to precede D and is an intermediate outcome for D with respect to the therapeutic effects of interest. However, we do not require I to be a surrogate for D in the Prentice [99] sense; we only need the two outcomes to be correlated, and thus we call it an 'intermediate outcome'. Further details on this correlation are given in Section 6.2.7. For a detailed discussion of composite intermediate endpoints see Chen et al. [17].

Suppose that k experimental treatments $E_1, ..., E_k$ are to be compared with a control treatment C. Let (Δ_{I_i}, Δ_D) be the log hazard ratios for pairwise comparisons of an experimental treatment with control under the intermediate and primary outcome measures respectively where $i = 1, ..., s-1$ and s gives the total number of stages in the trial. The hypotheses for a multi-stage trial are then as follows for each treatment arm:

$$H_0 : (\Delta_{I_1}, \Delta_{I_2}, ..., \Delta_{I_{s-1}}, \Delta_D) = (\Delta^0_{I_1}, \Delta^0_{I_2}, ..., \Delta^0_{I_{s-1}}, \Delta^0_D)$$

and

$$H_1 : (\Delta_{I_1}, \Delta_{I_2}, ..., \Delta_{I_{s-1}}, \Delta_D) = (\Delta^1_{I_1}, \Delta^1_{I_2}, ..., \Delta^1_{I_{s-1}}, \Delta^1_D)$$

An experimental treatment is deemed advantageous iff $\Delta^1_{I_i} < \Delta^0_{I_i}$ and $\Delta_{I_i} < 1$, where $i = 1, ..., s-1$, as well as $\Delta^1_D < \Delta^0_D$ and $\Delta_D < 1$.

Define e_{I_i} as the total number of I events in the control arm after Stage i in the trial and e_D as the total number of D events in the control arm. The trial then proceeds in s stages as outlined:

Stage 1 to Stage s-1

1. Define a critical value for the rejection of H_0, δ_{I_i}, so that an experimental treatment E will pass to Stage $i+1$ if the estimate of the log hazard ratio $\widehat{\Delta_{I_i}}$ is found to be smaller than $\ln \delta_{I_i}$.

2. Randomise N_i patients, $i = 1, ..., s-1$, between the control and k experimental arms. Patients are distributed using an equal allocation ratio in most cases. N_i

needs to be sufficient to expect e_{I_i} I events in the control arm.

3. Compute the hazard ratios using the Cox proportional hazards [92] for the I event once e_{I_i} events have been observed in the control arm and compare this with the value of δ_{I_i} to decide whether the experimental treatment arm will pass to the next stage.

Stage s

1. Define a critical value for the rejection of H_0, δ_D, so that E is deemed to be superior to the control if the estimate of the log hazard ratio $\widehat{\Delta_D}$ is found to be smaller than $\ln \delta_D$.
2. Randomise an additional N_D patients to both the control and each experimental treatment arm carried over into Stage s.
3. Compute hazard ratios for D again using the Cox proportional hazards once e_D events have been observed among the control arm.

The event numbers are cumulative across all stages. Assumptions made during the course of this approach are the proportionality of the hazards and the standard multivariate normal distribution of the log hazard ratios.

6.2.1 Sample size and power calculations

The overall type I error probability, the probability of falsely rejecting H_0, within this framework is given by

$$\begin{aligned}\alpha &= P(\widehat{\Delta_{I_1}} < \ln \delta_{I_1}, \widehat{\Delta_{I_2}} < \ln \delta_{I_2}, ..., \widehat{\Delta_{I_{s-1}}} < \ln \delta_{I_{s-1}}, \widehat{\Delta_D} < \ln \delta_D | H_0) \\ &= \Phi_s(z_{\alpha_{I_1}}, z_{\alpha_{I_2}}, ..., z_{\alpha_{I_{s-1}}}, z_{\alpha_D}, R)\end{aligned}$$

where $\Phi_s(.)$ denotes the standard multivariate Normal distribution function and $z_{\alpha_{I_i}}$ and z_{α_D} are normal deviates corresponding to a one-sided significance level α. The standard multivariate density function is given by

$$f_s(x_0, x_1, ..., x_p) = (2\pi)^{-(p+1)/2} |d(A)|^{1/2} \exp(-\frac{Q}{2})$$

Here $Q = x'Ax$ and A is a symmetric matrix of s rows and columns which is defined as the inverse of the correlation matrix R since the standard multivariate normal distribution has unit variance. Thus

$$A^{-1} = R = \left\{ \begin{array}{cccc} 1 & \rho_{12} & \dots & \rho_{1s} \\ \rho_{21} & 1 & \dots & \rho_{2s} \\ \dots & \dots & \dots & \dots \\ \rho_{s1} & \rho_{s2} & \dots & 1 \end{array} \right\}$$

In this model the correlation matrix R depends upon three things in particular. Firstly, it is dependent upon the time at which e_{I_i} events have been accrued in the control arm during Stage i. Secondly, the interval between that time point and the point in time at which e_D events have occurred is of importance. Lastly, there will be a built-in correlation if event I is a composite which includes D. One example of this would be the use of progression free survival as an intermediate marker.

Assuming that we have specified the type I error, power and the log hazard ratios Δ_{I_i}, Δ_D in all stages, we need to calculate the cut-off δ_{I_i} as well as the number of control arm events needed in all stages. It is intuitive that $\ln \delta_{I_i}$ should lie between $\Delta^0_{I_i}$ and $\Delta^1_{I_i}$. Let Φ^{-1} denote the standard Normal distribution function and $((\sigma^0_{I_i})^2, (\sigma^0_D)^2)$ the variances of the estimated log hazard ratios $(\widehat{\Delta_{I_i}}, \widehat{\Delta_D})$ under H_0. Hence by definition for all stages where I is the outcome

$$\begin{aligned} z_{\alpha_{I_i}} &= \frac{\ln \delta_{I_i} - \Delta^0_{I_i}}{\sigma^0_{I_i}} \\ &= \Phi^{-1}(\alpha_{I_i}) \end{aligned}$$

and

$$z_{\alpha_D} = \frac{\ln \delta_D - \Delta^0_D}{\sigma^0_D} = \Phi^{-1}(\alpha_D)$$

We define the overall power across all stages to be $1 - \beta$

$$\begin{aligned} 1 - \beta &= P(\widehat{\Delta_{I_1}} < \ln \delta_{I_1}, \widehat{\Delta_{I_2}} < \ln \delta_{I_2}, ..., \widehat{\Delta_{I_{s-1}}} < \ln \delta_{I_{s-1}}, \widehat{\Delta_D} < \ln \delta_D | H_1) \\ &= \Phi_m(z_{1-\beta_{I_i}}, z_{1-\beta_{I_2}}, ..., z_{1-\beta_{I_{s-1}}}, z_{1-\beta_D}, R) \qquad (6.1) \end{aligned}$$

Let $((\sigma^1_{I_i})^2, (\sigma^1_D)^2)$ denote the variances of the estimated log hazard ratios $(\widehat{\Delta_{I_i}}, \widehat{\Delta_D})$

under H_1. Again by definition for all stages where I is the outcome

$$\begin{aligned} z_{1-\beta_{I_i}} &= \frac{\ln \delta_{I_i} - \Delta^1_{I_i}}{\sigma^1_{I_i}} \\ &= \Phi^{-1}((1-\beta)_{I_i}) \end{aligned} \tag{6.2}$$

and

$$z_{1-\beta_D} = \frac{\ln \delta_D - \Delta^1_D}{\sigma^1_D} = \Phi^{-1}((1-\beta)_D)$$

[103]. The quantities $(1-\beta)_{I_i}$ and $(1-\beta)_D$ may be interpreted as the probability of an effective new treatment passing to the next stage when the alternative hypothesis is true and the power of the final significance test at Stage s respectively. As all s tests need to be passed, the overall power cannot exceed $(1-\beta)_{I_i}$ or $(1-\beta)_D$. Following a similar argument, the overall type I error may not exceed either α_{I_i} or α_D.

According to Tsiatis, 1981, [137] we can approximate the variance under H_0 using the following formula

$$(\sigma^0_{I_i})^2 = (\sigma^1_{I_i})^2 = \frac{2}{e_{I_i}} \tag{6.3}$$

and

$$(\sigma^0_D)^2 = (\sigma^1_D)^2 = \frac{2}{e_D}$$

where the es are the number of intermediate and primary outcome measure events required. Using this approximation and the type I and II error formulae given above we can calculate the number of events as

$$e_{I_i} = \frac{2}{(\sigma^0_{I_i})^2} = \frac{2}{(\sigma^1_{I_i})^2}$$

Following an argument similar to Royston et al. [103], who evaluate the above expression for the two-stage case, we find that

$$\sigma^0_{I_i} = \frac{\Delta^1_{I_i} - \Delta^0_{I_i}}{z_{\alpha_{I_i}} - z_{1-\beta_{I_i}}}$$

finally

$$e_{I_i} = \frac{2(z_{\alpha_{I_i}} - z_{1-\beta_{I_i}})^2}{(\Delta^1_{I_i} - \Delta^0_{I_i})^2} \tag{6.4}$$

and

$$e_D = \frac{2(z_{\alpha_D} - z_{1-\beta_D})^2}{(\Delta^1_D - \Delta^0_D)^2} \tag{6.5}$$

6.2.2 Relationship between number of events and trial time

After the trial has been launched, patients will be recruited gradually over time. If none of the experimental arms passes to the next stage, recruitment (but not necessarily follow-up) will cease. In the case where one or more experimental arms pass on to the next stage we assume that recruitment will continue at the same rate as in Stage 1. All available patients are then randomised between the remaining arms. This means that the more arms continue to the next stage the fewer patients each arm will receive and hence the longer and more expensive the trial will be.

Define $R(t)$ as the number of patients recruited to the control arm by time t. We can now take

$$r(t) = \frac{dR}{dt} \tag{6.6}$$

to represent the instantaneous recruitment rate of patients. The expected number of control arm survival events at time t in a simple parallel group trial is then given by

$$e(t) = \int_0^t F(t-u)r(u)du \tag{6.7}$$

where $F(t) = 1 - S(t)$, $S(t)$ being a survivorship function in the control group [103]. We assume this survival distribution to be exponential, i.e. $S(t) = e^{-\lambda t}$, since this is often used in sample size derivations and allows for tractability. If we set $r(t) = r$ with r being a constant, i.e. assuming a constant recruitment rate, we arrive at the expression

$$\begin{aligned} e(t) &= \int_0^t [1 - S(t-u)]r(u)du \\ &= \int_0^t [1 - e^{-\lambda(t-u)}]rdu \\ &= r\int_0^t [1 - e^{-\lambda(t-u)}]du \\ &= r[u - \frac{1}{\lambda}e^{-\lambda(t-u)}]_0^t \end{aligned}$$

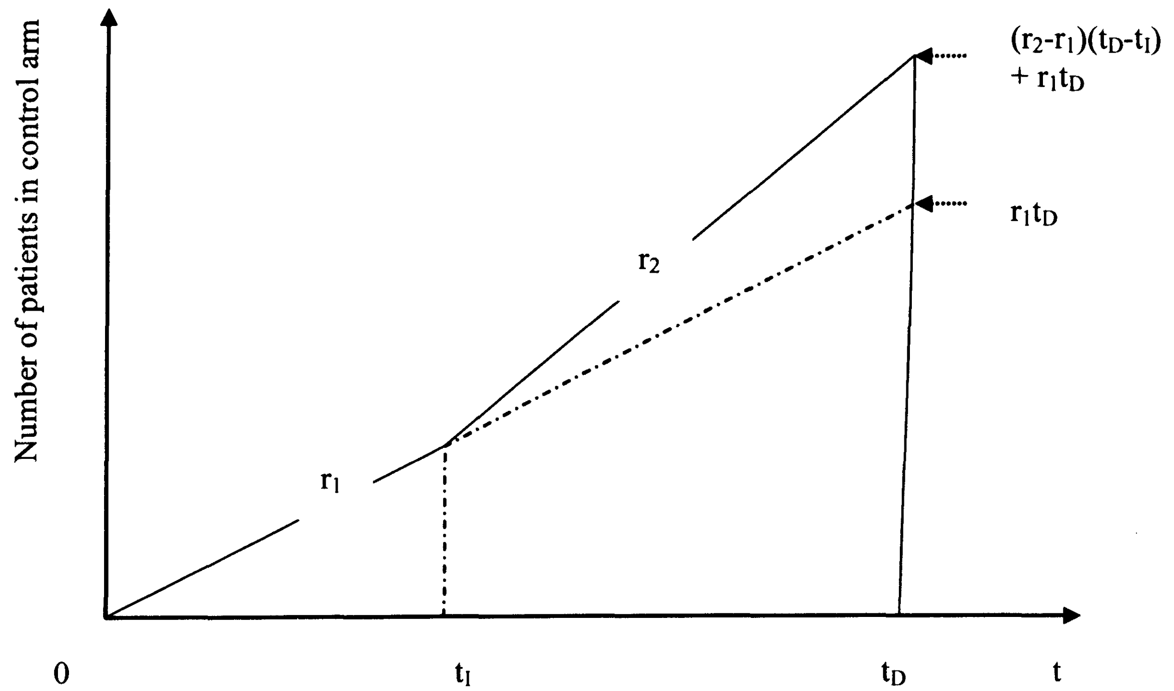


Figure 6-2: Accrual to control arm in a two-stage trial

Hence

$$e(t) = r(t - \frac{1 - e^{-\lambda t}}{\lambda}) \tag{6.8}$$

We now consider the total trial duration t_D which is the time at which e_D events are accrued in the control arm. Let t_I denote the time at which in a two-stage trial Stage 1 terminates, that is, the time to accrue e_I I-events in the control arm. Assume that the (constant) recruitment rates per arm per unit time are r_1 and r_2 in Stages 1 and 2 respectively where $r_2 \geq r_1$. Furthermore, let $F_D(t)$ be the distribution function for D-events in the control arm. Considering Figure 6-2 we can see that e_D may be calculated in two separate integrals. The first one calculates the area under the triangle from 0 to t_D and up to the first arrow, the second integral calculates the area of the smaller triangle lying above that. Thus combining this knowledge with Expression 6.7 above we get

$$e_D = r_1 \int_0^{t_D} F_D(t_D - u)du + (r_2 - r_1) * \int_0^{t_D - t_I} F_D(t_D - t_I - u)du$$

[103]. Since $S(t)$ is assumed to follow an exponential distribution the formula calculated

above may be modified to give

$$\begin{aligned}
e_D &= r_1\int_0^{t_D}[1-S(t_D-u)]du+(r_2-r_1)*\int_0^{t_D-t_I}[1-S(t_D-t_I-u)]du \\
&= r_1[u-\frac{1}{\lambda}e^{-\lambda(t_D-u)}]_0^{t_D}+(r_2-r_1)[u-\frac{1}{\lambda}e^{-\lambda(t_D-t_I-u)}]_0^{t_D-t_I} \\
&= r_1(t_D-\frac{1-e^{-\lambda_D t_D}}{\lambda_D})+(r_2-r_1)(t_D-t_I-\frac{1-e^{-\lambda_D(t_D-t_I)}}{\lambda_D}) \qquad (6.9)
\end{aligned}$$

The number of intermediate endpoint events needed is then given by

$$e_I=r_1(t_I-\frac{1-e^{-\lambda_I t_I}}{\lambda_I})$$

We can extend this derivation to multi-stage trials with more than two stages to give

$$\begin{aligned}
e_D &= r_1\int_0^{t_D}F_D(t_D-u)du+(r_2-r_1)*\int_0^{t_D-t_{I1}}F_D(t_D-t_{I_1}-u)du+... \\
&\quad +(r_s-r_{s-1})*\int_0^{t_D-t_{I_{s-1}}}F_D(t_D-t_{I_{s-1}}-u)du \\
&= r_1(t_D-\frac{1-e^{-\lambda_D t_D}}{\lambda_D})+(r_2-r_1)(t_D-t_{I_1}-\frac{1-e^{-\lambda_D(t_D-t_{I_1})}}{\lambda_D})+... \\
&\quad +(r_s-r_{s-1})(t_D-t_{I_{s-1}}-\frac{1-e^{-\lambda_D(t_D-t_{I_{s-1}})}}{\lambda_D})
\end{aligned}$$

Since we take k to represent the number of experimental treatments used in the study, Stage 1 will consist of $k+1$ treatments. Now we assume that the rate of accrual in Stage 1 will be equal to the rate of accrual in Stage 2 of our study [103]. Hence

$$(k+1)r_1=(k_2+1)r_2$$

where k_2 is the number of experimental treatments at Stage 2 of the trial. This holds for all s stages of the trial. Thus the total number of patients needed in the trial is given by

$$(k+1)r_1t_D \qquad (6.10)$$

6.2.3 Algorithm used for Stata program

Since Formulae 6.4 and 6.5 for the number of events e_{I_i} and e_D are based on an estimate of the variance under H_0, they will slightly underestimate the true sample size needed to achieve power $1-\beta$. Hence the computer program available for Stata is based on the following algorithm in order to adjust for this:

1. Calculate the number of control I events e_{I_i} needed based on Formula 6.4.
2. Calculate the critical log hazard ratio $\ln \delta_{I_i} = \Delta^0_{I_i} + z_{\alpha_{I_i}} * \sqrt{\left(\frac{p}{e_{I_i}}\right)}$
3. Calculate the time t_{I_i} needed to run the trial until the end of Stage i using Section 6.2.2
4. Calculate the number of events in the experimental arm(s) $e^*_{I_i}$ under H_1 by the end of Stage i using an exponential survival distribution
5. Calculate power for Stage i which can be achieved under e_{I_i} and $e^*_{I_i}$
   (a) If power is less than needed, replace e_{I_i} by $e_{I_i} + 1$ and rerun Steps 2. to 5.
   (b) If power is as desired, terminate the algorithm

6.2.4 Other accrual mechanisms

Section 6.2.2 has employed a uniform recruitment pattern to aid calculations. However, in a number of instances it is more appropriate to assume a different recruitment pattern (see Chapter 3, Section 3.5), such as an exponential accrual path. An example of such recruitment curves is given in Figure 3-3. Taking $r(u)$ as exponential with parameter a gives

$$\begin{aligned} e(t) &= \int_0^t e^{-au}[1 - e^{-\lambda(t-u)}]du \\ &= [-\frac{1}{a}e^{-au} - \frac{1}{\lambda - a}e^{-au-\lambda(t-u)}]_0^t \\ &= \frac{1}{a} - \frac{\lambda}{a(\lambda - a)}e^{-at} + \frac{1}{\lambda - a}e^{-\lambda t} \end{aligned}$$

Another possibility is to take $r(u)$ as piecewise linear, i.e.

$$r(u) = \begin{cases} a_1 + b_1 r & 0 < r \leq R_1 \\ a_2 + b_2 r & R_1 < r \leq R_2 \\ \dots & \dots \\ a_j + b_j r & R_{j-1} < r \leq R_j \\ \dots & \dots \end{cases}$$

This expression may then be fed back into the derivation of $e(t)$. Figure 6-3 gives examples of such accrual patterns with a target accrual of 103 patients.

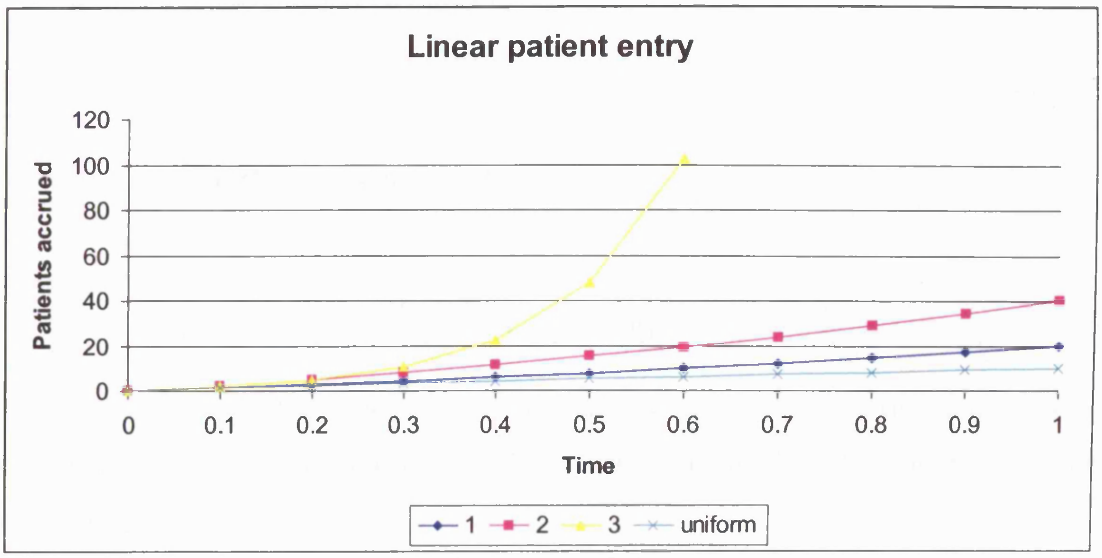


Figure 6-3: Piecewise linear patient accrual patterns 1 - a1 and b1 are equal to one in first time increment and increase by one in each increment, 2 - a1 and b1 are equal to two in first time increment and increase by two in each increment thereafter, 3 - a1 and b1 are equal to three in first time increment and double in each increment thereafter, uniform - uniform patient entry pattern starting at 1 in first time increment

6.2.5 Stopping accrual at a pre-specified time-point

So far we have assumed that recruitment may continue until the end of the trial, if needed. However, there may be situations where it is more appropriate to stop recruiting to the trial earlier on and after that only follow patients up. This allows one to restrict the required sample size. Furthermore, when implementing these sample size calculations in a computer program such as Stata, the following derivations are needed to account for those treatment arms which do not proceed on to the next stage, i.e. to which no further accrual takes place, but which are still being followed up.

Let t^* denote the time at which accrual is stopped. Starting from a two-stage trial only, there are two possible scenarios to consider. First we will look at the case $t^* > t_I$ where t_I denotes the end of the first stage. Here we only need to consider the primary outcome measure as the trial will proceed as before until the end of Stage 1. Denote the number of patients at risk at t^* by $N_D(0,t^*)$

$$\begin{aligned}
N_D(0,t^*) &= (\#\text{ patients recruited by } t^*) - (\#\text{ events by } t^*) \\
&= (\#\text{ patients recruited by } t_I) + (\#\text{ patients recruited between } t_I, t^*) \\
&\quad -(\#\text{ events by } t^*) \\
&= r_1 t^* + (r_D - r_1)(t^* - t_I) - e_D(0,t^*)
\end{aligned}$$

where

$$e_D(0,t^*) = r_1(t^* - \frac{F_D(t^*)}{\lambda_D}) + (r_2 - r_1)(t^* - t_I - \frac{F_D(t^* - t_I)}{\lambda_D})$$

which relates back to Formula 6.9. The probability of an event in the time interval between t^* and t_D is given by $F_D(t_D - t^*)$ and hence the number of events in this time interval is given by

$$e_D(t^*, t_D) = N(t^*, t_D)$$

The total number of events up to t_D is then given by

$$e_D(0,t^*) + e_D(t^*, t_D)$$

The second case to consider is $t^* < t_I$. First we derive the number of patients at risk based on the of primary outcome events

$$\begin{aligned}
N_D(0,t^*) &= r_1 t^* - e_D(0,t^*) \\
&= \frac{r_1}{\lambda_D} F_D(t^*)
\end{aligned}$$

The number of events in the interval 0 to t^* is given by

$$e_D(0,t^*) = r_1(t^* - \frac{F_D(t^*)}{\lambda_D})$$

and

$$e_D(t^*, t_D) = \frac{r_1}{\lambda_D} F_D(t^*) F_D(t_D - t^*)$$

For the intermediate events the number of patients at risk is given by

$$N_I(0,t^*) = \frac{r_1}{\lambda_I} F_I(t^*)$$

and

$$e_I(0,t^*) = r_1(t^* - \frac{F_I(t^*)}{\lambda_I})$$

as well as

$$e_I(t^*,t_I) = \frac{r_1}{\lambda_I} F_I(t^*) F_I(t_I - t^*)$$

This framework is easily extended to more than two stages. For the case that $t^* > t_{I_{s-1}}$ we have that

$$N_D(0,t^*) = r_1 t^* + (r_2 - r_1)(t^* - t_{I_1}) + ... + (r_s - r_{s-1})(t^* - t_{I_{s-1}}) - e_D(0,t^*)$$

and

$$\begin{aligned} e_D(0,t^*) &= r_1(t^* - \frac{F_D(t^*)}{\lambda_D}) + (r_2 - r_1)(t^* - t_{I_1} - \frac{F_D(t^* - t_{I_1})}{\lambda_D}) + ... \\ &\quad + (r_s - r_{s-1})(t^* - t_{I_{s-1}} - \frac{F_D(t^* - t_{I_{s-1}})}{\lambda_D}) \end{aligned}$$

as well as

$$e_D(t^*,t_D) = N(0,t^*) F_D(t_D - t^*)$$

For $t^* < t_{S_{s-1}}$ we define m as the number of stages in between t^* and s. Hence

$$N_D(0,t^*) = r_1 t^* + (r_2 - r_1)(t^* - t_{I_1}) + ... + (r_{s-m+1} - r_{s-m})(t^* - t_{I_{n-m}}) - e_D(0,t^*)$$

and

$$\begin{aligned} e_D(0,t^*) &= r_1(t^* - \frac{F_D(t^*)}{\lambda_D}) + (r_2 - r_1)(t^* - t_{I_1} - \frac{F_D(t^* - t_1)}{\lambda_D}) + ... \\ &\quad + (r_{s-m+1} - r_{s-m})(t^* - t_{I_{s-m}} - \frac{F_D(t^* - t_{I_{s-m}})}{\lambda_D}) \end{aligned}$$

as well as

$$e_D(t^*,t_D) = N_D(0,t^*) F_D(t_D - t^*)$$

The number of events for the intermediate endpoint can be derived in a similar manner at all $s - 1$ stages.

6.2.6 Probability of research arms continuing recruitment in Stage i

When planning a multi-stage trial it is important at the outset to consider the potential number of research arms in each of the stages. This allows the implementation of safeguards for the cost and length of the trial.

When dealing with two stages only, we can calculate the probability of a research arm progressing into the second stage using the Binomial distribution. In this case we are calculating the probability of k or more research arms out of the total number of arms in Stage 1 progressing into Stage 2 when the probability of 'success' for a single research research arm is given by α_I under H_0 and $1 - \beta_I$ under H_1. Thus under H_0

$$P(k \geq x) \sim Bin(k, \alpha_I)$$

and under H_1

$$P(k \geq x) \sim Bin(k, 1 - \beta_I)$$

However, these probabilities do not take the correlation between the hazard ratios for the experimental arms compared with control into account. This exists since the same control arm is used in each comparison. Furthermore, there is the correlation between the log hazard ratios after each stage which should be taken into consideration. If an experimental arm has passed the hurdle after Stage 1, it is more likely to have a significant result in comparison with the control arm after Stage 2. Hence we compared probabilities calculated using the Binomial distribution with simulation results based on 100,000 replications. The significance level α was taken as 0.05 and 0.025 in Stages 1 and 2 respectively. In addition, power is taken to be 95% in the first stage and 90% in the last. For the simulation set-up the hazard ratio under H_1 for all experimental arms compared to control was set at 0.752. This hazard ratio is based around the ICON5 trial described in Section 7.3.1. These simulation set-ups are described in more detail in Chapter 7. Results from these studies are given in Tables 6.1 and 6.2. In general, the binomial approximation performs pretty well. However, we can observe that the results from the simulation studies give a flatter distribution of the probabilities over the number of arms in Stage 2 of the trial.

An extension of this idea to more than two stages needs to take into account that in order for a research arm to progress into Stage 3 it needs to have progressed from Stage 1 to Stage 2 already. Hence the number of 'successes' in Stage 3 is dependent on the

Approx. prob. of k experimental arms reaching Stage 2				
k (# arms)	0	1	2	3
Under H_0 calculated	0.857	0.135	0.007	0.000
Under H_0 simulated	0.778	0.156	0.051	0.015
Under H_1 calculated	0.000	0.007	0.135	0.858
Under H_1 simulated	0.005	0.026	0.105	0.864

Table 6.1: Probabilities for number of experimental arms reaching Stage 2 of a two-stage trial

		Experimental arm 2			
		Under H_0		Under H_1	
		Experimental arm 3		Experimental arm 3	
		Under H_0	Under H_1	Under H_0	Under H_1
Exp. arm 1	Under H_0	0.778	0.056	0.057	0.013
	Under H_1	0.056	0.013	0.014	0.005

Table 6.2: Probabilities for zero experimental arms reaching Stage 2 of a two-stage trial from simulation results run on 100,000 replications while varying the simulation of the experimental arms under H0 and H1

number of 'successes' in Stage 2. The same argument follows for Stages 4, 5, etc.. This is illustrated in the decision tree in Figure 6-4. Hence if k_i denotes the number of treatment arms in a given stage and x is the number of 'successes'

$$P(k_s = x) = P(k_s = x | k_{s-1} \geq x, ..., k_2 \geq x)$$

Each of these probabilities may be calculated using the binomial distribution. As an example consider a trial run in three stages with three experimental arms and one control. The significance level α was taken as 0.25, 0.1 and 0.025 in Stages 1, 2 and 3 respectively. In addition, power is taken to be 95% in the first two stages and 90% in the last. Using Figure 6-4 we can calculate the probability of having two arms in Stage 3 under H_1 using the Binomial distribution as follows

$$\begin{aligned} P(k_3 &= 2) = P(k_3 = 2 | k_2 = 3) * P(k_2 = 3) + P(k_3 = 2 | k_2 = 2) * P(k_2 = 2) \\ &= 0.859 * 0.134 + 0.134 * 0.9025 \\ &= 0.237 \end{aligned}$$

The complete results are displayed in Table 6.3. The table also gives probabilities calculated from simulation studies using 100,000 replications. For the simulation set-up the hazard ratio under H_1 was set at 0.752 for all arms in all three stages. The results indicate that while the calculations for the probabilities using the binomial distribution and

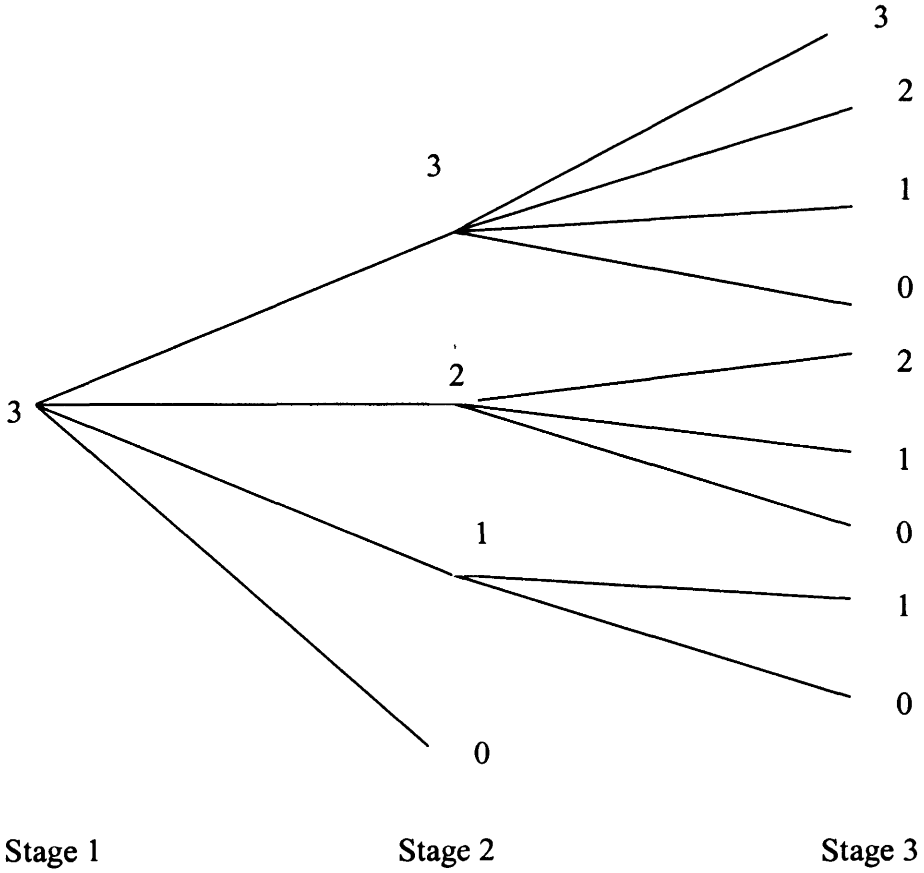


Figure 6-4: Decision tree for a multi-stage trial with three experimental arms and three stages

decision tree approach are not exact, they are a good indication of what might happen in a trial setting. However, there are some important differences for small probabilities. For example, under H_0 the calculated probability of 3 arms reaching Stage 3 is zero but in the simulation setting we still observed three arms in Stage 3 in 1% of trials.

6.2.7 Estimation of the correlation matrix

In order to estimate the correlation matrix R needed for the calculation of the overall type I error and power defined in Section 6.2.1 we bootstrapped patient data from the previously conducted trial ICON3. Estimates of the log hazard ratio for trials with more than two stages were obtained by dividing ICON3 into several periods at which the hazard ratio for the intermediate outcome was calculated. The elements of R were then estimated using the bootstrap results and are based on 1,000 replications. Table 6.4 illustrates the

Approx. prob. of k experimental arms reaching Stage 2				
k (# arms)	0	1	2	3
Under H_0 calculated	0.422	0.422	0.141	0.016
Under H_0 simulated	0.528	0.251	0.145	0.075
Under H_1 calculated	0.000	0.007	0.134	0.859
Under H_1 simulated	0.014	0.052	0.166	0.768
Approx. prob. of k experimental arms reaching Stage 3				
k (# arms)	0	1	2	3
Under H_0 calculated	0.928	0.042	0.000	0.000
Under H_0 simulated	0.803	0.141	0.044	0.012
Under H_1 calculated	0.001	0.025	0.237	0.737
Under H_1 simulated	0.027	0.085	0.214	0.673

Table 6.3: Probabilites for number of experimental arms reaching Stages 2 and 3 of a three-stage trial

	$\ln\Delta_{I_1}$	$\ln\Delta_{I_2}$	$\ln\Delta_{I_3}$	$\ln\Delta_{I_4}$	$\ln\Delta_{I_5}$	$\ln\Delta_{I_6}$	$\ln\Delta_D$
$\ln\Delta_{I_1}$	1						
$\ln\Delta_{I_2}$	0.6772	1					
$\ln\Delta_{I_3}$	0.5692	0.8182	1				
$\ln\Delta_{I_4}$	0.5155	0.7292	0.9054	1			
$\ln\Delta_{I_5}$	0.4671	0.6781	0.8369	0.9230	1		
$\ln\Delta_{I_6}$	0.2799	0.4407	0.5504	0.6106	0.6785	1	
$\ln\Delta_D$	0.2024	0.3099	0.3902	0.4386	0.4786	0.6625	1

Table 6.4: Bootstrap results for correlation matrix P based on ICON3 results $\ln\Delta_{I_i}$ gives the log hazard ratio for the intermediate outcome after 50, 100, 150, 200, 250 and 500 I events and 830 D events in the control group respectively, $\ln\Delta_D$ gives the log hazard ratio for the primary outcome based on the full dataset available from the study

results. Figure 6-5 illustrates these results graphically. Hence we can see that the strength of correlation increases the closer the log hazard ratios lie together in terms of numbers of intermediate and primary events in the control arm. Additionally, we can see that for stages with I outcomes it is not sufficient to assume a correlation of one, i.e. perfect correlation between the stages. Another important observation is that the test statistics for early stages based on I events in the control arm have a low correlation with the test statistic at the end of the trial. Therefore, in very early stages in such a trial treatments should not be rejected unlesss they are shown to be worse than control.

In addition, Figure 6-5 demonstrates that an assumption of bivariate normality between the hazard ratios is reasonable. These plots show the ellipsoidal swarm of points which is characteristic of the bivariate Normal distribution (Rice, p. 82 [100]). This is very apparent in the plot for the relationship between a log hazard ratio after 200 and 250 I events.

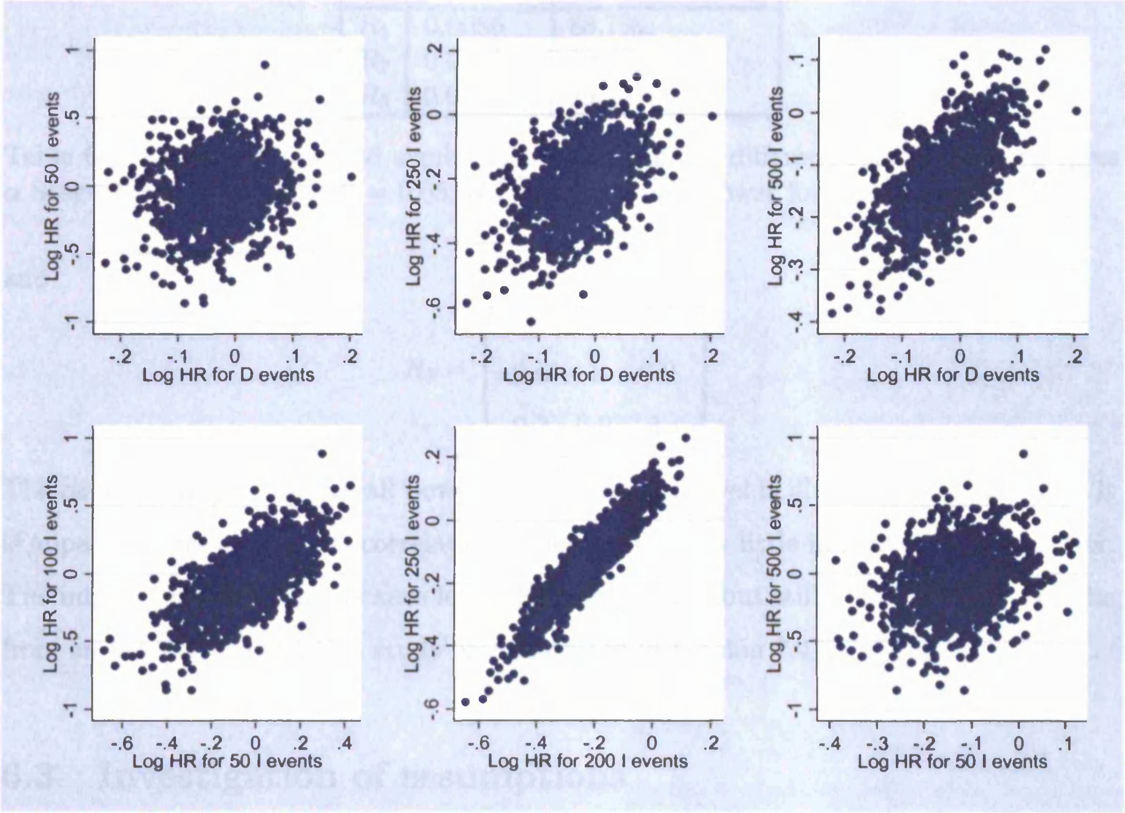


Figure 6-5: Joint distributions of the log HR for D and I events - illustrating strength of correlation at varying time points

This analysis also illustrates the difficulty in choosing an adequate specification of the correlation matrix. Hence more trials would need to be analysed to obtain a clearer picture in different disease areas. However, the choice of this matrix is important in order to calculate overall power and significance level of a multi-stage trial.

In order to see the impact on overall power and significance level of specifying a certain correlation structure we varied R in a three-stage trial three times to get

$$R_1 = \begin{pmatrix} 1 & 0.6 & 0.6 \\ 0.6 & 1 & 0.6 \\ 0.6 & 0.6 & 1 \end{pmatrix}$$

$$R_2 = \begin{pmatrix} 1 & 0.6 & 0.5 \\ 0.6 & 1 & 0.7 \\ 0.5 & 0.7 & 1 \end{pmatrix}$$

	α overall	**Power overall**
R_1	0.0086	88.7%
R_2	0.0102	88.8%
R_3	0.0123	89.3%

Table 6.5: Overall power and significance level α under different correlation structures α Stage 1 = 0.25, α Stage 2 = 0.05, α Stage 3= 0.025, power for all stages = 0.95%

and

$$R_3 = \begin{pmatrix} 1 & 0.6 & 0.3 \\ 0.6 & 1 & 0.9 \\ 0.3 & 0.9 & 1 \end{pmatrix}$$

The resulting impact on overall power and significance level is illustrated in Table 6.5. It is apparent that varying the correlation structure has very little impact on overall power. The impact on overall significance level is more apparent but still not very large. Results from an extensive simulation study are illustrated in Section 7.4.3.

6.3 Investigation of assumptions

6.3.1 Variance of the log hazard ratio

The calculations for sample size in a multi-stage trial using the Stata program whose algorithm is given in Section 6.2.1 centre around an approximation for the variance of the log hazard ratio given by

$$var(\Delta) = \frac{1}{e_1} + \frac{1}{e_2} \tag{6.11}$$

where e_1 and e_2 are the number of events in the first and second treatment group respectively [92]. Hence we decided to investigate whether this relationship breaks down at any point.

Simulations were performed in Stata 8. Design specifications for all sets of simulations were two years of accrual, two years of follow-up, uniform accrual, exponential survival and one year median survival in the control group. The probability p of being allocated to the control treatment group was varied between $\frac{1}{2}$, $\frac{2}{3}$, $\frac{4}{5}$ and $\frac{10}{11}$ and the hazard ratio was varied between 0.5, 0.7 and 0.9 in favour of the experimental treatment group. The variance estimate given in Expression 6.11 was calculated as well as an estimate from the Cox model available from Stata. In Figure 6-6 the mean difference for the variance estimates is based on 10,000 simulated trials. This difference is also illustrated in percentage terms in Figures 6-7 and 6-8. As sample sizes come close to zero simulations become less robust due to small sample issues and variance values shoot off the scale. Hence, some observations have been omitted for small values of N.

From the tables it is apparent that for high N, e.g. 1000 patients (and thus a high number of events), there is little difference between either calculation of variance. However, lower sample sizes, which result in 40 events or less, do display a greater degree of disparity between Equation 6.11 and the variance given by the Cox model. This is important since the multi-stage trials may in some cases have stages where less than 40 events are accrued per stage. Future work may hence include an improvement of this variance approximation.

6.3.2 Exponential survival

The derivation of the sample size calculations given above as well as their implementation in Stata 8 rely on exponential survival distributions. This assumption is common to many sample size formulae for time-to-event outcomes, such as Schoenfeld [114] and Freedman [41], as it eases calculations and is applicable in many trial settings. In some trial situations the assumption of exponential survival patterns, however, may not be appropriate. Ignoring this may then lead to underpowered trials as events come in later than expected. An example of a trial in breast cancer is given in Figure 6-9. In this case the actual survival distribution follows a flatter pattern than the single exponential distribution during the first year. After that, the rate of death increases, causing both curves to cross during year 4. One possible solution to this problem would be the implementation of a piecewise exponential distribution. This follows the Kaplan-Meier curve very closely in this example. The implementation of this methodology in the case of parallel group trials was described in Chapter 3. Further work is required to incorporate it into the multi-stage framework.

Another option would be to allow the user to read in the actual survival distribution for the control group taken from previous trials. By transforming this into the cumulative hazard function, working out the required sample size and other quantities on that time scale and then transforming back to the original scale, the above calculations would then still be valid, regardless of the form of the actual distribution. After transforming to the cumulative hazard function a model needs to be found which fits this function closely. Two methods were explored for the ICON3 trial. In the first instance a fractional polynomial

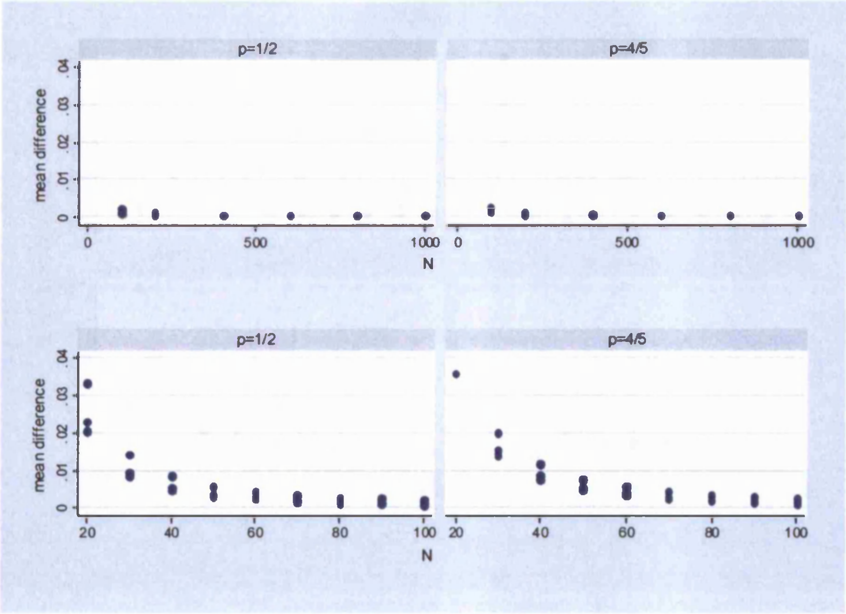


Figure 6-6: Overview of difference in variance under approximation and from Cox model - for hazard ratios 0.5, 0.7 and 0.9 mean difference - mean difference between variance calculated under approximation and from Cox model, p - probability of being allocated to control treatment group, N - number of patients, based on 10,000 replications

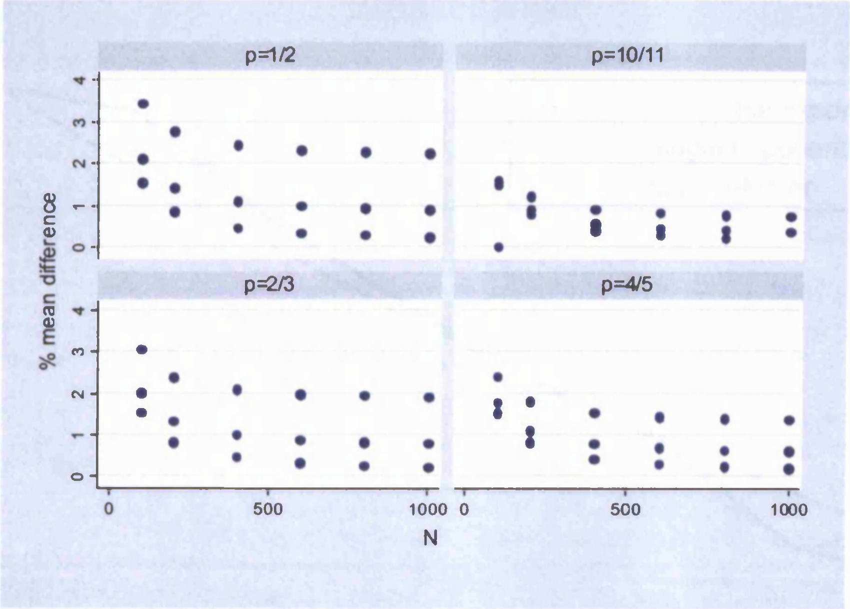


Figure 6-7: Detail of % difference in variance under approximation and from Cox model for N between 100 and 1000 - for hazard ratios 0.5, 0.7 and 0.9 % mean difference - mean difference in percent between variance calculated under approximation and from Cox model, p - probability of being allocated to control treatment group, N - number of patients, based on 10,000 replications

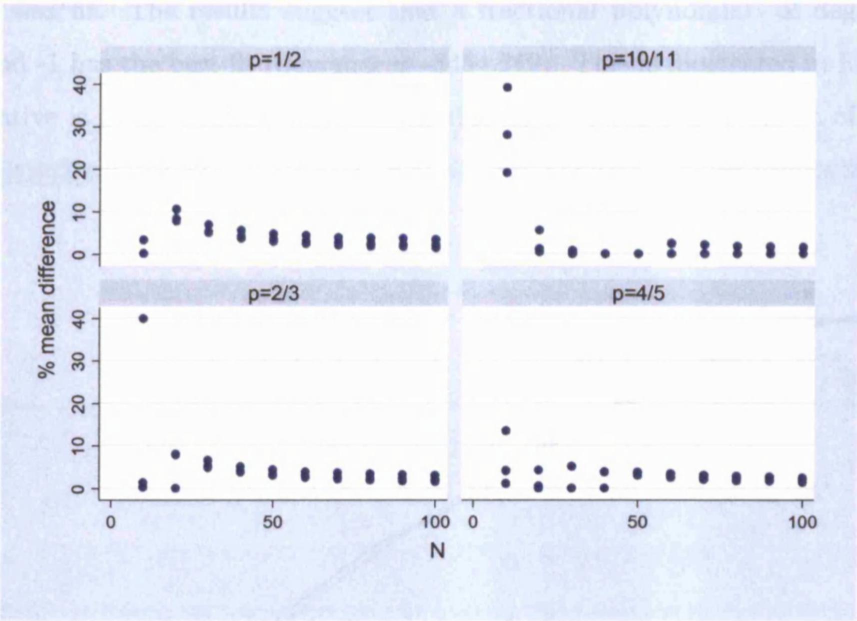


Figure 6-8: Detail of % difference in variance under approximation and from Cox model for N between 10 and 100 - for hazard ratios 0.5, 0.7 and 0.9 % mean difference - mean difference in percent between variance calculated under approximation and from Cox model, p - probability of being allocated to control treatment group, N - number of patients, based on 10,000 replications

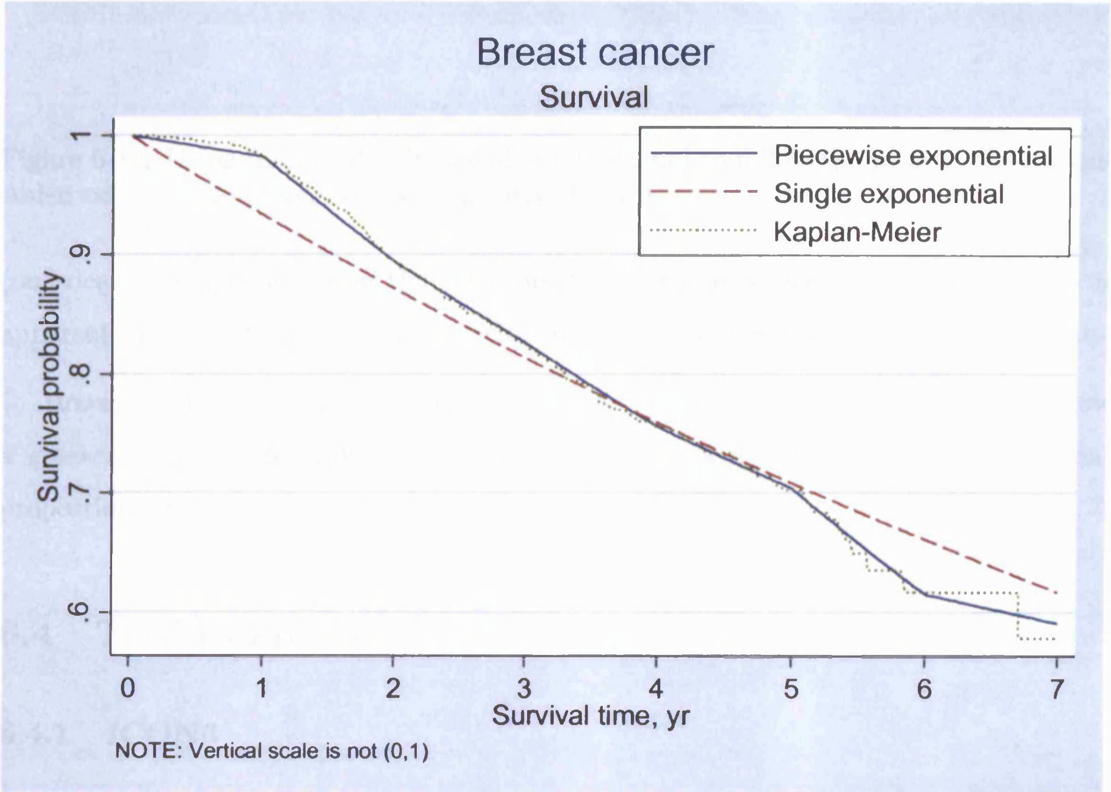


Figure 6-9: Trial example for non-exponential survival

regression was fit. The results suggest that a fractional polynomial of degree 2 with power 0 and -1 has the best fit (deviance = -3454.249). This is illustrated in Figure 6-10. An alternative is to fit a spline function. In this case we compared the fit of a Weibull

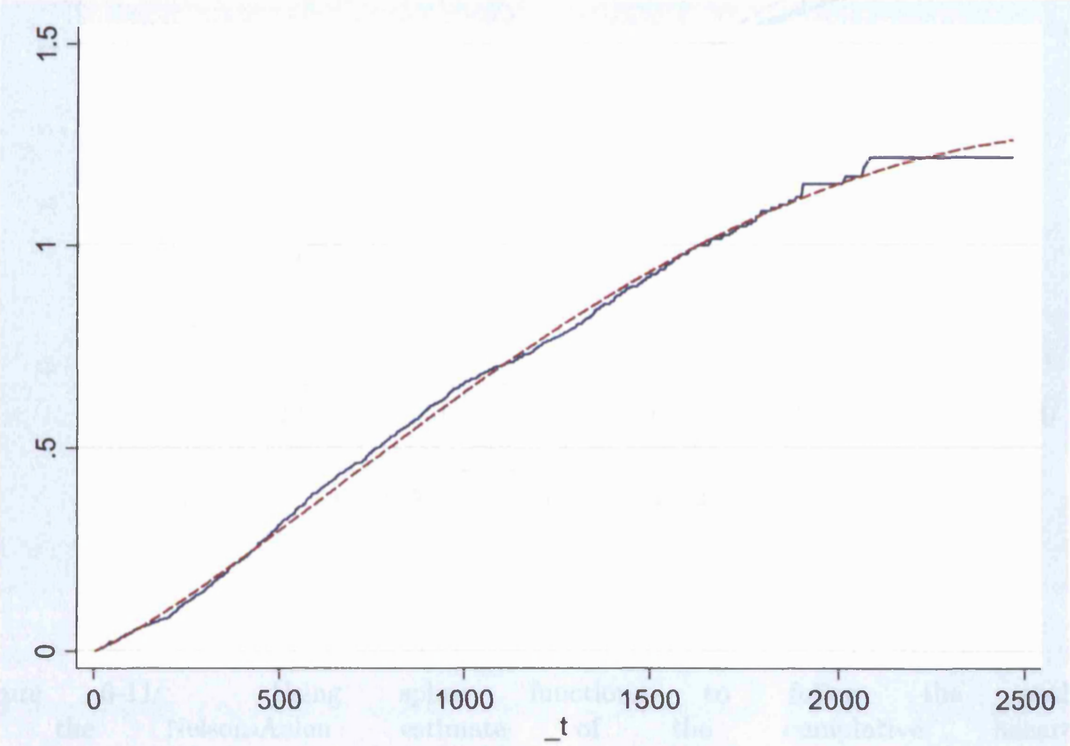


Figure 6-10: Using fractional polynomials (dotted line) to follow the path of the Nelson-Aalen estimate (solid line) of the cumulative hazard

(one degree of freedom) with that of a function with three degrees of freedom. As is apparent in Figure 6-11, the function with three degrees of freedom is more appropriate.

However, the piecewise exponential method may be preferable since it will allow a greater degree of flexibility, for example, it provides for the incorporation of non-proportional hazards.

6.4 Trial examples

6.4.1 ICON6

ICON6 is a proposed multi-stage, multi-arm clinical trial in ovarian cancer. The main objectives of this trial are to compare the efficacy of each experimental arm consisting

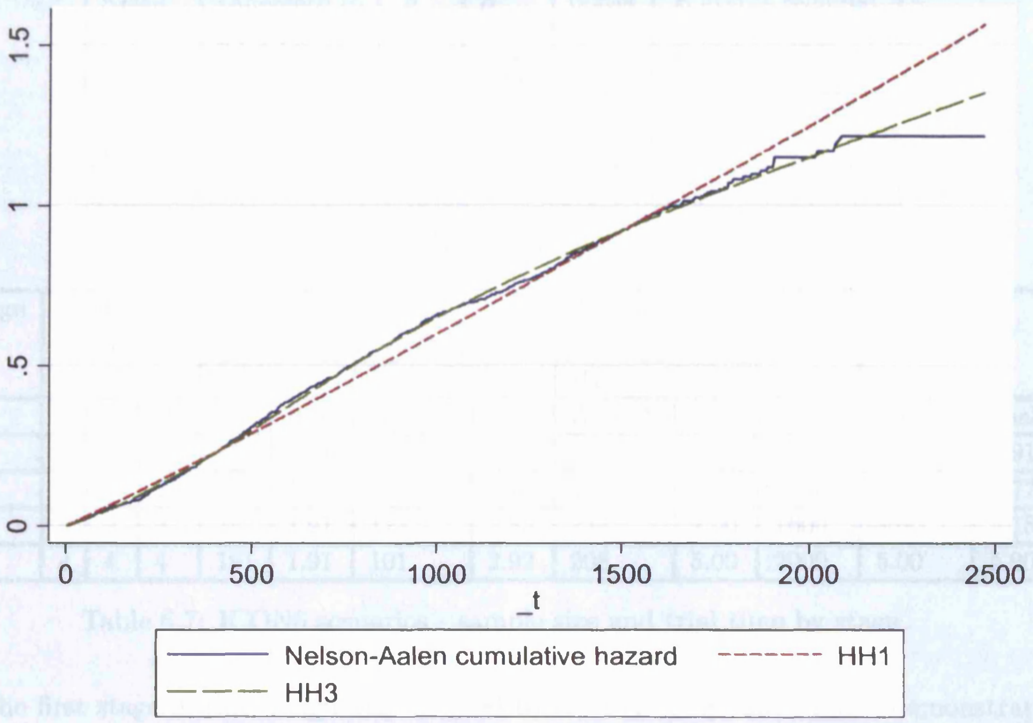


Figure 6-11: Using spline functions to follow the path of the Nelson-Aalen estimate of the cumulative hazard HH1 - spline with one degree of freedom (Weibull), HH3 - spline with three degrees of freedom

of chemotherapy plus a biological agent with the reference arm of chemotherapy alone in patients with relapsed ovarian cancer. Efficacy is to be compared through analysis of overall survival at the final stage and progression free survival at the intermediate stages. Possible design characteristics are given in Table 6.6 for a trial being conducted over three stages. Overall power and significance level were calculated with a correlation structure R

$$R = \begin{pmatrix} 1 & 0.6 & 0.5 \\ 0.6 & 1 & 0.7 \\ 0.5 & 0.7 & 1 \end{pmatrix}$$

Corresponding sample size and time requirements for several possible scenarios are illustrated by Table 6.7.

Stage	Difference to be detected		δ	α	power	# events required in control arm
I	HR	0.70	0.91	0.25	95%	94
II	HR	0.70	0.84	0.05	95%	181
III	HR	0.75	N/A	0.025	95%	330
Overall	Pairwise			0.010	89%	

Table 6.6: ICON6 design characteristics

Design	# arms in stage			# patients in reference arm and accrual period (years) in stage						Total N	Total Time	
				I		II		III				
	I	II	III	N	time	Add. N	time	Add. N	time		accrual	analysis
1	4	0	0	191	1.91	-	-	-	-	764	1.91	1.91
2	4	2	2	191	1.91	158	2.70	260	4.00	1600	4.00	4.77
3	4	3	2	191	1.91	123	2.83	234	4.00	1600	4.00	5.18
4	4	4	4	191	1.91	101	2.92	208	5.00	2000	5.00	5.90

Table 6.7: ICON6 scenarios - sample size and trial time by stage

The first stage in this design may be used to identify compounds which demonstrate sufficient activity and have acceptable toxicity. This is similar to a Phase II trial. For all scenarios given in Table 6.7 it was decided that the total accrual time should not exceed four years, except if all arms go through to Stage 3, in which case accrual is to be stopped after five years. This provides a safeguard on the number of patients to be accrued in total. In addition, all calculations assume that 400 patients may be accrued to the trial per calendar year.

6.4.2 STAMPEDE

STAMPEDE is a multi-stage, multi-arm trial in men with prostate cancer conducted at the MRC Clinical Trials Unit. This trial aims to assess three alternative classes of treatments in men starting androgen suppression. Five experimental arms are compared with a control of androgen suppression alone in five stages. In this case the first stage is used as a randomised pilot phase carried out to confirm feasibility and safety of treatments when used in combination with androgen suppression. Stages 2 to 4 are a randomised comparison of compounds shown to be safe using the intermediate outcome measure of failure free survival. The final analysis is then carried out in Stage 5 as a comparison of all those arms carried over from Stage 4 with the control based on overall survival as the primary outcome measure. Hence in terms of the multi-arm, multi-stage design and its calculations we are only dealing with four stages.

Stage	Difference to be detected		δ	α	power	# events required in reference arm	Expected total # patients
I	Pilot	n/a	n/a	n/a	n/a	n/a	210
II	HR	0.75	1.00	0.5	95%	115	1200
III	HR	0.75	0.92	0.25	95%	225	1800
IV	HR	0.75	0.89	0.1	95%	350	2400
V	HR	0.75	n/a	0.025	90%	440	3200
Overall	Pairwise			0.017	84%		

Table 6.8: STAMPEDE design characteristics

The operating characteristics and resulting number of events required at the end of each stage are illustrated in Table 6.8. A similar correlation structure R to that for ICON6 above was used to calculated overall power and significance level, i.e.

$$R = \begin{pmatrix} 1 & 0.6 & 0.5 & 0.4 \\ 0.6 & 1 & 0.7 & 0.7 \\ 0.5 & 0.7 & 1 & 0.8 \\ 0.4 & 0.7 & 0.8 & 1 \end{pmatrix}$$

This is based on the discussion of the correlation matrix in Section 6.2.7. As we can see in Table 6.8 high levels of the significance level α were chosen for Stages 2 to 4. The aim here is to avoid rejecting a potentially promising treatment arm too early on in the trial while at the same rejecting any treatments which are worse than the critical value δ. Due to the parameter values chosen a treatment will therefore pass from Stage 2 to Stage 3 if it shows any beneficial effect in comparison with the control arm. A higher significance level early on in the trial also means that we will not have to wait too long for the first comparisons while maintaining a reasonable power.

6.5 Discussion

The methodology presented in this chapter aims to address the pressing need for new additions to the 'product development toolkit' for clinical trials to achieve reliable results more quickly. The approach has two distinguishing characteristics: we compare many new therapies at once against a control treatment and we reject ineffective therapies on the basis of an intermediate outcome measure, by a randomised comparison of each new arm against the control.

A design first introduced by Royston et al. [103] has been extended to more than two stages. The underlying assumptions were examined and further improvements to the methodology were suggested particular in the area of non-exponential survival.

The calculation of overall Type I and II errors in this methodology depends upon the ability to adequately specify the correlation structure between the intermediate and primary outcomes. While some progress has been made in this area using studies on ovarian cancer with two survival-type outcomes, other cancer types and other outcome measures still require further work. Hence we propose to examine the correlation between the intermediate outcome measure and the final outcome measure, using already completed studies, to assess the likely ranges to help design multi-stage trials.

One area of criticism raised at presentations of this methodology is the potential introduction of bias. This may arise because treatments are selected at one or more interim stages and the trial is sequentially monitored. Both of these could lead to an overestimation of the treatment effect at the end of the trial. While the original publication states that such bias is avoided by reporting treatment effects for those treatments which were dropped early at the end of the trial also, this issue warrants further examination. In an academic setting such as ours patients on treatments which are not taken forward into the next stage would still be followed up and analysed at the end of the trial. However, if this design were taken to a pharmaceutical setting where economic considerations are more pressing, such treatment arms could potentially be disregarded in the final analysis which may then lead to bias in the estimates for the dropped treatment arms.

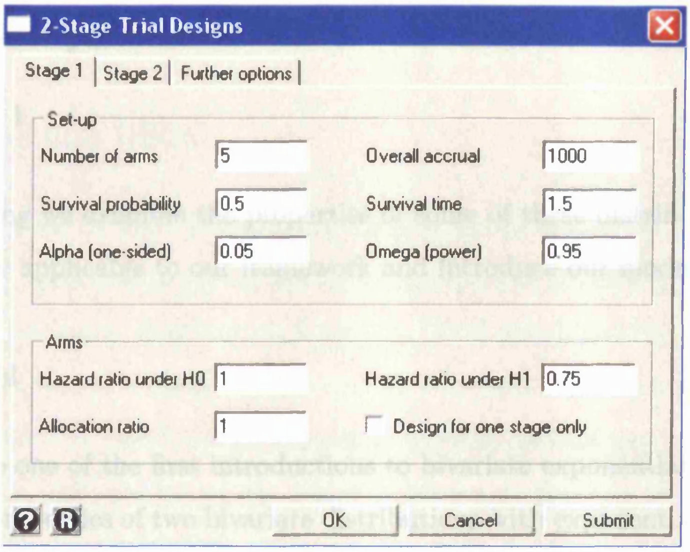
These types of trials are complex to set up since negotiations may need to be held with many stakeholders, perhaps for example many companies and several national groups in order to allow a number of different experimental treatments arms to be tested at once. Furthermore, patients could be deterred by the more complex design, though that has not been the experience to date. In addition a design such as this will require more training for participating physicians and research staff.

Chapter 7

Robustness of multi-stage trials

7.1 Introduction

This chapter provides an assessment of the validity of the sample size calculations for multi-stage trials as illustrated in Chapter 6 using simulation studies. In particular, we wish to investigate the performance of the implementation of the methodology in Stata.



Set-up

Number of arms 5

Overall accrual 1000

Survival probability 0.5

Survival time 1.5

Alpha (one-sided) 0.05

Omega (power) 0.95

Arms

Hazard ratio under H0 1

Hazard ratio under H1 0.75

Allocation ratio 1

□ Design for one stage only

Figure 7-1: Two-stage sample size program designed for Stata 8

Parameters of key importance are the power and significance level in each stage and in the trial overall. In this context, the literature concerning bivariate exponential distributions is reviewed and a bivariate exponential model based on an extension to the bivariate standard normal distribution introduced. Furthermore, assumptions of the sample size method are investigated and 'shocks' to the model explored.

7.2 Bivariate exponential distributions

We wish to simulate patients who experience disease progression, e.g. tumour growth in cancer, and death at a certain time point. These time points need to be randomly generated and there is a correlation between time to progression, X, and time to death, Y, for each patient. Since both time to progression or death and time to death are assumed to follow an exponential distribution in the derivation of the sample size formula in Chapter 6 it is convenient to require that both time to progression and time to death stem from a bivariate exponential distribution with marginal exponential distributions. As shown by Fréchet [40] this bivariate exponential distribution will not be unique; indeed he has proven that for a given marginal distribution there exist infinitely many bivariate distributions which can be defined by these margins. The desired properties for our bivariate exponential distribution are as follows:

- proportional hazards and X and Y both follow exponential marginal distributions
- $\min(X,Y) \sim \mathit{Exponential}$
- $0 \leq \rho_{X,Y} \leq 1$

In the following we examine the properties of some of these distributions. We assess whether these are applicable to our framework and introduce our model.

7.2.1 Gumbel

Gumbel [53] gave one of the first introductions to bivariate exponential models whereby he analysed the properties of two bivariate distributions with exponential margins. Up to that point most bivariate distributions studied were based around the normal distribution with concentric ellipses forming the curves of equal probability densities and straight line regression curves which intersect at the origin. Gumbel's bivariate exponential density functions are given by

$$F(x,y) = 1 - e^{-x} - e^{-y} + e^{-x-y-\delta xy} \tag{7.1}$$

for $x \geq 0$, $y \geq 0$, where $0 \leq \delta \leq 1$, and

$$F(x,y) = (1 - e^{-x})(1 - e^{-y})[1 + \alpha e^{-x-y}] \tag{7.2}$$

where $-1 \leq \alpha \leq 1$. For the first distribution as defined by the density function in Equation 7.1 the conditional expectation of one variable, X or Y, decreases to zero with increasing values of the other. The coefficient of correlation $\rho(x, Y)$ is a function of the parameter δ

$$\rho = -\frac{e^{1/\delta}}{\delta}\,\mathrm{Ei}(-\delta^{-1}) - 1$$

whereby $\mathrm{Ei}(-\delta^{-1})$ represents the exponential integral of $-\delta^{-1}$. The correlation ρ is in this case never positive and lies in the interval $-0.4 \leq \rho \leq 0$. In the case of the second distribution, Equation 7.2, the conditional expectation of one variable, X or Y, increases or decreases with increasing values of that variable, X or Y, whereby this depends on the sign of the correlation. Here the correlation lies in the interval $-0.25 \leq \rho \leq 0.25$ and is a function of α such that

$$\rho = \frac{\alpha}{4}$$

Due to the range of the correlation coefficients both these distributions are not applicable to our simulation problem.

7.2.2 Marshall and Olkin

While the derivations of the bivariate exponential distributions by Gumbel were not motivated by one particular practical problem, Marshall & Olkin [83] decided to obtain a multivariate exponential distribution based on 'fatal shock models'. Three different methods of derivation all leading to the same distribution are provided in their paper, whereby the first two are based around the 'shock models' and the last on the requirement that residual life is independent of age which is known as the loss of memory property (LMP). These different derivations underline the wide range of possible uses of their distribution. The density function common to all three derivations is given by

$$F(x,y) = \exp[-\lambda_1 x - \lambda_2 y - \lambda_{12}\max(x,y)] \tag{7.3}$$

for $x, y > 0$ whereby X and Y follow exponential marginal distributions with parameters λ_1 and λ_2 respectively. This density is often referred to in the literature as

$BVE(\lambda_1, \lambda_2, \lambda_{12})$. However, the BVE allows for the possibility that $X = Y$ occurs with positive probability. This property arises since the distribution has both an absolutely continuous and a singular part whereby the singular part is a reflection of the fact that $X = Y$. Hence according to Theorem 3.1 of their paper if density $F(x, y)$ is $BVE(\lambda_1, \lambda_2, \lambda_{12})$ and $\lambda = \lambda_1 + \lambda_2 + \lambda_{12}$ then

$$F(x,y) = \frac{\lambda_1 + \lambda_2}{\lambda} F_\alpha(x,y) + \frac{\lambda_{12}}{\lambda} F_s(x,y) \tag{7.4}$$

with the absolutely continuous part of the density

$$F_\alpha(x,y) = \frac{\lambda}{\lambda_1 + \lambda_2} \exp\left[-\lambda_1 x - \lambda_2 y - \lambda_{12} \max(x,y)\right] - \frac{\lambda_{12}}{\lambda_1 + \lambda_2} \exp\left[-\lambda \max(x,y)\right] \tag{7.5}$$

and the singular part given by

$$F_s(x,y) = \exp\left[-\lambda \max(x,y)\right] \tag{7.6}$$

In the context of 'shock models' this situation may arise when failure is caused by a shock felt by both items or if an essential input fails which is common to both items. In our situation it is difficult to imagine a situation where such an event may arise since the detection of disease progression and death are unlikely to occur in the same instance.

Still, the density provides a correlation between X and Y which is in the range $0 \leq \rho \leq 1$ and can be calculated as

$$\rho = \frac{\lambda_{12}}{\lambda} \tag{7.7}$$

where $\lambda = \lambda_1 + \lambda_2 + \lambda_{12}$. Another useful property of the distribution is that $\min(X, Y)$ follows an exponential distribution with parameter λ and $\min(X, Y)$ is independent of $X - Y$. This is illustrated by simulation studies in Figure 7-2. The distribution also retains the loss of memory property (LMP). It was proven by Block & Basu [12] that the only absolutely continuous bivariate distribution with exponential marginals and the LMP is a bivariate distribution with independent exponential marginals. Hence we need to sacrifice the LMP if we want to obtain a distribution which is absolutely continuous and has correlation in the range $0 \leq \rho \leq 1$ as required for our simulations.

7.2.3 Downton

The bivariate exponential distribution defined by Downton [28] was motivated by the BVE, however, he required absolute continuity of the density in his derivation. The model itself is based on successive damage whereby the damage is supposed to accumulate until it reaches a level sufficient to cause failure in the component. Downton assumes that a single component receives successive shocks with times between these being independent identically distributed random variables. This leads to the joint density function of the two marginally exponential distributed component lifetimes as

$$f(x,y) = \frac{\mu_1\mu_2}{1-\rho} \exp\left(-\frac{\mu_1 x + \mu_2 y}{1-\rho}\right) I_0 \left\{ \frac{2\sqrt{(\rho\mu_1\mu_2 xy)}}{1-\rho} \right\} \tag{7.8}$$

with $\mu_1, \mu_2 > 0$ and $0 \leq \rho \leq 1$. This density is a special case of the bivariate gamma distribution as discussed by Kibble [65].

However, it can be shown that $\min(X,Y)$ is not exponential, although it is stated in the paper to be a close approximation through simulation studies. Our own studies illustrated in Figure 7-3 show that while a histogram of the actual $\min(X,Y)$ follows the exponential distribution relatively well, its backtransformation to a uniform does not. Downton's paper provides a comparison with the bivariate exponential distribution by Marshall & Olkin in terms of the effect of the correlation on both the mean and variance of the smaller and larger of the two variables. We can note that while the effect of an increasing correlation is linear on the mean under the 'fatal shock model' it follows a more gradual path in the case of the 'successive damage model' whereby the mean of the larger variable decreases with increasing correlation and the mean of the smaller variable increases with increasing correlation. The effect on the variance of the two variables is very similar under both models, however, it is interesting to note that the effect of the correlation on the larger variable is to cause an initial rise in the variance in the 'fatal shock model' while the variance gradually decreases under the 'successive damage model'.

7.2.4 Sarkar

Another potential bivariate exponential model for our simulation studies was provided by Sarkar [108] under the name of $ACBVE_2$ (Absolutely Continuous BVE). This distribution is closely based on the BVE in its properties, however, the requirement was for it to be absolutely continuous as the name suggests. Hence the LMP needs to be abandoned as explained above. Apart from that it retains the properties of the BVE that X and Y are marginally exponential and that the $\min(X, Y)$ is again exponential. If X and Y are $ACBVE_2(\lambda_1, \lambda_2, \lambda_{12})$ where $\lambda_1 > 0$, $\lambda_2 > 0$ and $\lambda_{12} \geq 0$ then the density function is given by

$$F(x,y) = \begin{cases} \exp\{-(\lambda_2+\lambda_{12})y\}\left\{1-[1-\exp(-\lambda_1 y)]^{-\nu}[1-\exp(-\lambda_1 x)]^{1+\nu}\right\} & 0 < x \leq y \\ \exp\{-(\lambda_1+\lambda_{12})x\}\left\{1-[1-\exp(-\lambda_2 x)]^{-\nu}[1-\exp(-\lambda_2 y)]^{1+\nu}\right\} & x \geq y > 0 \end{cases} \tag{7.9}$$

where $\nu = \lambda_{12}/(\lambda_1 + \lambda_2)$. If $\lambda_{12} = 0$, X and Y are independent. Furthermore, the correlation is in the range $0 \leq \rho_{BVE}(X, Y) \leq \rho_{ACBVE_2}(X, Y) < 1$ whereby $\rho_{BVE}(X, Y) = \rho_{ACBVE_2}(X, Y)$ iff X and Y are independent. Simulation results show that the maximum absolute discrepancy for a given parameter combination $\Delta(\lambda_1, \lambda_2, \lambda_{12})$ between this distribution and the BVE is 1/16. One important drawback of this distribution is, however, that it is difficult to simulate from it.

7.2.5 Normal bivariate exponential (NBVE)

Due to the difficulty to simulate from those distributions described above which do hold the properties found to be critical for our simulation studies, we derived the model described in the following. So far we have found no literature references for this approach. A number of closely related distributions are described in Patil et al. [93].

The bivariate exponential model we chose for our initial simulation studies is based on a transformation of the bivariate standard normal distribution with pdf defined as

$$f(u, v, \rho) = \frac{1}{2\pi\sqrt{(1-\rho^2)}} \exp\left\{-\frac{1}{2(1-\rho^2)}(u^2 - 2\rho uv + v^2)\right\} \tag{7.10}$$

We first simulate U and V from a bivariate standard normal distribution. By definition U and V then follow marginal standard normal distributions [100], i.e.

$$U \sim N(0,1)$$

and

$$V \sim N(0,1)$$

By first transforming these into uniform random numbers A and B and then taking the logarithm

$$X = -\ln(A) * \frac{1}{\lambda_1}$$

and

$$Y = -\ln(B) * \frac{1}{\lambda_2}$$

we obtain X and Y which are marginally exponential distributed with parameters λ_1 and λ_2 and retain the range for the correlation of the bivariate standard normal distribution. The functional form may be found by using a Jacobian transformation, i.e. by treating X and Y as a transformation of U and V. Hence this approach will be referred to as $NBVE$. Since U and V have correlation $0 \leq \rho_{U,V} \leq 1$, the transformed variables X and Y will also have correlation $0 \leq \rho_{X,Y} \leq 1$. However, although the $\min(X, Y)$ is close to an exponential distribution, it is not exactly exponential as simulation studies illustrated in Figure 7-4 have shown.

The approximation appears to be sufficient though for an initial assessment of the robustness of the sample size calculations when we compare it to simulation results of the BVE as given in Figure 7-2.

7.3 Performance of the methodology

7.3.1 Simulation designs

Our simulations studies were conducted in Stata 8 and results are based on 5000 replications of each trial set-up. We simulated time to progression and time to death as variables X and Y stemming from the $NBVE$ as explained above. Progression free survival time was then taken as $\min(X, Y)$. Results for the significance level α are obtained from simulations run under the null hypothesis, i.e. with a hazard ratio of one. In order to make our scenarios as realistic as possible we based the parameters around those of ICON5. This is a trial in ovarian cancer recently conducted at the MRC in collaboration with centres in the USA, mainland Europe and Australia using the two-stage design with four experimental arms and one control arm. Sample size and number of event requirements for this trial are illustrated in Figure 7-5 for one experimental arm going through to Stage 2 only.

7.3.2 General comments on the calculation of power

In all of the tables we provide a calculation of pairwise power for each stage separately as well as an estimate of overall power in the trial. The pairwise power for Stage 2 is conditional on that treatment arm having passed to Stage 2.

Pairwise power in Stage 1 is calculated for each comparison of an experimental arm with the control by counting the number of times that the χ^2 statistic is greater than its reference value from the χ^2 tables. From this we then subtract the number of times that the hazard ratio is greater than one and divide the result by the number of repetitions used in each dataset, i.e.

$$Power1 = \frac{\#(\chi^2 > \chi^2_{1-\alpha}) - \#(HR > 1)}{\#(repetitions)}$$

This resulting power should be the same or close to the power calculated by using Formula 6.2 in Chapter 6. We will get a very similar result if we count the number of times that the hazard ratio is smaller than the cut-off δ under the alternative hypothesis H_1, divide by the number of repetitions and subtract that from one, i.e.

$$Power1\ alternative = 1 - \frac{(\#(HR < \delta) - \#(HR > 1))}{\#(repetitions)}$$

Pairwise conditional power for Stages 2 and 3 is calculated in a similar way to power for Stage 1 but we need to subtract the number of times that the arm was stopped at the previous stage from the number of repetitions, i.e.

$$Power2|arm\ passed\ to\ Stage\ 2 = \frac{\#(\chi^2 > \chi^2_{1-\alpha}) - \#(HR > 1)}{\#(repetitions) - \#(arms\ stopped)}$$

Overall power after Stage 2 or Stage 3, which is the probability that the log hazard ratio is smaller than the cut-off in all Stages under H_1, is obtained in the same way as *Power*2 without subtracting the number of trials stopped. This result should be close to overall power calculated through Formula 6.1 in Chapter 6.

We illustrate the above description with an example which corresponds to the second line of results in Table 7.3. In this case we get for Stage 1 that

$$Power1 = \frac{4675 - 2}{5000} = 0.935$$

or

$$Power1\ alternative = 1 - \frac{305 - 2}{5000} = 0.939$$

For conditional power in Stage 2 the example gives

$$Power2 = \frac{4610 - 306}{5000 - 305} = 0.917$$

and overall power may be calculated as

$$Power\ overall = \frac{4610 - 306}{5000} = 0.861$$

The significance level may be obtained in a similar manner from the simulation datasets.

7.3.3 One stage only

This set of simulations was run to assess the performance of the program for designing a standard parallel group trial, i.e. using the first stage calculations of the program only. Our aim was to find out whether target power and significance level are attained for a variety of trial scenarios. Hence the simulation sets included variations of the hazard ratio (HR), accrual rates of patients per unit time, and target power to ascertain that there is no particular combination which performs best / worst. All calculations were designed to achieve a 5% one-sided significance level α. Since results in Tables 7.1 and 7.2 are based on 5000 replications, these have a standard error of approximately 0.4% and hence a confidence interval around the nominal power of 90% ranging from 89.2 to 90.8%. Similarly, the confidence interval around a significance level of 5% ranges from 4.2 to 5.8%. Median survival in the control treatment group was taken to be one year.

We can observe that target power is maintained for all scenarios as displayed in Table 7.1. Since the approximation of the variance of the log hazard ratio as given in Formula 6.3 in Chapter 6 is derived from the Cox proportional hazards test, we also explored whether using that test instead of the logrank test in our simulation studies would give us similar results. However, the resulting power was almost identical.

Results for the significance level are also encouraging. As Table 7.2 illustrates, the significance level is robust to variations in target power.

HR	accrual rate	p	N	target power	power
0.6	700	0.5	596	90	89.8
0.6	800	0.5	638	90	90.3
0.6	900	0.5	674	90	90.2
0.6	1000	0.5	706	90	89.2
0.7	1000	0.5	1022	90	89.6
0.8	1000	0.5	1718	90	89.8
0.9	1000	0.5	4494	90	90.2

Table 7.1: Simulation results for power for one stage only, one control and one experimental group
HR - hazard ratio in favour of experimental group in comparison with control, accrual rate - rate of patients accrued per unit time, p - probability of being allocated to control treatment group, N - sample size calculated for target power and 5% significance level, power - power achieved through simulation with sample size N

HR	accrual rate	p	N	target power	target α	α
0.6	700	0.5	424	70	5	4.6
0.6	700	0.5	496	80	5	5.1
0.6	700	0.5	596	90	5	4.8
0.6	700	0.5	686	95	5	5.3
0.6	800	0.5	638	90	5	4.8
0.6	900	0.5	674	90	5	5.5
0.6	1000	0.5	706	90	5	4.8
0.7	1000	0.5	1022	90	5	5.1
0.8	1000	0.5	1718	90	5	4.8
0.9	1000	0.5	4494	90	5	4.9

Table 7.2: Simulation results for the significance level for one stage only, one control and one experimental group
HR - hazard ratio in favour of experimental group in comparison with control, accrual rate - rate of patients accrued per unit time, p - probability of being allocated to control treatment group, N - sample size calculated for target alpha and 90% power, α - significance level achieved through simulation with sample size N

7.3.4 Two arms

Our next performance assessment was based around a two-stage trial where we have one experimental and one control treatment group in both stages of the trial. The main concern was again the robustness of power and the significance level in a variety of trial settings. One particular focus here was on the assessment of overall power and significance level achieved.

Time to progression and time to death were simulated as correlated exponentials from the *NBVE* model. In all simulated datasets median survival for the progression free survival time was taken to be one unit of time and median survival for overall survival was set at two units of time. In addition, the correlation ρ between the primary and intermediate outcome was fixed at 0.6. We refer the reader to Section 7.4.3 for the impact of variations in ρ on both power and the significance level.

For the analysis of power the accrual rate in Stages 1 and 2 was varied as well as target power after Stage 2. The significance level was taken as fixed at 5% in Stage 1 and 2.5% in Stage 2. As Table 7.3 shows power in Stage 1 is below the confidence bounds while power in Stage 2 is close to the nominal power though it overshoots in some cases. Overall power is slightly lower than estimated by the program which is given at approximately 87% if we have a power of 95% in Stage 1 and 90% in Stage 2. However, we do not expect to get exact results with these simulations since $\min(X, Y)$ is only close to an exponential distribution in the bivariate exponential distribution NBVE. Again, we ran a second set of simulations using the Cox proportional hazards test for analysis instead of the logrank test. These results are given in Table 7.4. In general, this provides no improvement.

HR	accrual	accrual	N 1	N 2	target	power 1	target	power 2	power overall	power
	rate 1	rate 2			power 1		power 2		program	overall
0.752	700	700	1290	1424	95	92.9	80	81.5	78.4	76.2
0.752	700	700	1290	1689	95	93.5	90	91.7	87.4	86.1
0.752	700	700	1290	1913	95	93.8	95	96.2	91.6	90.6
0.752	700	1000	1290	1476	95	93.7	80	80.6	78.4	75.9
0.752	700	1000	1290	1824	95	93.4	90	91.3	87.4	85.7
0.752	700	1000	1290	2105	95	93.6	95	96.3	91.6	90.5
0.752	1000	700	1494	1620	95	93.3	80	82.1	78.4	77.0
0.752	1000	700	1494	1848	95	93.4	90	92.3	87.4	86.7
0.752	1000	700	1494	2050	95	93.4	95	96.2	91.6	90.3
0.752	1000	1000	1494	1670	95	92.9	80	81.2	78.4	75.9
0.752	1000	1000	1494	1974	95	93.4	90	91.6	87.4	86.0
0.752	1000	1000	1494	2233	95	93.8	95	95.8	91.6	90.4

Table 7.3: Simulation results for power for two stages, one control and one experimental group
HR - hazard ratio in favour of experimental group in comparison with control, accrual rate - rate of patients accrued per unit time, N - sample size calculated for target power and 5% and 2.5% significance level in Stages 1 and 2 respectively, power - power achieved through simulation with sample size N, power overall program - overall power as given by sample size program

Simulation sets to assess the robustness of the significance level focused on variations in the rate of patient accrual in Stages 1 and 2 as well as variations in the significance levels of Stages 1 and 2. Power was fixed at 95% and 90% in Stage 1 and 2 respectively. Results from this analysis are displayed in Table 7.5. Overall it is apparent that the significance level is more robust than power. The significance level is nearly always within the confidence intervals of the corresponding target level. The only exception is

HR	accrual	accrual	N 1	N 2	target	power 1	target	power 2	power overall	power
	rate 1	rate 2			power 1		power 2		program	overall
0.752	700	700	1290	1424	95	93.2	80	81.3	78.4	76.3
0.752	700	700	1290	1689	95	92.9	90	91.6	87.4	85.5
0.752	700	700	1290	1913	95	93.3	95	96.4	91.6	90.4
0.752	700	1000	1290	1476	95	93.4	80	81.1	78.4	76.2
0.752	700	1000	1290	1824	95	92.7	90	91.5	87.4	85.5
0.752	700	1000	1290	2105	95	93.3	95	96.7	91.6	90.7
0.752	1000	700	1494	1620	95	93.2	80	82.1	78.4	77.1
0.752	1000	700	1494	1848	95	93.5	90	91.9	87.4	86.5
0.752	1000	700	1494	2050	95	92.9	95	96.4	91.6	90.3
0.752	1000	1000	1494	1670	95	93.6	80	80.1	78.4	75.4
0.752	1000	1000	1494	1974	95	93.2	90	91.0	87.4	85.2
0.752	1000	1000	1494	2233	95	94.0	95	95.9	91.6	90.8

Table 7.4: Simulation results for power for two stages, one control and one experimental group, analysed using Cox proportional hazards test HR - hazard ratio in favour of experimental group in comparison with control, accrual rate - rate of patients accrued per unit time, N - sample size calculated for target power and 5% and 2.5% significance level in Stages 1 and 2 respectively, power - power achieved through simulation with sample size N, power overall program - overall power as given by sample size program

when a significance level of 50% is to be attained in the first stage, however, this may be due to the difficulty in obtaining an accurate estimate of the χ^2 statistic in this case.

HR	accrual	accrual	N 1	N 2	critical	target	alpha 1	target	alpha 2
	rate 1	rate 2			HR	alpha 1		alpha 2	
0.752	700	700	872	987	0.92	25	25.3	25	24.9
0.752	700	700	872	1310	0.92	25	25.5	10	9.7
0.752	700	700	872	1509	0.92	25	25.4	5	5.0
0.752	700	700	606	987	1.00	50	49.0	25	24.6
0.752	700	700	1024	1024	0.90	15	15.4	25	25.1
0.752	700	1000	1024	1411	0.90	15	14.8	10	9.8
0.752	1000	700	1194	1447	0.90	15	15.0	10	9.8
0.752	1000	1000	1194	1539	0.90	15	14.4	10	9.9

Table 7.5: Simulation results for alpha for two stages, one control and one experimental group
HR - hazard ratio in favour of experimental group in comparison with control, accrual rate - rate of patients accrued per unit time, N - sample size calculated for target alpha and 95% and 90% power in Stages 1 and 2 respectively, alpha - significance level achieved through simulation with sample size N

7.3.5 More than two stages

Following the extension of the design as outlined in Chapter 6 we wanted to assess the robustness of those extensions for a three-stage trial with one experimental arm and one control. Our main aim was to assess the robustness of power using the logrank test. In addition, simulation runs were conducted using the Cox proportional hazards test for analysis.

In this case, time to progression, X, and time to death, Y, were simulated using the NBVE model. However, in this set of simulations both Stage 1 and Stage 2 used time to progression free survival as the outcome and Stage 3 was simulated using time to death. Thus events obtained by taking $\min(X, Y)$ were employed for analysis in Stages 1 and 2.

The results displayed in Table 7.6 were obtained by stopping each stage after the required number of events for that stage had been reached, whereby the number of I-events were counted for the first two stages and the number of D-events for the final stage. A further approximation, however, enters these results since as illustrated using correlation results from ICON3 (Chapter 6, Table 6.4) the strength of the correlation between the log hazard ratios for progression free survival and overall survival may change between Stages 1 and 2. These changes have not been taken into consideration in the calculations for the results illustrated since they are not easily quantifiable and depend on the disease area. Required power for Stages 1 and 2 was 95% and a significance level of 10%, 5% and 2.5% was desired for Stages 1, 2 and 3 respectively. The correlation between the test statistics after each of the stages was set at 0.6. Median survival was fixed at one unit of time for progression free survival and two units of time for overall survival.

Table 7.6 shows that in general power for Stage 1 in this design is slightly lower than desired, while both Stages 2 and 3 are overpowered. However, considering the approximations made in the simulation studies, we believe that these results are satisfactory. Results obtained by using the Cox proportional hazards test as shown in Table 7.7 are again not significantly better. In addition we may observe that the overall power does not decrease significantly by adding an extra stage.

7.4 'Shocks' to the design

When assessing the robustness of a certain design, we not only want to know how it performs under optimum conditions but also test which situations may cause it to falter. This allows us to safeguard against these circumstances when applying the methodology in practice. In the following three likely situations are considered.

accrual	accrual	accrual	N 1	target	power1	N 2	target	power2	N 3	target	power3	overall
rate 1	rate 2	rate 3		power1			power2			power3		power
700	700	1000	1128	95	93.1	1291	95	97.7	1476	80	81.5	74.8
700	700	1000	1128	95	93.0	1291	95	97.5	1824	90	92.1	84.2
700	700	1000	1128	95	93.6	1291	95	97.3	2105	95	96.7	88.9
700	1000	1000	1128	95	92.8	1351	95	97.4	1529	80	81.3	74.0
700	1000	1000	1128	95	94.0	1351	95	97.3	1863	90	92.2	85.2
700	1000	1000	1128	95	92.8	1351	95	97.6	2137	95	96.1	87.8
1000	700	1000	1134	95	93.4	1445	95	97.1	1626	80	81.4	74.6
1000	700	1000	1134	95	93.6	1445	95	97.3	1939	90	92.0	84.5
1000	700	1000	1134	95	94.0	1445	95	97.4	2203	95	96.4	89.0
1000	1000	1000	1134	95	93.8	1495	95	97.6	1670	80	80.5	74.3
1000	1000	1000	1134	95	93.3	1495	95	97.4	1974	90	92.0	84.5
1000	1000	1000	1134	95	93.0	1495	95	97.0	2233	95	96.3	87.9

Table 7.6: Simulation results for power for three stages, one control and one experimental group
accrual rate - rate of patients accrued per unit time, N - sample size calculated for target power and 10%, 5% and 2.5% significance level in Stages 1, 2 and 3 respectively, power - power achieved through simulation with sample size N

7.4.1 Number of arms in Stage 2

When designing a two-stage trial we need to estimate the number of arms carried over into Stage 2 in order to arrive at a sensible sample size estimate. Thus we ran an investigation into the impact on power of a mis-specification of the number of arms in Stage 2 of the trial.

Trials were simulated with up to three experimental and one control treatment arm and a similar set of parameters as before. Hence median time to progression or death was taken as one year and median time to death as three years with hazard ratios in both cases at 0.752 in favour of the experimental arm under H_1. Furthermore an accrual rate of 900 patients per year was assumed for both stages and a correlation of ρ=0.6 between the two test statistics. Sample size was calculated for trials with at most three experimental arms in Stage 2. Thus in some cases the number of simulated experimental arms exceeded the number of experimental arms assumed in the sample size calculations in Stage 2 and vice versa.

Figure 7-6 illustrates the impact on power at the end of Stage 2 and overall power. The upper and lower lines indicate the confidence interval around the nominal power of 90%. We can observe that there is a near linear relationship in terms of power whereby we only achieve power within the confidence interval if the actual number of arms in Stage 2 corresponds to the number of arms that the sample size was calculated for.

accrual	accrual	accrual	N 1	target	power1	N 2	target	power2	N 3	target	power3	overall
rate 1	rate 2	rate 3		power1			power2			power3		power
700	700	1000	1128	95	93.0	1291	95	97.5	1476	80	81.5	74.7
700	700	1000	1128	95	94.2	1291	95	97.1	1824	90	91.4	84.1
700	700	1000	1128	95	93.6	1291	95	97.0	2105	95	96.1	88.1
700	1000	1000	1128	95	93.3	1351	95	96.7	1529	80	83.5	76.2
700	1000	1000	1128	95	93.3	1351	95	97.1	1863	90	91.8	83.8
700	1000	1000	1128	95	93.2	1351	95	97.1	2137	95	96.2	88.0
1000	700	1000	1134	95	93.8	1445	95	97.2	1626	80	82.0	75.4
1000	700	1000	1134	95	93.5	1445	95	97.1	1939	90	92.0	84.2
1000	700	1000	1134	95	94.1	1445	95	97.3	2203	95	96.2	88.9
1000	1000	1000	1134	95	93.1	1495	95	97.3	1670	80	81.4	74.6
1000	1000	1000	1134	95	93.5	1495	95	97.3	1974	90	91.8	84.3
1000	1000	1000	1134	95	93.6	1495	95	97.1	2233	95	96.7	88.6

Table 7.7: Simulation results for power for three stages, one control and one experimental group, analysed under Cox proportional hazards test accrual rate - rate of patients accrued per unit time, N - sample size calculated for target power and 10%, 5% and 2.5% significance level in Stages 1, 2 and 3 respectively, power - power achieved through simulation with sample size N

In order to have a spectrum of sample size requirements we urge the user to run the sample size program using different scenarios. In addition, it is possible to obtain an estimate of the probability of the number of arms in the stages by ticking this option on the program menu.

7.4.2 The actual accrual rate

During the conduct of ICON5 it was observed that the actual accrual rate for Stage 1 was faster than that which was used for the original sample size calculations. This meant that accrual for Stage 2 was started before a Stage 1 analysis could be run as the necessary number had not been observed by the time that patients for Stage 1 had been accrued. Therefore, arms which were stopped after the first Stage (in this case all) had had too many patients accrued to them. Hence we wanted to know the impact of a lower or higher than anticipated accrual rate on either the number of events accrued in the control arm by the predicted end of Stage 1 or the time by which the required number of events would be accrued in Stage 1.

Figure 7-7 illustrates the impact of the actual accrual on both number of events and time. In this figure the solid line gives the calculated number of events needed in the control treatment arm and the predicted time needed to run Stage 1 under an anticipated accrual rate of 900 patients per year. The underlying parameters were the same as in

Section 7.4.1 with $\rho = 0.6$. The relationship between the rate of accrual and the number of events or the total length of Stage 1 appears to be near linear in both cases, whereby the relationship with time experiences more of a levelling off towards higher rates of accrual. We may say that a 10% change in the accrual rate causes a change of approximately 5% in either the number of events or time.

7.4.3 Impact of correlation on power

In all the simulation sets in the previous sections we have taken the correlation between the Stage 1 and Stage 2 test statistics to be a fixed number, usually 0.6. However, we now wanted to assess whether a) the strength of the correlation coefficient has an impact on power and b) whether mis-specifying the correlation at the trial planning stage has an impact on power in a two stage trial.

Simulation studies for which the results are illustrated in Figures 7-8, 7-9 and 7-10 were run using a hazard ratio of 0.752 in both stages. Also, as before the median survival time for progression free survival was taken to be one unit of time and median survival for overall survival was fixed at two units of time. The target power for Stage 1 was set at 95% and at 90% for Stage 2. Furthermore, the significance level to be attained in Stage 1 was set at 5% and at 2.5% in Stage 2.

There is no obvious relationship between the strength of the correlation coefficient and overall power as illustrated in Figure 7-8. However, from this figure we may say that power in Stage 2 increases as the correlation coefficient reaches 0.4 and above. From Figure 7-9 we can observe that mis-specifying the correlation coefficient appears to have no effect on power. In general in this figure, however, power appears to be higher for a correlation coefficient of 0.8. The results for the significance level as depicted in Figure 7-10 also show no particular influence of the correlation coefficient. In most cases the pattern appears random apart from the results for an actual correlation of 0.8 and the significance level at Stage 2. In this case the significance level appears to decrease with an increasing specified correlation coefficient. However, these results do not fall outside the confidence bounds.

In general the strength of the correlation coefficient is important for the design as a whole though as running a trial with a low correlation between the two test statistics may be dangerous.

7.5 Rejection sampling

In order to improve our simulation studies we investigated the use of rejection sampling. As Figure 7-4 shows, the minimum of the two exponential distributions is not exactly exponentially distributed when we sample from the NBVE. This is especially apparent in the transformation to the uniform as illustrated in the density histogram. If the distribution was exact, all density bins would have a height of one. The shape of the transformation and the degree of variation from the uniform varies for different values of median survival as illustrated in Figure 7-11.

The principle of the rejection sampling method is that a distribution function $f(x)$ is approximated by another distribution function $h(x)$ which is easier to calculate, and then a correction is made by randomly accepting values with a probability $p(x)$, and rejecting x values with a probability $1 - p(x)$ [46]. For our purposes, the algorithm is therefore as follows:

1. Draw x and y from the $NBVE$, create $z = \min(x, y)$
2. For this z evaluate whether to reject or not, i.e. whether the density is greater than one
3. If we need to reject z, resample x and y from $NBVE$ and re-evaluate new z

Using this algorithm, we reproduced Figure 7-4 to get the improved Figure 7-12. However, while there are improvements for the distribution of $\min(X, Y)$, these come at a high cost. The simulations to obtain Figure 7-12 took over four days, compared to two minutes without rejection sampling.

There were also problems with the algorithm running infinitely in some simulation cases. Hence we relaxed the rejection criterium to reject z only if the density of a particular bin was greater than 1.02. This number was arrived at by simulating 100,000 observations from a $uniform(0, 1)$ distribution and plotting the resulting histogram. Variations in density between 0.98 and 1.02 could be observed. Using this relaxed criterion, we arrive at Figure 7-13 in just over 90 minutes.

Thus, simulations to assess power on 5,000 replications would take 625 days, compared to an average of 15 minutes before. These times will increase for parameter combinations which result in further diversions from the uniform as illustrated in Figure 7-11. The method is therefore highly inefficient in terms of computer time and would be impractical for large scale simulation studies.

7.6 Discussion

Overall the simulations have shown that the sample size calculations underlying the multi-stage model perform satisfactorily. Particular attention needs to be paid to the accrual rate and a sensitivity analysis to the number of arms in Stage 2 (and the following stages), with the first potentially leading to problems with the feasibility of a Stage 1 analysis and mis-specification in the second area causing over- or underpowered studies.

Future work in this area could include a set of simulations using the BVE or Downton method for simulating time-to-event data in order to assess whether the calculations are robust under these assumptions also. The use of correlated frailty models may also be considered. A good introduction to their application is given by Wienke [142]. These were originally developed for the analysis of bivariate failure time data in which two associated random variables are used to characterise the frailty effect for each pair.

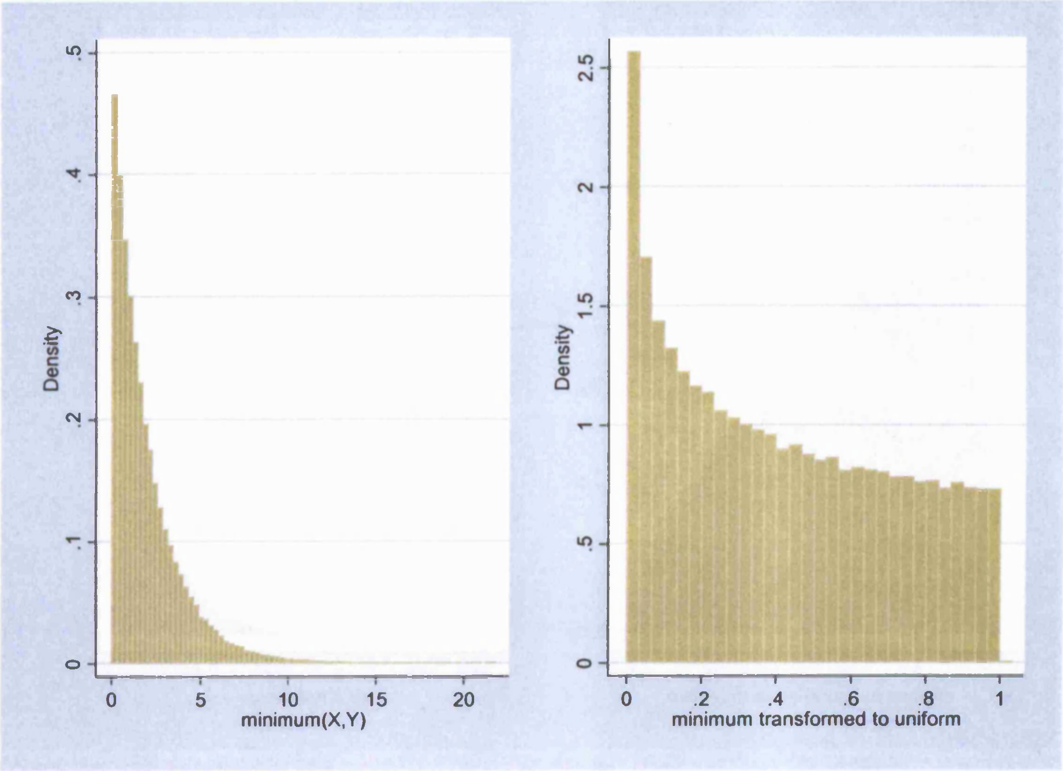


Figure 7-2: Assessment through simulation studies of whether min(X,Y) as given by the BVE follows an exponential distribution based on 100,000 replications

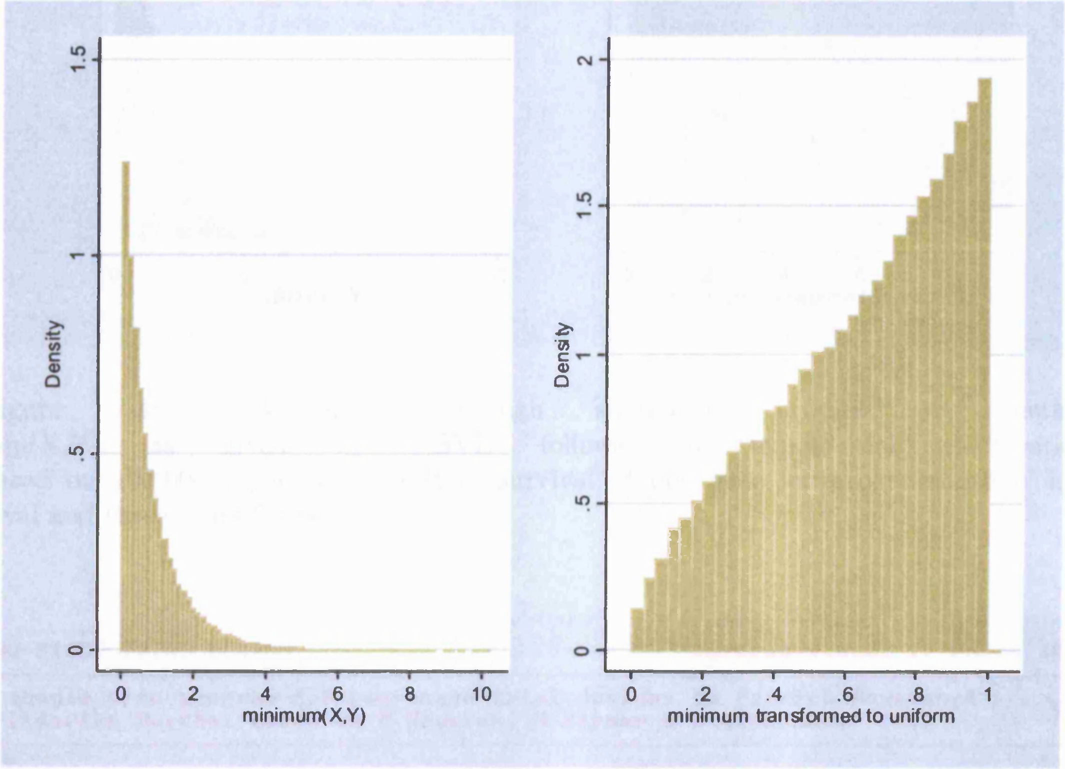


Figure 7-3: Assessment through simulation studies of whether min(X,Y) as given by Downton follows an exponential distribution based on 100,000 replications, median survival of one year for progression-free survival and three years for death

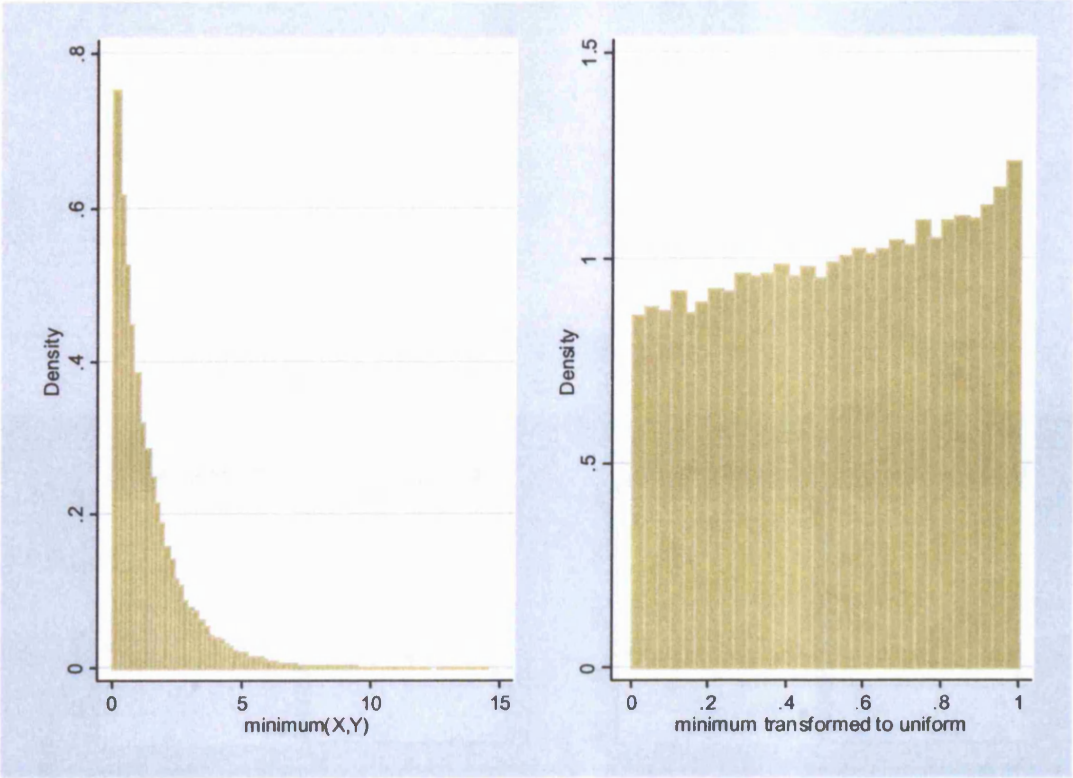
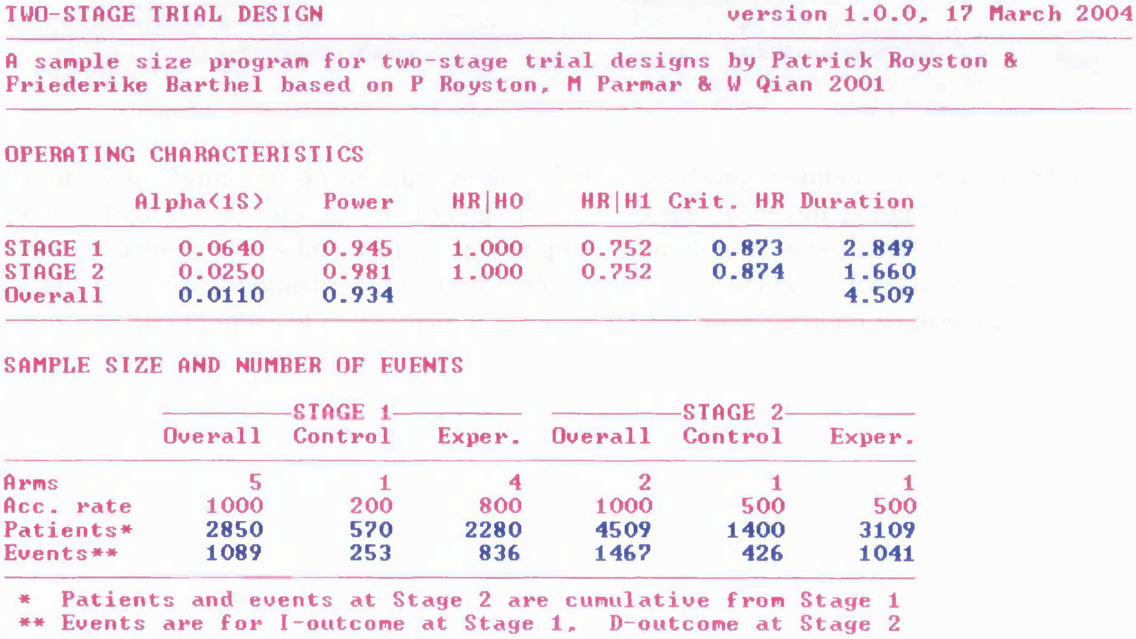


Figure 7-4: Assessment through simulation studies of whether min(X,Y) as given by NBVE follows an exponential distribution based on 100,000 replications, median survival of one year for progression-free survival and three years for death



```
TWO-STAGE TRIAL DESIGN                               version 1.0.0, 17 March 2004
---------------------------------------------------------------------------------
A sample size program for two-stage trial designs by Patrick Royston &
Friederike Barthel based on P Royston, M Parmar & W Qian 2001
---------------------------------------------------------------------------------

OPERATING CHARACTERISTICS

         Alpha(1S)    Power    HR|H0    HR|H1 Crit. HR Duration
----------------------------------------------------------------
STAGE 1     0.0640    0.945    1.000    0.752    0.873    2.849
STAGE 2     0.0250    0.981    1.000    0.752    0.874    1.660
Overall     0.0110    0.934                               4.509
----------------------------------------------------------------

SAMPLE SIZE AND NUMBER OF EVENTS

           ---------STAGE 1---------  ---------STAGE 2---------
           Overall  Control   Exper.  Overall  Control   Exper.
----------------------------------------------------------------
Arms             5        1        4        2        1        1
Acc. rate     1000      200      800     1000      500      500
Patients*     2850      570     2280     4509     1400     3109
Events**      1089      253      836     1467      426     1041
----------------------------------------------------------------
 *  Patients and events at Stage 2 are cumulative from Stage 1
 ** Events are for I-outcome at Stage 1,  D-outcome at Stage 2
```

Figure 7-5: Sample size requirements for ICON5 as output by Stata program

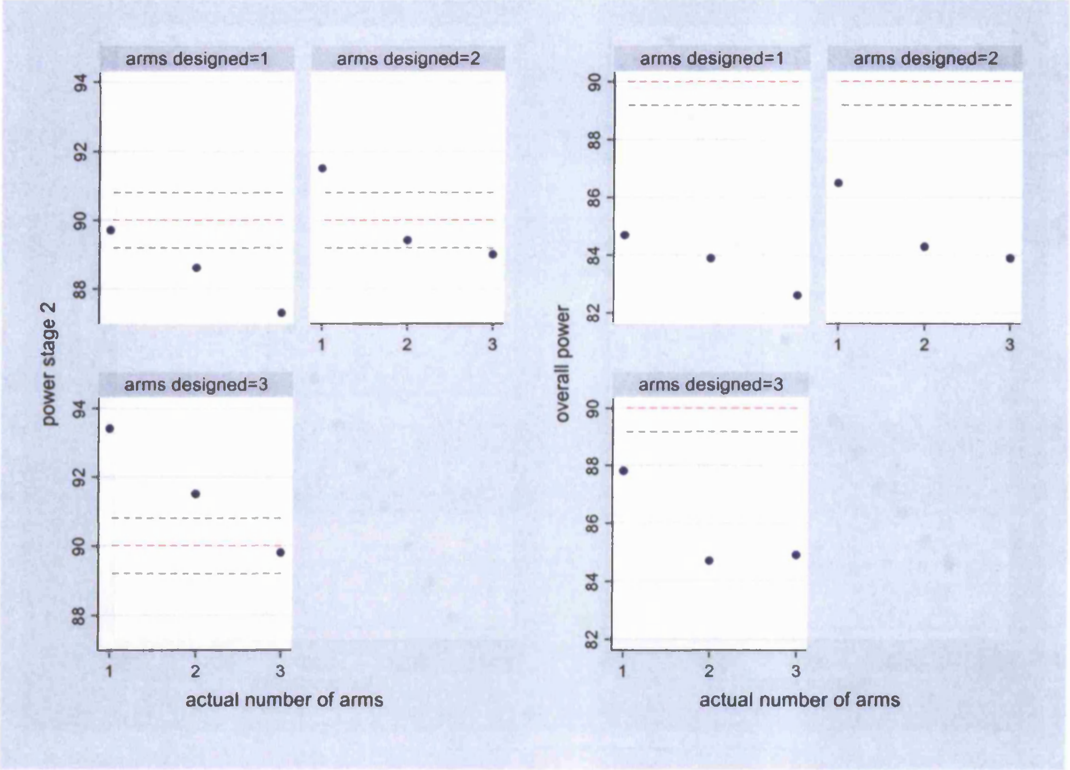


Figure 7-6: Influence on average power of mis-specifying number of arms in Stage 2 power stage 2 - average power over all arms for Stage 2, overall power - average power over all arms for the whole trial, arms designed - number of experimental arms in Stage 2 that trial was designed for (i.e. that sample size was calculated for), actual number of arms - actual number of experimental arms that have gone through to Stage 2

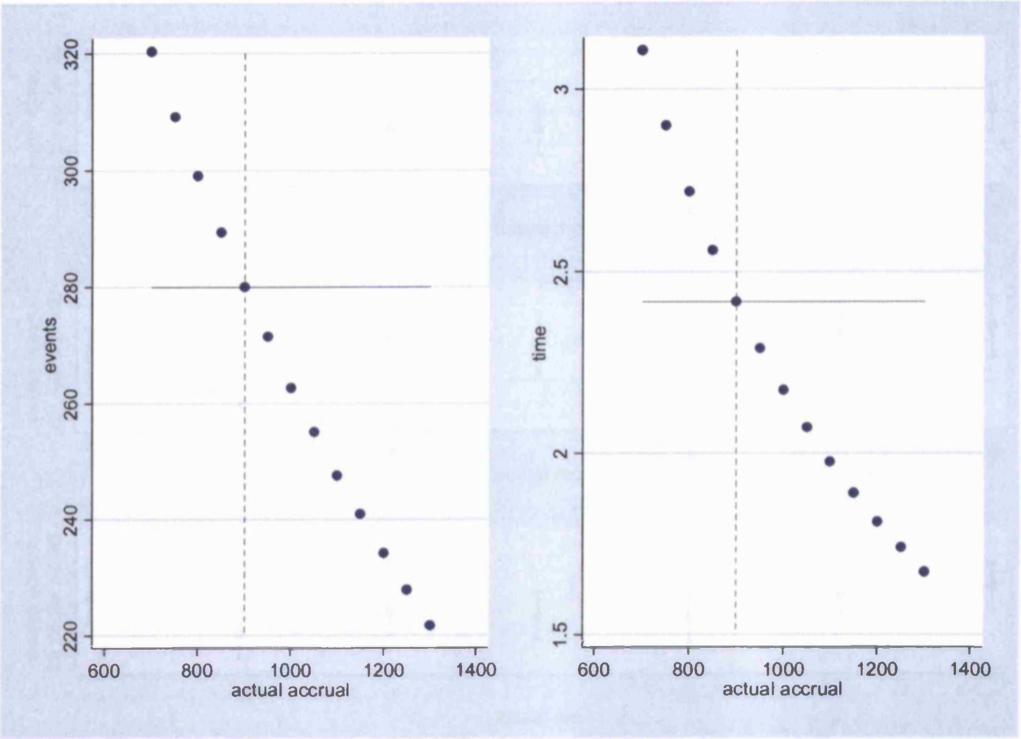


Figure 7-7: Influence of mis-specifying accrual on number of mean events accrued during Stage 1 and total duration of Stage 1 events - average number of events obtained by projected end of Stage 1 under actual rate of accrual, time - average time taken to accrue required number of patients for Stage 1 under actual rate of accrual, actual accrual - actual rate of accrual per unit of time

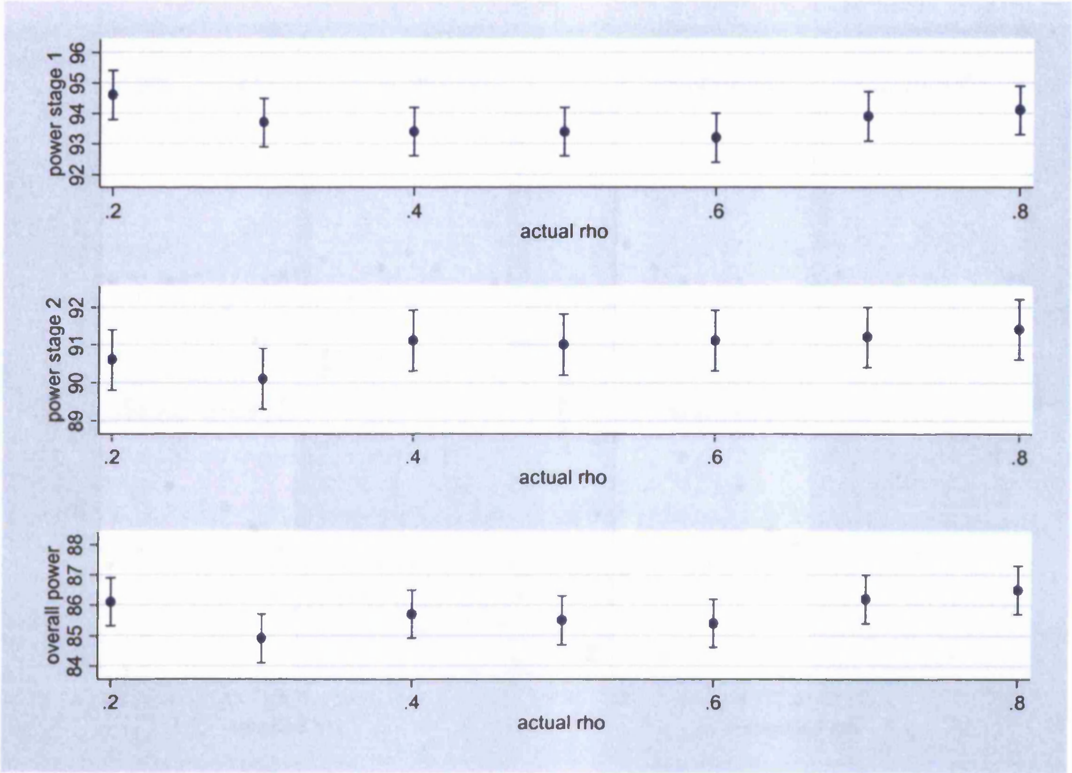


Figure 7-8: Relationship between power and correlation coefficient power stage 1 - power for Stage 1 with 95% confidence intervals, power stage 2 - power for Stage 2 with 95% confidence intervals, overall power - overall power for the trial with 95% confidence intervals, rho - correlation coefficient between test statistics after Stages 1 and 2

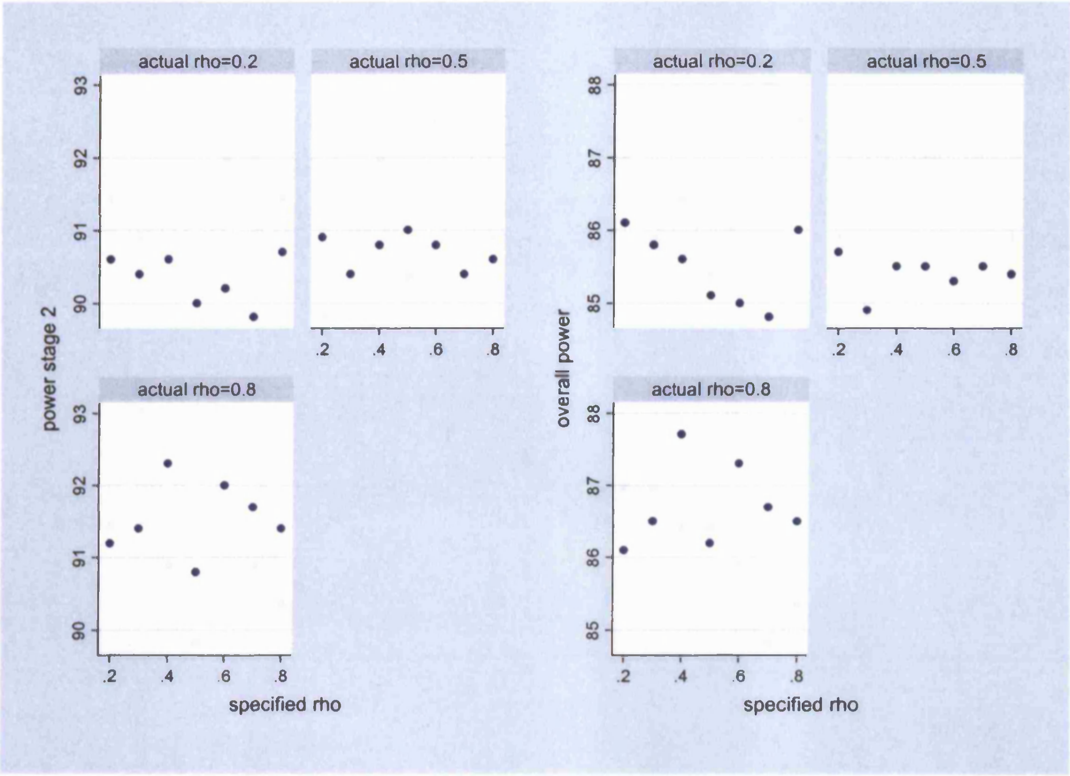


Figure 7-9: Relationship between mis-specified correlation and power specified rho - correlation used for sample size calculations, actual rho - correlation used in simulations, power stage 2 - power for Stage 2, overall power - overall power for the trial

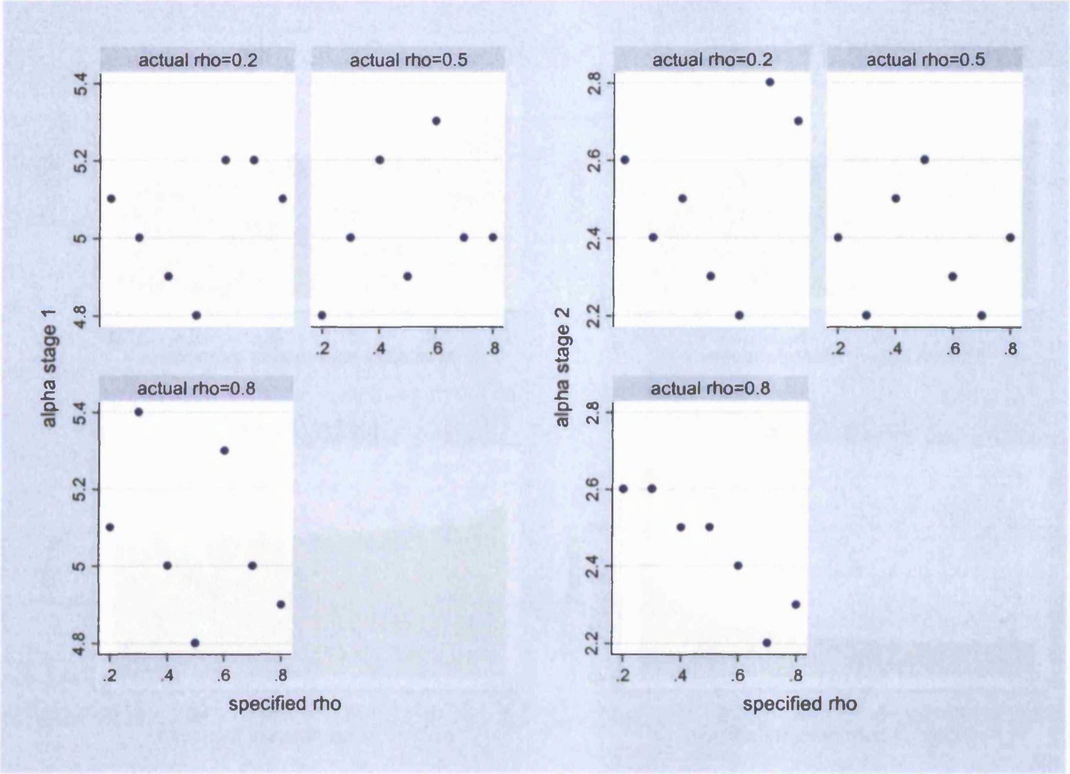


Figure 7-10: Relationship between mis-specified correlation and alpha specified rho - correlation used for sample size calculations, actual rho - correlation used in simulations, alpha stage 1 - significance level for Stage 1, alpha stage 2 - significance level for Stage 2

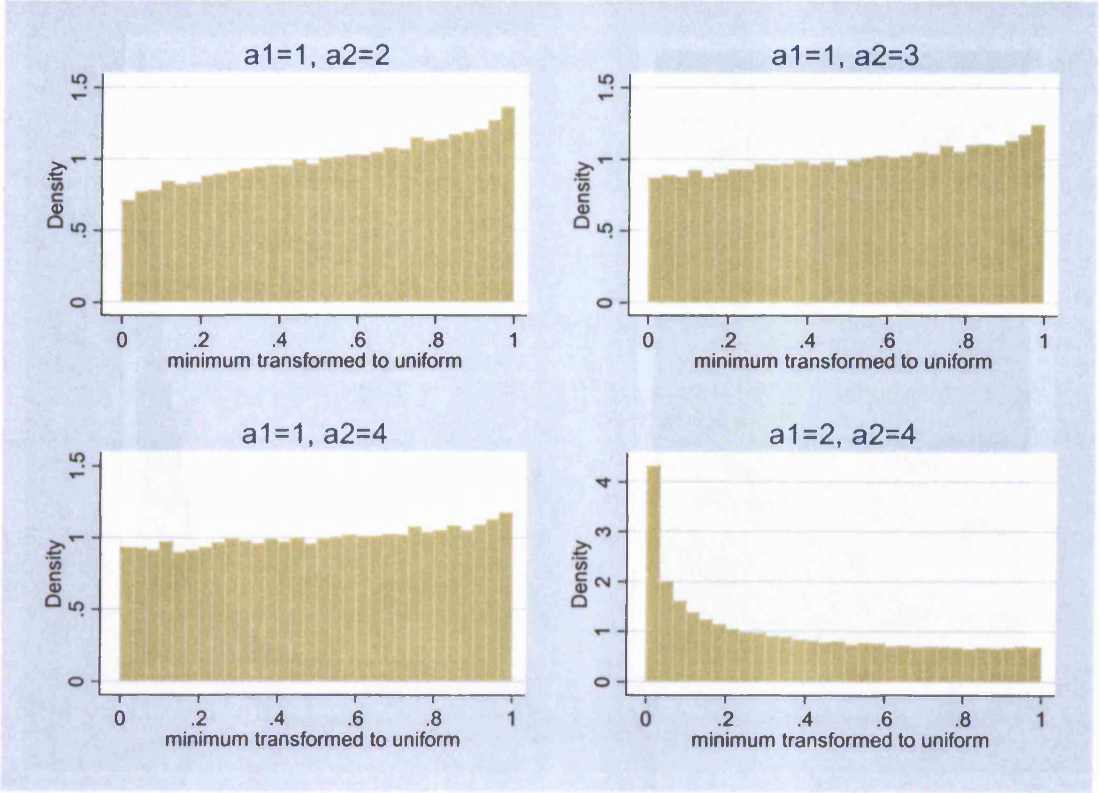


Figure 7-11: Assessment through simulation studies of shape of uniform transformation of min(X,Y) using NBVE results based on 100,000 replications, a1 - median survival for progression-free survival, a2 - median survival for overall survival

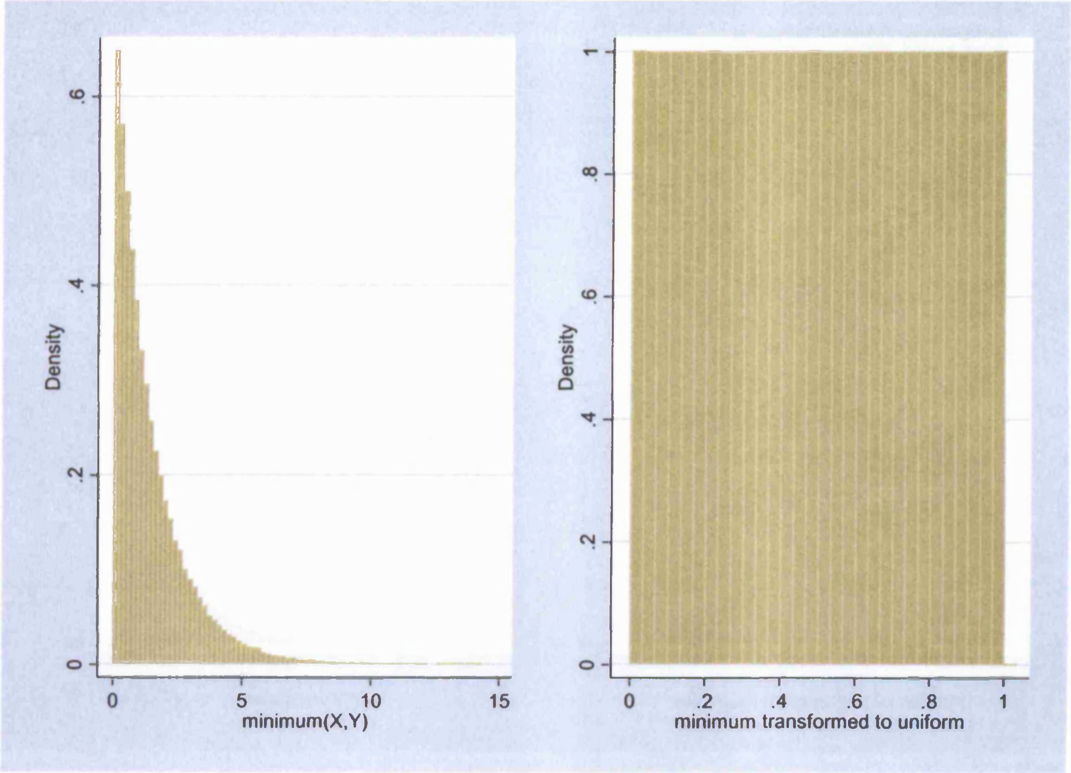


Figure 7-12: Improvement of min(X,Y) as given by NBVE using rejection sampling based on 100,000 replications, median survival of one year for progression-free survival and three years for death

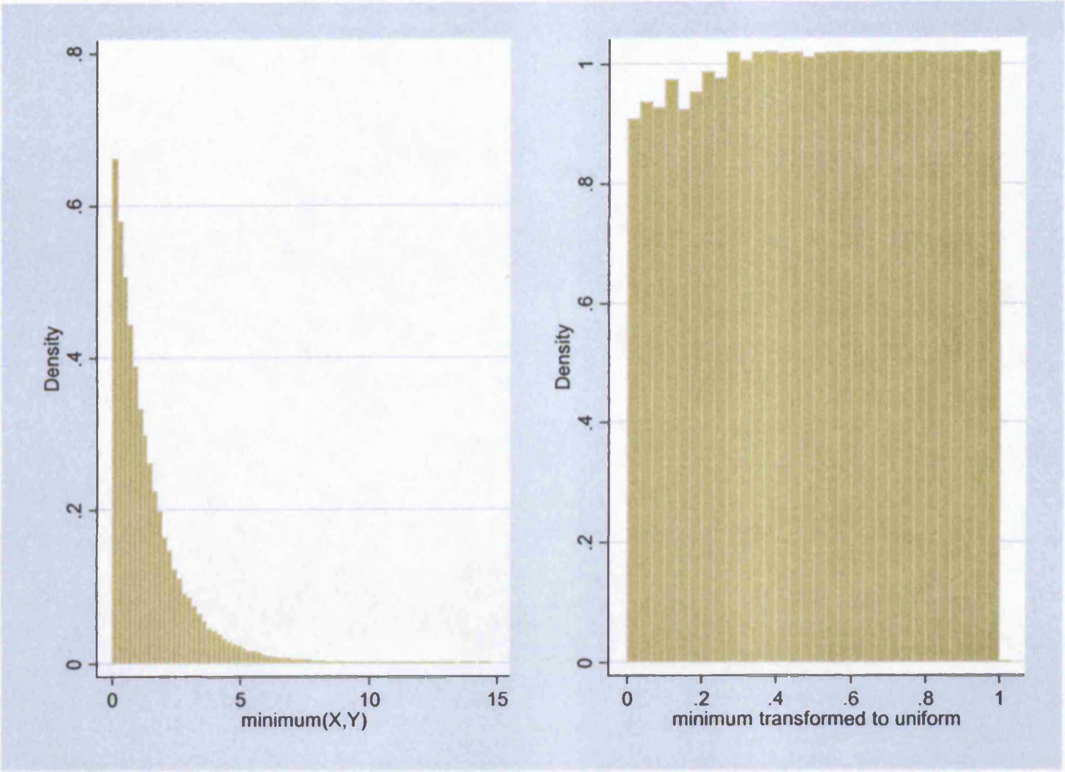


Figure 7-13: Improvement of min(X,Y) as given by NBVE using rejection sampling with relaxed criterion based on 100,000 replications, median survival of one year for progression-free survival and three years for death

Chapter 8

Sample sizes for time-to-event outcomes: implications of the variability in events and time

8.1 Introduction

Sample sizes for trials with a time-to-event outcome are usually derived using three main components: i) the given total duration of the trial consisting of accrual and follow-up time, ii) the survival distribution for the control treatment and iii) the hazard ratio which we hope to see in the experimental arm(s). These lead to the calculation of the number of events as a fixed quantity. Early examples of such sample size calculations were given by George & Desu [50] and Freedman [41]. The following formula given by Schoenfeld [114] forms the basis of most of our calculations. For a given log hazard ratio Δ, a probability ψ of not being censored and a probability p of being allocated to the control treatment group the required sample size is given by

$$N = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{\Delta^2 \psi p(1-p)} \tag{8.1}$$

where $z_{1-\alpha/2}$ and $z_{1-\beta}$ are the normal deviates corresponding to a two-sided significance level α and type II error probability $1-\beta$ respectively.

Korn & Simon [69] examine the problems with the above approach of treating either the number of events or trial time as a fixed quantity. Their analysis of the situation is taken from the perspective of a data monitoring committee. Difficult decisions will have to be made if the variability in the length of the trial or the number of events accrued by a certain time point are not taken into account at the planning stages. In particular they focus on the problems arising if accrual is slower than expected or events accrue slower than anticipated. In these cases modifications to the design need to be contemplated since the trial will have less than the designed power by the time it is meant to be analysed. Gould [52] focuses on the subject of sample size re-estimation in particular whereby sample size is adjusted on the basis of variability. He proposes that the sample size needed for a trial depends on the significance level and power, the magnitude of the log hazard ratio Δ and the variability of the response variable, i.e.

$$\textit{sample size} = \textit{variance} \times f(\textit{error rates})/\Delta^2$$

where

$$f(\textit{error rates}) = (z_{\alpha/2} + z_\beta)^2 \times \textit{ inflation factor}$$

Gould's approach is aimed at dealing with group sequential designs, where the inflation factor is expected to be greater than one. For the standard parallel group trials he proposes it should equal unity. One option suggested by Gould to protect from interim adjustments is to overpower the trial initially. In order to do this, however, one needs to have an idea of the inherent variability of the trial parameters and its effect on power. He suggests not to adjust a trial unless the increase in sample size would be equal to or greater than 33%. While we can avoid the power problem by pre-specifying the number of events to be attained, we need to be aware, however, that this does cause variability in terms of the resources of the trial since the trial duration now becomes a random variable.

This chapter seeks to explore possible ways to assess the potential inherent variability in trial time and / or number of events and provides tools to assess the variability at the beginning of the trial as well as update these estimates throughout patient and event accrual.

8.2 Literature approaches

This topic has so far received very little attention in the literature apart from the angle of sample size re-estimation. We have hence not been able to locate any papers which provide a comprehensive a-priori assessment of the variability in events or time for time-to-event type trials. Shuster [122] considers the topic from the perspective of trials with low event rates. She recommends fixing the number of events rather than the number of patients and models the distribution of the number of events in the treatment group using an unconditional binomial distribution. Sample sizes are then obtained through an iterative solution.

The use of internal pilot studies to assess variability is discussed by Birkett & Day [11]. This primarily addresses the problem of differing patient groups in preceding trials as compared to the current trial. In this setting the first few patients entering the study are used to assess variability and then act as a basis for the overall sample size calculations. However, as the authors point out, this is not a feasible solution in studies where we have long treatment periods and events only start to accrue at a relatively late time-point after enrollment, as for example in many Cancer or HIV trials. Additionally, this approach is problematic since we usually need to know sample size in advance for practical reasons.

8.3 Analysis of variation at the planning stage

We first concentrate on assessing the amount of variability in total trial time at the outset of the trial. We derive the most simple case mathematically in Section 8.3.1 while the rest of the analysis and programming concentrates on simulation studies. The reason for this, as Section 8.3.1 illustrates, is that the resulting distribution is complex and does not lend itself easily to extensions.

8.3.1 Modelling the distribution

Graphs of the distribution of the total length of the trial as displayed in Figure 8-1 suggest an underlying approximate normal distribution. In the following we derive the exact distribution of time to required numbers of events.

Define Y as the time at which a patient is accrued which follows a uniform distribution, i.e. $Y \sim unif(0, t_1)$ where t_1 denotes the time at which the accrual period ends. In addition, define X as the survival time of a patient which is modelled using an exponential distribution, i.e. $X \sim Exp(\lambda)$ where $\lambda = \frac{a}{\log 2}$ in the control group with a being defined as the median survival time. The log hazard ratio in favour of the experimental treatment group is denoted by Δ, that is $\Delta = \log(\lambda_e/\lambda)$ where λ_e is defined as the hazard in the experimental treatment group. Furthermore assume that X and Y are independent of

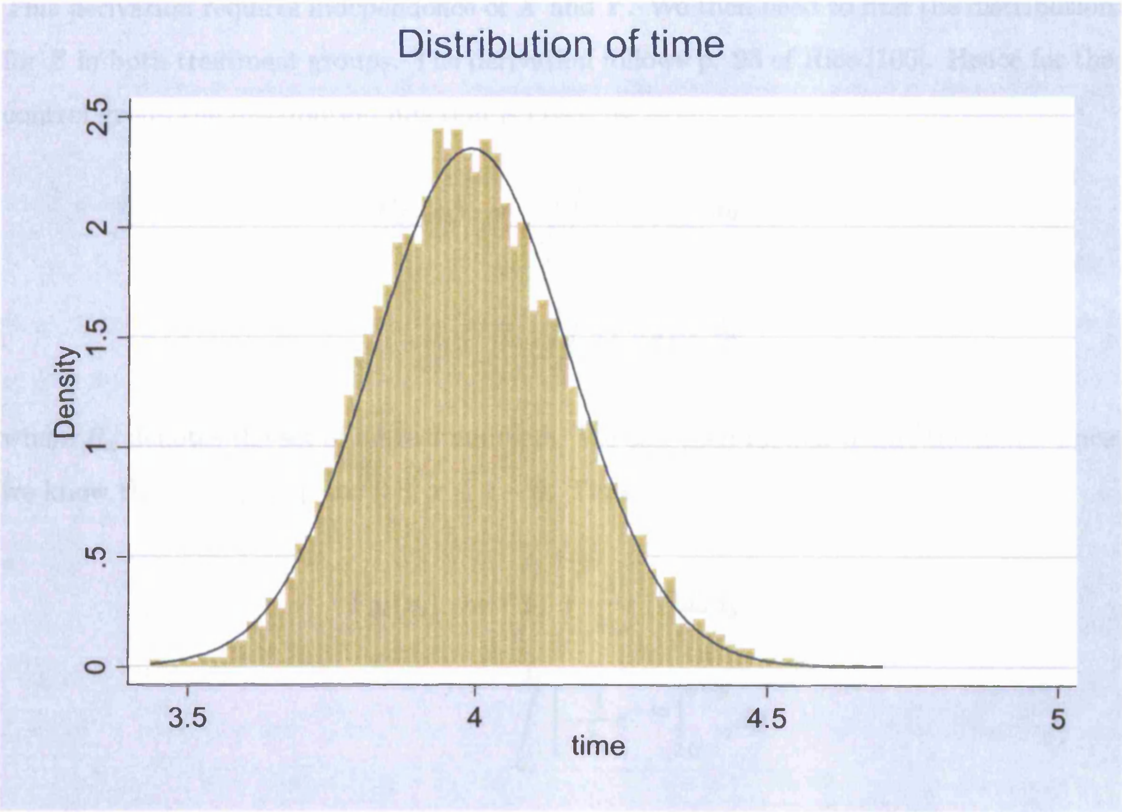


Figure 8-1: Distribution of time to required number of events based on 10,000 replications

each other. Let the total trial time be given by $t_1 + t_2$ where t_2 denotes the length of the follow-up period after recruitment has terminated. The real time at which a patient experiences an event is thus given by $Z = X + Y$.

As a first step we find the joint distribution for X and Y

$$\begin{aligned} f(x,y) &= f_X(x)f_Y(y) \\ &= \begin{cases} \frac{\lambda}{t_1}e^{-\lambda x} & \textit{for control treatment group} \\ \frac{\lambda}{\Delta t_1}e^{-\lambda x/(\exp\Delta)} & \textit{for experimental treatment group} \end{cases} \end{aligned}$$

This derivation requires independence of X and Y. We then need to find the distribution for Z in both treatment groups. The derivation follows p. 93 of Rice [100]. Hence for the control group the distribution function is given by

$$\begin{aligned} F_{Zc}(z_c) &= \iint\limits_{R_z} f(x,y)dxdy \\ &= \iint\limits_{0}^{\infty} f(x,y)dxdy \end{aligned}$$

where R_z denotes the set of all real numbers. We can then further define the limits since we know that $0 \leq y \leq t_1$ and $0 \leq x \leq z - y$. Thus

$$\begin{aligned} F_{Zc}(z_c) &= \int\limits_0^{t_1}\int\limits_0^{z-y} \frac{\lambda}{t_1}e^{-\lambda x}dxdy \\ &= \int\limits_0^{t_1}\left[-\frac{1}{t_1}e^{-\lambda x}\right]_0^{z-y} dy \end{aligned}$$

$$\begin{aligned} &= \int\limits_0^{t_1}\left(-\frac{1}{t_1}e^{-\lambda(z-y)} + \frac{1}{t_1}\right) dy \\ &= -\frac{1}{t_1}\int\limits_0^{t_1}\left(e^{-\lambda z}e^{\lambda y} - 1\right) dy \\ &= -\frac{1}{t_1}\left[e^{-\lambda z}\frac{1}{\lambda}e^{\lambda y} - y\right]_0^{t_1} \\ &= -\frac{1}{t_1}\frac{1}{\lambda}e^{-\lambda z}(e^{\lambda t_1} - 1) + 1 \end{aligned}$$

and

$$\begin{aligned} f_{Z_c}(z_c) &= \frac{dF_{Zc}(z)}{dz} \\ &= \frac{1}{t_1}e^{-\lambda z}e^{\lambda t_1} - \frac{1}{t_1}e^{-\lambda z} \\ &= \frac{1}{t_1}e^{-\lambda z}(e^{\lambda t_1} - 1) \end{aligned}$$

for $0 \leq z \leq t_1 + t_2$. By a similar argument we can derive for the experimental group that

$$F_{Z_e}(z_e) = -\frac{1}{t_1}\frac{(\exp \Delta)}{\lambda}e^{-\frac{\lambda}{(\exp \Delta)}z}(e^{\frac{\lambda}{(\exp \Delta)}t_1} - 1) + 1$$

and

$$f_{Z_e}(z_e) = \frac{1}{t_1}e^{-\frac{\lambda}{(\exp \Delta)}z}(e^{\frac{\lambda}{(\exp \Delta)}t_1} - 1)$$

for $0 \leq z \leq t_1 + t_2$.

Let k denote the total number of events needed out of a total of N patients in the trial. When considering the time to the required number of events we are in effect looking at the time at which the kth patient experiences an event, i.e. the kth order statistic. According to p. 101 in Rice [100] the density of the kth order statistic $Z_{(k)}$ is given by

$$f_k(Z_{(k)}) = \frac{N!}{(k-1)!(N-k)!}f(z)F^{k-1}(z)\left[1 - F(z)\right]^{N-k}$$

where $f(z)$ and $F(z)$ are as defined above for one treatment group only. However, for two groups we need to take into account that the kth event may come from either group, control or experimental. The event $z \leq Z_{(k)} \leq z + dz$ occurs if $k - 1$ observations are smaller than z, one observation is in the interval $[z, z + dz]$, and $N - k$ observations are greater than $z + dz$. This event may occur in either of the two treatment groups. The probability of such an arrangement under an equal allocation ratio to both groups is given in the control group by

$$f(z_c)F^{k-1}(z_c)\left[1 - F(z_c)\right]^{\frac{N}{2}-k} dz_c$$

and in the experimental group by

$$f(z_e)F^{k-1}(z_e)\left[1 - F(z_e)\right]^{\frac{N}{2}-k} dz_e$$

These events are mutually exclusive since if one of the groups provides the kth patient, the other will not be able to. However, this also applies to the previous $(k-1)$ events. We therefore need to sum over all possible scenarios of allocating k events between the two treatment groups. The two extreme scenarios are i) if $k \leq N/2$ all events come from one group and none from the other and ii) if $k > N/2$ k events come from one group and $k - N/2$ from the other. In order to express all scenarios, we assume that i events occur in one group and $k-i$ in the other. Using the multiplication principle, we have if $k \geq N/2$

$$\sum_{i=0}^{\frac{N}{2}} \frac{(\frac{N}{2})!}{(k-i)!1!\left(\frac{N}{2}-i\right)!} * \frac{(\frac{N}{2})!}{(k-i-1)!1!\left(\frac{N}{2}-k+i\right)!}$$

and if $k < N/2$

$$\sum_{i=0}^{k} \frac{(\frac{N}{2})!}{(k-i)!1!\left(\frac{N}{2}-i\right)!} * \frac{(\frac{N}{2})!}{(k-i-1)!1!\left(\frac{N}{2}-k+i\right)!} \qquad (8.2)$$

such possible arrangements over the trial population as a whole. Hence the density of $Z_{(k)}$ is then for $k \geq N/2$ given by

$$\begin{aligned} f_k(Z_{(k)}) &= \sum_{i=0}^{\frac{N}{2}} \frac{(\frac{N}{2})!}{(k-i)!\left(\frac{N}{2}-i\right)!} * \frac{(\frac{N}{2})!}{(k-i-1)!\left(\frac{N}{2}-k+i\right)!} \\ &\quad *F^{k-1}(z_c)\left[1-F(z_c)\right]^{\frac{N}{2}-k} dz_c \\ &\quad *F^{k-1}(z_e)\left[1-F(z_e)\right]^{\frac{N}{2}-k} dz_e \\ &\quad *(f(z_c)+f(z_e)) \qquad (8.3) \end{aligned}$$

and similarly for $k < N/2$ using Equation 8.2. However, this distribution underlies stringent simplifying assumptions, e.g. in reality the last patient needed may arrive just before or after $t_1 + t_2$.

The close resemblance of the distribution of time to a normal distribution as displayed in Figure 8-1 may be explained by the similarities between the exact distribution of time to the end of the trial as given in Equation 8.3 and a negative binomial distribution (Rice, p. 38 [100]). Just as in the case of a negative binomial distribution, the distribution of time to the end of the trial will follow the normal approximation more closely as k increases, i.e. as the number of events needed in the trial becomes larger.

8.3.2 Simulation methods

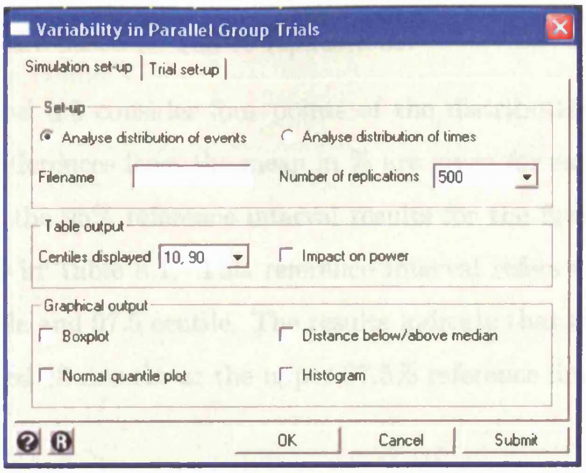
A Stata 8 program `varsim - simulation` with an accompanying dialog was written which analysis variability in events and time for both single stage parallel group trials and multi-stage, multi-arm trials. One of the dialog menus and the program output are given in Figure 8-2. This part of the program is based on simulations run using sample sizes calculated for the input parameters. In the case of parallel group trials the sample size is given by ART (Analysis of Resources for Trials) with calculations based on Barthel et al. [7] which are described in more detail in Chapters 3 and 4. For multi-stage trials simulations are run using sample size calculations based on an extension of Royston et al. [103] as illustrated in Chapter 6. The user may specify trial parameters as well as the type of analysis as illustrated on the dialog window in Figure 8-2. In general, tabular output is given for the distribution of either time or events as well as the parameters chosen to calculate sample size for the trial. Graphics then include a histogram, a boxplot and an assessment of normality.

All of the possible simulation set-ups employ a uniform accrual mechanism and exponential survival where the exponential distribution is parameterised using median survival and the hazard ratio as specified by the user. The set-up for the survival distribution in multi-stage trials is based on the Normal Bivariate Exponential (NBVE) distribution whose characteristics are described in Chapter 7. Hence, as is described in that chapter, simulation results are more accurate for the last stage of any trial than for the previous ones. The program allows one to specify a wide range for the number of replications to be used in the simulations; however, we recommend the use of 5,000 to 10,000 replications to ensure sufficient accuracy and speed.

This simulation method may be extended to bring it in line with the options implemented in ART as described in Chapter 4. This would then allow for more complex survival functions, loss to follow-up and cross-over.

8.3.3 Some important results

Tables 8.1, 8.2 and 8.3 illustrate a simulation study of the variability in time for a parallel group trial comparing a control with one experimental treatment. Common to all trials are an accrual and follow-up time of two years. The hazard ratio in each of the trials is given by HR and p denotes the probability of being allocated to the control treat-



● Analyse distribution of events ○ Analyse distribution of times

Filename:

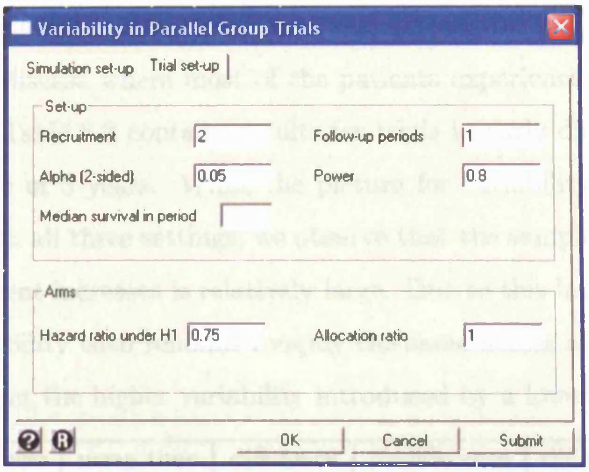
Number of replications: 500

Centiles displayed: 10, 90

□ Impact on power

□ Boxplot □ Distance below/above median

□ Normal quantile plot □ Histogram



Recruitment: 2

Follow-up periods: 1

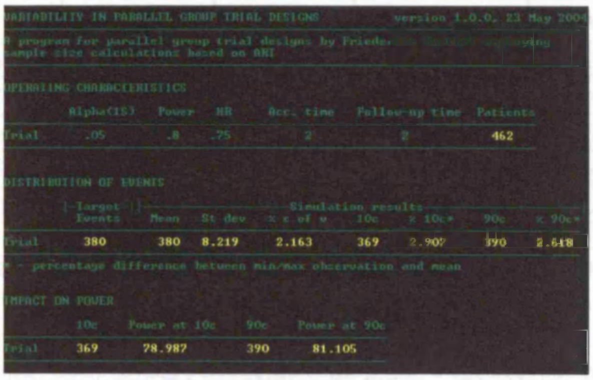
Alpha (2-sided): 0.05

Power: 0.8

Median survival in period:

Hazard ratio under H1: 0.75

Allocation ratio: 1



```
VARIABILITY IN PARALLEL GROUP TRIAL DESIGNS          version 1.0.0, 23 May 2004
A program for parallel group trial designs by Friede
sample size calculations based on ART

OPERATING CHARACTERISTICS

        Alpha(1S)   Power   HR     Acc. time   Follow-up time  Patients
Trial     .05         .8   .75        2              2           462

DISTRIBUTION OF EVENTS

         |-Target-||-----------------Simulation results----------------------
          Events     Mean  St dev    % c of v    10c    % 10c*     90c    % 90c*
Trial      380        380  8.219      2.163      369    2.907      390    2.618
* - percentage difference between min/max observation and mean

IMPACT ON POWER

          10c    Power at 10c    90c    Power at 90c
Trial     369      78.987        390      81.105
```

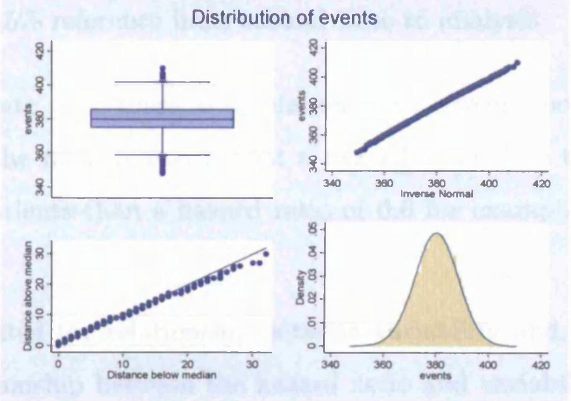


Figure 8-2: Output from Stata program Varsim - Simulation

ment group. The trials were designed to attain 90% power and a 5% two-sided level of significance. Results are based on 10,000 replications.

Tables 8.1, 8.2 and 8.3 consider four points of the distribution of time to required number of events. Differences from the mean in % are given for each of these points. As an example consider the 95% reference interval results for the first simulated trial with a hazard ratio of 0.6 in Table 8.1. This reference interval refers to the range of values between the 2.5 centile and 97.5 centile. The results indicate that this trial could take 54 instead of the expected 48 months at the upper 97.5% reference limit, an increase of half a year.

All three tables contain results for different disease stages. Table 8.1 considers a trial setting in advanced disease where most of the patients experience an event by the end of the trial whereas Table 8.3 contains results for trials in early disease, in this case for median survival time of 3 years. While the picture for variability in terms of time to events looks similar in all three settings, we observe that the sample size increase needed as median time to event increases is relatively large. Due to this large increase in sample size percentage variability then remains roughly the same across all tables as the higher sample size cancels out the higher variability introduced by a lower event rate.

p	HR	sample size	mean time to analysis	difference at 2.5%	% difference at 2.5%	difference at 97.5%	% difference at 97.5%
0.5	0.6	154	3.99	- 0.46	- 11.4	0.52	13.0
1/3	0.6	167	4.01	- 0.46	- 11.6	0.50	12.3
0.5	0.7	305	4.00	- 0.34	- 8.4	0.36	9.0
0.5	0.8	758	3.99	- 0.22	- 5.4	0.22	5.6
0.5	0.9	3327	4.00	- 0.10	- 2.5	0.10	2.6

Table 8.1: Simulation results for distribution of time to end of trial (or analysis of trial results) in advanced disease i.e. median survival one year p - allocation ratio, HR - hazard ratio in favour of experimental group, sample size - sample size required for 90% power, difference at 2.5% - difference between mean time to analysis and lower 2.5% reference limit around time to analysis

All tables illustrate that there is a relationship between the hazard ratio and the variation in time. The main reason is that a hazard ratio close to one requires a much higher number of patients than a hazard ratio of 0.6 for example and hence variability decreases.

We also investigated the relationship between variability and power. Figure 8-3 displays both the relationship between the hazard ratio and variability as well as between power and variability in more detail. One may observe that the coefficient of variation

p	HR	sample size	mean time to analysis	difference at 2.5%	% difference at 2.5%	difference at 97.5%	% difference at 97.5%
0.5	0.6	221	4.02	- 0.44	- 10.9	0.48	12.0
1/3	0.6	235	4.02	- 0.44	- 10.9	0.47	11.6
0.5	0.7	431	3.99	- 0.31	- 7.8	0.33	8.2
0.5	0.8	1054	4.00	- 0.20	- 4.9	0.20	5.0
0.5	0.9	4561	4.00	- 0.10	- 2.4	0.10	2.4

Table 8.2: Simulation results for distribution of time to end of trial (or analysis of trial results), median survival two years p - allocation ratio, HR - hazard ratio in favour of experimental group, sample size - sample size required for 90% power, difference at 2.5% - difference between mean time to analysis and lower 2.5% reference limit around time to analysis

p	HR	sample size	mean time to analysis	difference at 5%	% difference at 5%	difference at 95%	% difference at 95%
0.5	0.6	294	4.01	- 0.43	- 10.7	0.46	11.4
1/3	0.6	310	3.99	- 0.43	- 10.7	0.47	11.8
0.5	0.7	567	4.00	- 0.30	- 7.5	0.32	8.0
0.5	0.8	1378	4.00	- 0.19	- 4.8	0.20	5.0
0.5	0.9	5992	4.00	- 0.09	- 2.3	0.10	2.4

Table 8.3: Simulation results for distribution of time to end of trial (or analysis of trial results) in early disease i.e. median survival three years p - allocation ratio, HR - hazard ratio in favour of experimental group, sample size - sample size required for 90% power, difference at 2.5% - difference between mean time to analysis and lower 2.5% reference limit around time to analysis

decreases linearly with increasing power while there appears to be a relationship of exponential decay between the coefficient of variation and an increasing hazard ratio. This is to be expected since an increase in power, similar to an increase in the hazard ratio, leads to an increase in the sample size required for the trial. As we noted above, this increase in sample size leads to a decrease in variability of the total trial time.

8.4 Updating estimates using trial data

Once a clinical trial has started accruing patients and events, trialists may wish to obtain up-to-date estimates of the time by which the trial will have accrued the necessary number of events. This may deviate from earlier estimates since these were likely to be based on previous trial results, experimental data for the new treatment regimen or Phase II trials and not the current trial.

Bagiella & Heitjan [5] consider the prediction of analysis times in the context of trials with planned interim analyses. They introduce two model based approaches. The first

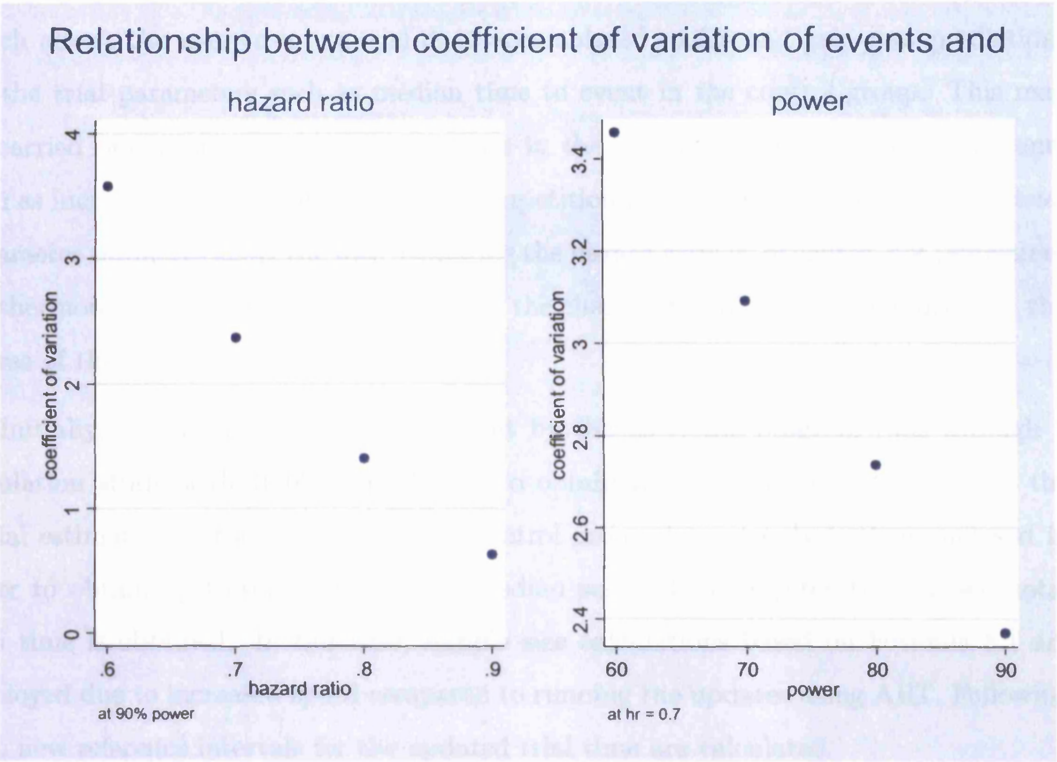


Figure 8-3: Variability relationships
results illustrated for equal allocation to both treatment arms and median survival of one year

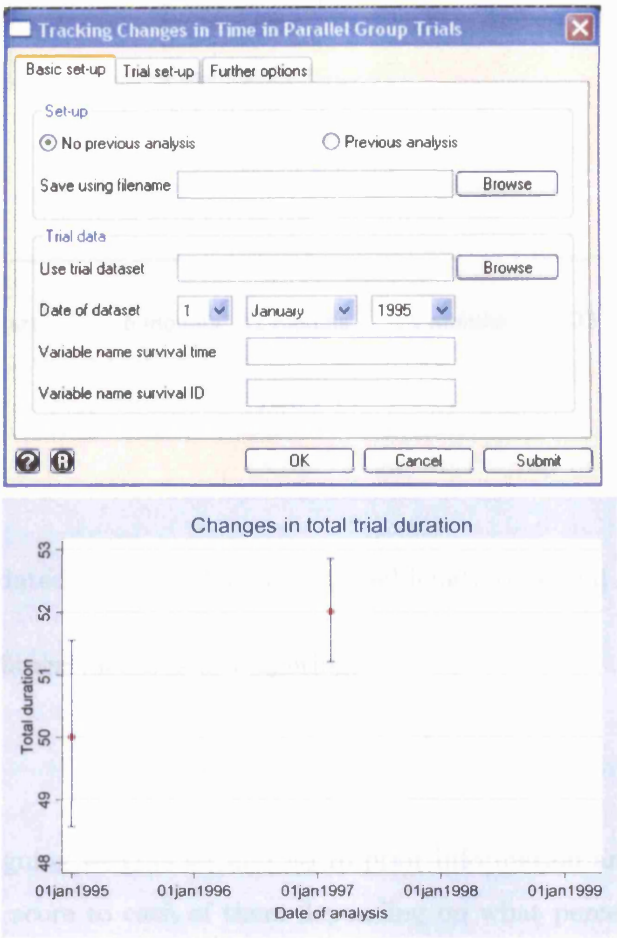
is based on a point prediction of analysis time by extrapolating the cumulative mortality into the future and selecting the date at which the expected number of deaths is equal to the required number of events. The second method uses a Bayesian simulation scheme to generate a predictive distribution of milestone times. Prediction intervals are then given by the quantiles of that distribution. Drawbacks of their methods include potential bias if the underlying accrual and failure time distributions differ from prior assumptions as well as the assumption of a constant accrual rate.

We have developed a software tool **`varsim - trial update`** shown in Figure 8-4 which allows the user to input trial data accumulated so far and gain new predictions for the trial parameters such as median time to event in the control group. This may be carried out at several time points, since in the case of longer term trials incidents such as increased advertising of a trial or competition from other trials may alter the trial parameter estimates along the way. Updating the hazard ratio is, of course, not permitted. Furthermore, a graphical tool allows to plot the changes in projected variability over the course of the trial.

Initially, based on the parameters input by the user, the program runs through a simulation study with 10,000 replications to obtain the reference intervals around the initial estimates for the trial time. The control group dataset only is then analysed in order to obtain up to date estimates of median survival. Using iterations a new total trial time is obtained. In this case, sample size calculations based on Formula 8.1 are employed due to increased speed compared to running the updates using ART. Following this, new reference intervals for the updated trial time are calculated.

8.4.1 Point estimates

In the following we explore a few possible ways of incorporating trial information into our parameter estimates. All variables are defined as described in Sections 8.1 and 8.3.1. Figure 8-5 illustrates the process of updating the estimated trial time $t_1 + t_2$. Circles give point estimates with 95% reference intervals around them. At the end of the trial, in this case at 33 months, we will have a point estimate only as this corresponds to a calculation of trial time at the end of the trial itself. The point estimate at the start is based on previous information only while the last point estimate at 33 months is based on trial data only. Estimates in between are based on a combination of trial data and prior information. This section considers the combination of prior information and trial



Set-up

● No previous analysis ○ Previous analysis

Save using filename

Trial data

Use trial dataset

Date of dataset 1 January 1995

Variable name survival time

Variable name survival ID

Figure 8-4: Output from Stata program Varsim - Trial Update with data taken from ICON3

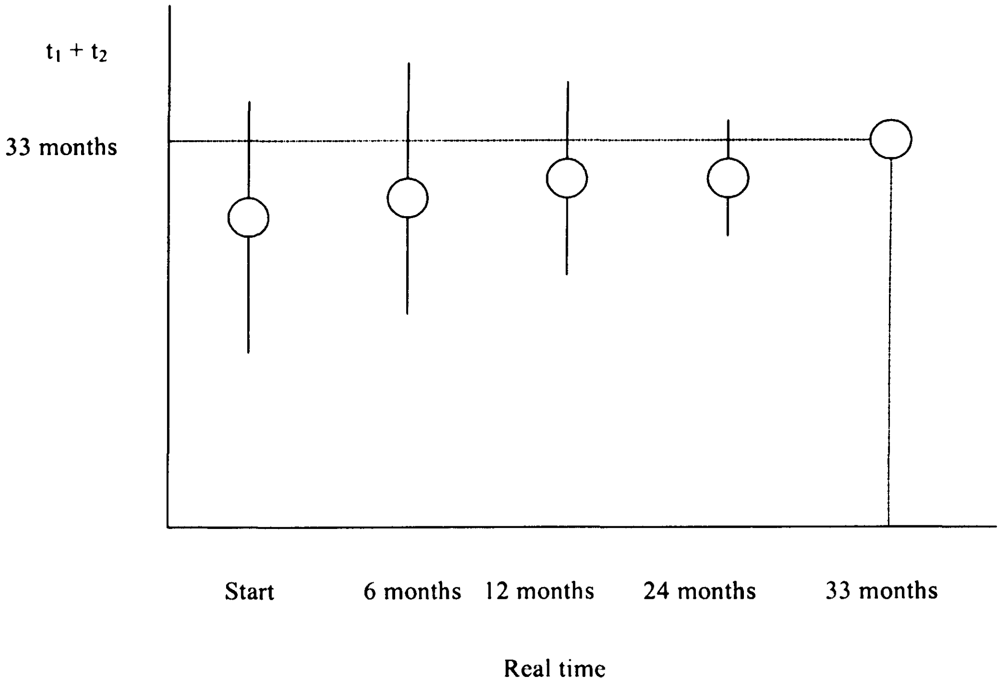


Figure 8-5: Updated estimates for the required length of a trial using trial data

data using three different methods of weighting.

Weights

A crude way of assigning weights w_1 and w_2 to prior information and trial data respectively is to assign a score to each of them depending on what percentage of the overall number of events required has been obtained at the time of analysis. This loosely follows the method of Tan et al. [129]. Define a_{prior} as the estimate of median survival in the control group obtained from prior information and a_{data} as the estimate of median survival in the control obtained by analying data on the present trial. Thus if we are updating estimates for the median time to event a we get

$$a_{estimated} = w_1 a_{prior} + w_2 a_{data}$$

where the weights $w_1 = 1 - w_2$ and

$$w_2 = \frac{\#\ events\ obtained\ so\ far}{total\ \#\ events}$$

The reason for combining the estimates in that way is that early trial events may be accumulated from patients who are sicker, and therefore may bias estimates of median time to event. Overall, our parameters early on in the trial would be based on very little information if we did not include prior knowledge. As the trial continues, more and more weight is assigned to trial data and prior information becomes less influential. This method has been implemented in our Stata program.

The above crude method may be improved upon by making the weight w_2 inversely dependent on the variance of the estimates, e.g. the variance of the median time to event a, and then still assigning $w_1 = 1 - w_2$ as the weight for prior information [3]. This method may give a realistic and data dependent weighting structure, however, at the same time w_2 may increase and then decrease again over the course of the trial.

A third possibility is to view the situation as a 'missing data' problem [80]. In that case we can use weights which are inversely related to the probability of the data being observed similar to Preissler et al. [98].

Our initial simulation studies around the initial knowledge of the total trial time provide us with the percentage variation on either side of the median total trial time. In order to obtain reference intervals for each of the new estimates, this percentage variation is reduced by the percentage amount of weighting w_2. Thus if we define v as the percentage variation on either side of the median total trial time we have

$$v_{new} = v_{old} - (v_{old} * \frac{w_2}{100})$$

8.5 Conclusions

From the analysis conducted in this chapter it is apparent that it is important to take into account the variability in trial duration since it may have a significant impact on the total length of the trial. The distributions of both the variability in events and time are wider for smaller trials. One approach in these situations may be to design for the lower bound of the reference interval around the number of events to ensure adequate power by the expected end of the study. Similarly, it makes sense in those cases to consider the upper bound of the reference interval around trial length to anticipate possible higher trial costs and to ensure adequate funding in advance. This knowledge is of particular importance in multi-stage trials where the stages itself are often relatively short and have a smaller sample size. It is necessary to realise that this degree of variability already arises if we have estimated all other parameters correctly when calculating sample sizes, a scenario which in many instances proves not to be the case.

These methods may be extended to incorporate options provided in ART, such as loss to follow-up and cross-over, as outlined in Chapter 4. Furthermore, it may be of interest to examine the impact of non-uniform accrual on the variability, especially when initially mis-specified. These are all likely to add uncertainty to the time at which the planned number of events are likely to be observed. Explicit assessment of these may give a more realistic and appropriate timeframe and sample size for studies.

The second part of the program currently utilises trial data from the control group only. However, in a number of trials only the data from all groups as a whole may be available due to issues of unblinding. Thus it may be useful to allow for the input of the dataset as a whole in order to obtain an overall median survival. By assuming the hazard ratio used for sample size calculations at the outset we may then calculate median survival in the control group only and employ the tool as described above.

Chapter 9

fintplot: Forest plots for interaction

9.1 Motivation

During the course of examining a particular treatment in a trial setting we often want to know the consistency of an observed relationship across two or more subgroups of patients in the study. We might suspect that a treatment works better in older patients compared with younger ones or that due to the genetic make-up of men and women there is a difference in its effect on the two genders. In the medical literature this type of heterogeneity is often referred to as synergy whilst in statistics we know it under the name interaction [4]. Examining the relationship can be helpful later when developing guidance on how to use that particular treatment in practice. One study that is currently being conducted by the MRC Clinical Trials Unit seeks to identify an interaction of the prevalence of the mutated gene p53 and the results of chemotherapy in colorectal cancer patients.

As outlined by Shuster et al. [123], tests for such interactions can have two uses. By retrospective analysis of possible interaction effects one can formulate interesting hypotheses for future trials. In planning a prospective trial, one may incorporate a test of an interaction effect if it is suspected that the therapies manipulate important factors differently. Hence the analysis of interactions in a trial or study can either be of an exploratory nature or consist of a test for interactions as defined in the protocol.

Our estimate of the interaction effect is based on a ratio of hazard ratios or a ratio

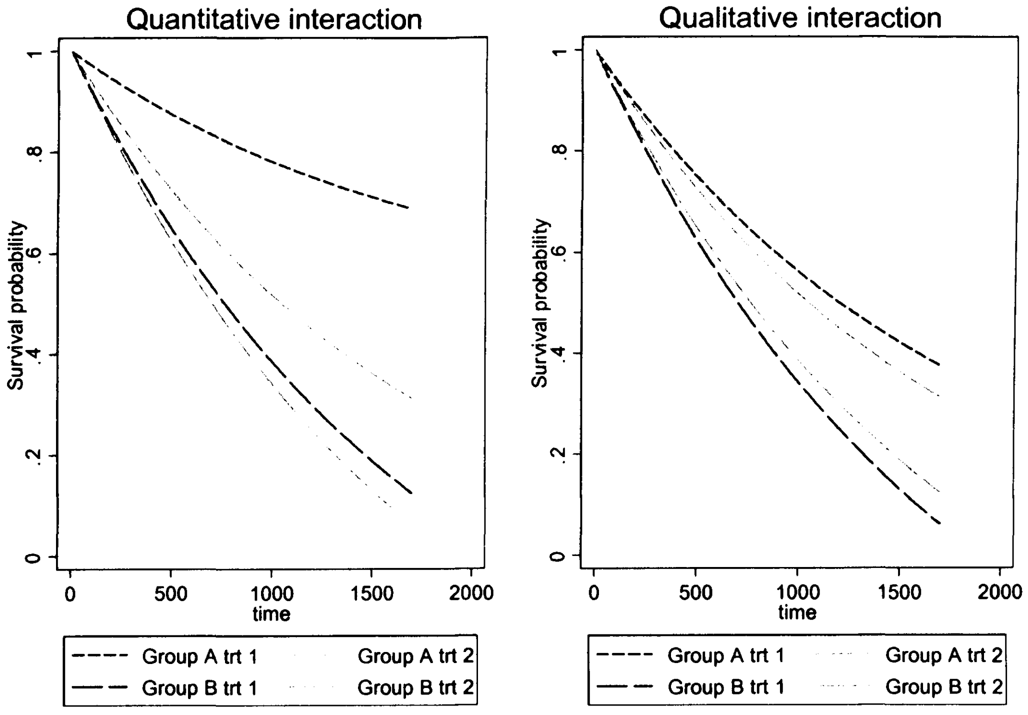


Figure 9-1: Quantitative and qualitative interactions illustrated using Kaplan-Meier survival curves

of odds ratios derived from a 2*2 table as described in Section 9.4. The definition is similar to that of Peterson et al. [94]. This ratio of hazard ratios describes quantitative interactions. A Stata program has been designed in order to ease the visualisation of interactions during the analysis of a clinical trial or study. It provides both numerical and graphical output in the form of a forest plot for this purpose while giving a choice of employing either the Cox proportional hazards model or logistic regression.

9.2 Treatment of interactions in the literature

Since it has been recognised that treatments might have different effects on subgroups in a clinical trial, attention has been devoted to the development of tests for such interactions. Most authors concentrate on one kind of interaction, either a quantitative or a qualitative one. A good illustration of the differences was provided by Byar [14]. In the case of a quantitative interaction, the magnitude of the treatment effect may vary with a patient's characteristics, however, the direction of the treatment effect will stay the same as illustrated in the first panel of Figure 9-1. By contrast, with a qualitative interaction a change in the direction of the treatment effect is involved as the second panel of Figure 9-1 shows. This type of interaction is also often referred to as cross-over or a reversal.

Gail & Simon [44] presented a test for qualitative interactions based on likelihood ratios. A more recent paper by Piantadosi & Gail [95] compared this likelihood ratio test with a range test and they found that whilst the likelihood ratio test has greater power when the new treatment is harmful in several subsets, the range test will have greater power when the new treatment is only harmful in a few subsets. Both computations do, however, indicate that even a very large trial would not have enough power to detect statistically significant occurrences of cross-over.

An approach based on proportional hazards regression models was proposed by Thall & Lachin [131]. The model was then applied to a clinical trial in prostate cancer in order to find the optimal treatment for a patient's set of covariates. Models based on pre-stratification and non-stratification were derived. Another possible approach developed by Uesaka [138] utilizes logarithmic generalised odds ratios. He states that even in the case of a sample size of 20 patients, power would be high enough for this test.

Pan & Wolfe [90] generalised interaction tests to the more practical problem of detecting an interaction effect which corresponds to a minimal treatment difference of clinical significance. They grade possible interaction effects in three classes. The first is the case where a treatment is superior to another across all subsets, which means that this case includes quantitative interactions. In the second class we have a slight qualitative interaction which means that one of the treatments is superior to the other across some subsets and is only inferior by a small amount d for the remaining subsets. The third class contains a severe qualitative interaction. In this case the reversal of treatment effects is so great that even the addition of d to the addition of the effect of the inferior treatment will not make it uniformly superior to the other across all subsets. Pan & Wolfe believe that most trials will be able to detect the second class and hence the alternative hypothesis should be that of a severe qualitative interaction. The test developed in their paper centres around confidence intervals about a clinically significant interaction d.

The situation of 2*k factorial experiments was examined by Xiang et al. [145] whose test statistic is based on a weighted residual sum of squares. In order to estimate the parameters of the test statistic they employe the Mantel-Haenszel and maximum likelihood methods.

Bayesian subset analysis is suggested by Simon [125]. In his approach the subset specific treatment effects are being estimated as an average of overall differences and observed within-subset differences. Following that the two components are weighted by an a-priori estimate of the likelihood of qualitative treatment by subset interactions. Hence this enables statisticians to incorporate an a-priori belief that qualitative interactions are unlikely. Underlying the approach is the proportional hazards model and prior distributions for the interaction effects are assumed to be normal and independent. Simon outlines an application of gender/treatment interactions.

Caution regarding such tests was expressed by Byar [14], shortly after the design by Gail & Simon [44] was made public. He believed that these tests need to take into account the fact that multiple comparisons are being made, and that therefore we need to ensure adequate power. Furthermore he suggests that interactions should be looked at in the context of exploratory analysis rather than that of formal hypothesis testing. Arguments such as this outline the need for rigorous sample size calculations to ensure adequate power of the tests. One such sample size calculation is provided by Schmoor et al. [112]. From their calculations we can deduce that ordinary sample sizes for a parallel group trial would have to be multiplied by a factor of four under equal allocation ratios both in the treatment and covariate groups in order to attain adequate power for such interaction tests.

9.3 Analysis of trials with treatment-covariate interactions present or suspected

The following analyses of trials with possible interaction effects were run to gain an understanding of the magnitude of interaction effects and the best way to represent these. Analyses were run using both the Cox proportional hazards model and the logrank test as well as Kaplan-Meier survival curves. To run the formal interaction analysis, an interaction variable of treatment and a covariate was created.

9.3.1 AXIS

From the AXIS trial 396 patients were selected to participate in this analysis. Sample size for this study was restricted due to cost and practical issues. Patients which were included in the analysis had to have been randomised before 1st January 1995 and had to have curatively resectable Duke's B or C tumours in primary colon cancer. The main trial compared the effect of postoperative portal vein infusion of fluorouracil (5FU) to the effect of no infusion in patients who underwent a planned resection of colon or rectal cancer. Due to cost and practical issues the number of patients in this study was restricted to 400 patients [6]. Apart from survival time in the two groups data was collected on LOH at p53, DP1, D18S61and D18S851, nmifinal, DNA ploidy, sex, age and Duke's stage B or C. Furthermore, a combination variable was formed out of p53, D18S61 and D18S851 and named hetany2. Hence the test for potential treatment-covariate interactions was pre-specified in the protocol. This is generally desireable to avoid multiple testing issues. If not pre-specified, interactions may be analysed for covariates which are thought to be relevant but particular care has to be taken to adjust the type I error.

During the AXIS trial 171 out of 396 patients died. The overall hazard ratio of the trial was observed to be 0.73 which indicates a reduction in the risk of death of 27% following 5FU with a confidence interval ranging from 0.54 to 0.98 and a significance level of 0.038.

We can identify hetany2, P53, D18S61, D18S851, nmifinal, DNA ploidy and Duke's stage as having potential interaction effects with treatment. Each of the first five variables are split into three categories consisting of retained heterozygosity, loss of heterozygosity and not informative. The distribution of patients among the categories for four of the variables is portrayed in Table 9.1. Both D18S61 and D18S851 have very similar distributions with the least number of patients in the first category and the other two categories containing roughly equal numbers. DNA ploidy is split into the two categories of roughly equal numbers. These discrepancies in the distribution of patients among the categories call for caution in the later interpretation of the analysis due to a lack of power. The Kaplan-Meier survival curves in Figure 9-2 for the combination of treatment and Duke's stage illustrate why we might suspect an interaction effect in this case. As we can see in the graph, patients with Duke's stage B generally have better survival rates regardless of whether they are in treatment group 1 or 2 in comparison to patients with Duke's stage C.

Covariate	retained heterozygosity	loss of heterozygosity	N/A
hetany2	157	159	80
p53	40	93	250
nmifinal	279	89	28
Duke's stage	240	156	-

Table 9.1: Distribution of patients among categories for the covariates in AXIS

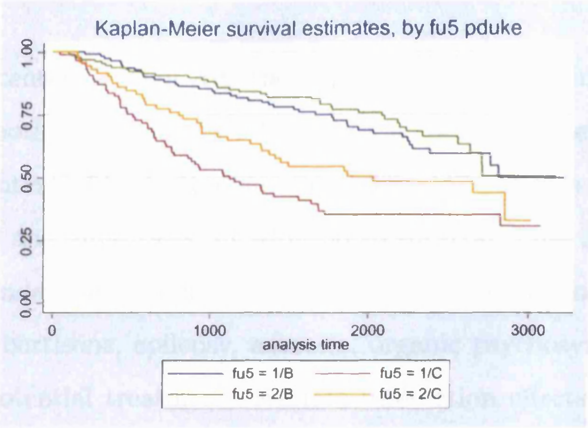


Figure 9-2: Kaplan-Meier survival estimates, by treatment (fu5) and Duke's stage (pduke) in the AXIS study

While looking at possible interaction effects only the effect for hetany2 was found to be significant at the 5% level (p-value 0.03) with a hazard ratio for the interaction effect of 0.74 and a confidence interval ranging from 0.55 to 0.97. The Kaplan-Meier survival curve for the interaction of hetany2 with treatment is illustrated in Figure 9-3. This plot was created by multiplying the treatment and the covariate indicator and then plotting the Kaplan-Meier survival curves for each of the three categories. We can see that one of the categories (fu5_hetany2 = 1) lies above the other two and hence indicates a potential interaction effect.

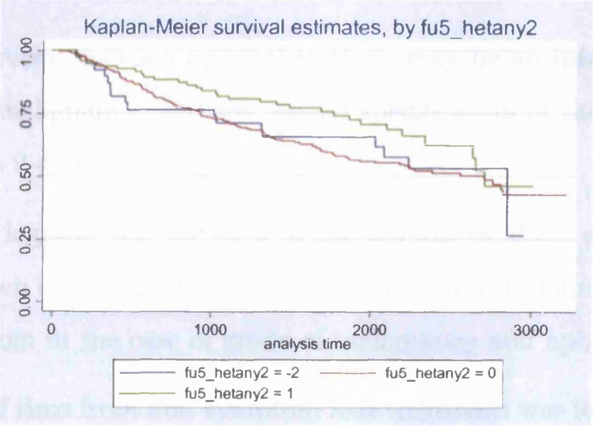


Figure 9-3: Kaplan-Meier survival estimates for interaction between treatment and hetany2 (fu5_hetany2) in the AXIS study

9.3.2 Glioma2

Glioma2 was a multicenter German-Austrian randomised trial conducted to test the standard therapy of Monotherapy with BCNU against a combined chemotherapy of BCNU and VM26 in the context of brain tumours in adults. 447 patients were randomised between February 1983 and June 1988. In addition to survival times data was collected on age, sex, Karnofsky-index, time from first symptom, grade of malignancy, type of surgical resection, convulsia, cortisone, epilepsy, amnesia, organic psychosyndrome and aphasia. Again the test for potential treatment-covariate interaction effects was pre-specified in the protocol.

During the trial 274 out of 411 patients died. The overall hazard ratio of the trial was observed to be 0.89 in favour of chemotherapy with a confidence interval ranging from 0.71 to 1.14 and a significance level of 0.38. Hence there was no evidence of a significant improvement in survival depending on treatment.

We can identify the time from first symptom, grade of malignancy, Karnofsky index and aphasia as possible interaction candidates. Investigation of these variables was done by Ulm et al. [139] and Sauerbrei [109]. Each of these variables has been split into two levels, with the Karnofsky index having two different level definitions. The grade of malignancy and the second definition of the Karnofsky index show big discrepancies in the numbers of patients present in each group. Therefore power for the comparison is relatively low.

Kaplan-Meier survival curves suggest that there may be an interaction especially in the case of grade of malignancy and the second specification of the Karnofsky-index as illustrated in Figures 9-4 and 9-5 respectively.

When running a logrank test for each of the covariates alone as prognostic factors, the differences between the categories in terms of survival were found to be significant at the 5% level apart from in the case of grade of malignancy and aphasia.

The interaction of time from first symptom and treatment was found to be significant at the 5% level (p-value 0.03) with a hazard ratio for the interaction term of 0.58 and a confidence interval from 0.35 to 0.96 which is very wide. Similarly both specifications of the Karnofsky index were found to have a significant interaction with treatment (p-values of 0.002 and 0.031) and very similar interaction hazard ratios of 0.64 and 0.66 (CI 1: 0.49 - 0.82, CI 2: 0.49 - 0.89). The Kaplan-Meier survival curves for the interaction term also

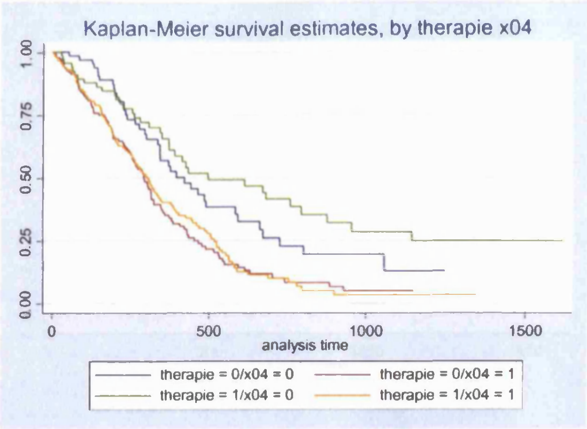


Figure 9-4: Kaplan-Meier survival estimates, by treatment (therapie) and grade of malignancy (x04) in the Glioma2 study

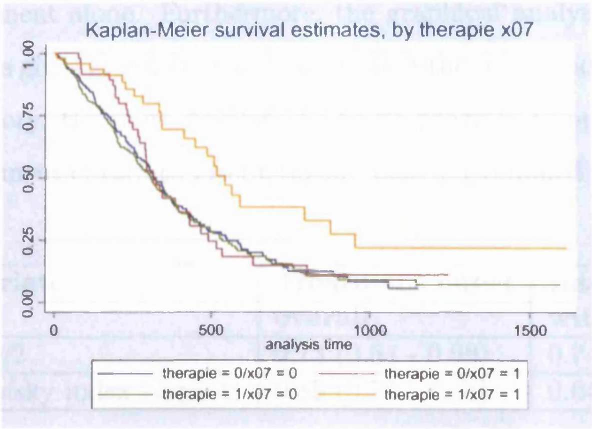


Figure 9-5: Kaplan-Meier survival estimates, by treatment (therapie) and Karnofsky index (type 2) (x07) in the Glioma2 study

follow a very similar path, with the one for the Karnofsky index (type 1) shown in Figure 9-6.

The interactions of grade of malignancy and aphasia with treatment were, however, not found to be significant (p-values 0.37 and 0.39).

9.3.3 Summary of main effects

Table 9.2 provides a summary of the main treatment effect in each trial and one of the major interaction effects of treatment with a covariate. What is apparent is that having done the main analysis we are interested in interaction effects of each of the levels of the

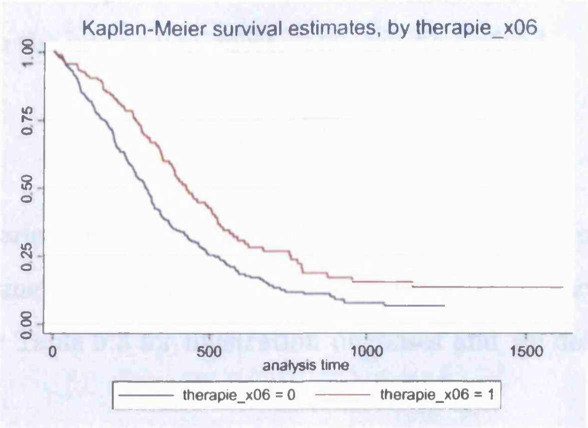


Figure 9-6: Kaplan-Meier survival estimates for interaction between treatment and Karnofsky index (type 1) (therapie_x06) in the Glioma2 study

covariate with treatment alone. Furthermore, the graphical analysis using the Kaplan-Meier survival graphs gives graphical representation of the magnitude of the effects but is not entirely satisfactory. Hence we developed a Stata program to provide more detail for the analysis of treatment-covariate interactions. This is explained in detail in the next section.

Study	**Covariate**	**Treatment effect overall**	**Interaction effect with covariate**
AXIS	hetany2	0.73 (0.54 - 0.98)	0.74 (0.55 - 0.97)
Glioma2	Karnofsky index (type 1)	0.89 (0.71 - 1.14)	0.64 (0.49 - 0.82)

Table 9.2: Summary of main effects together with a 95% confidence interval

9.4 Model and computation

A Stata 8 program and dialog were written to aid the visualisation of treatment-covariate interaction effects in clinical trials. The program produces tabular output of the interaction effects as well as graphics. This and the following two sections describe first the mathematical background for the calculations and then the program set-up. Two trial examples are given at the end.

The model underlying the calculations is based on a 2*2 table for interactions as illustrated in Table 9.3. For the Cox proportional hazard model we can see that the hazard ratio between treatment = 1 and treatment = 0, while the covariate is equal to 0, is λ. Similarly, we arrive at a hazard ratio of υ between the covariate being equal to 1 and 0, whilst treatment is equal to 0. We then define the ratio of hazard ratios (RHR) as τ which illustrates the interaction effect. This can be derived

$$RHR = \frac{\left(\frac{\lambda v \tau}{v}\right)}{\left(\frac{\lambda}{1}\right)} = \tau \qquad (9.1)$$

A similar definition arises when looking at the logistic regression model since the parameters remain the same but we are dealing with odds ratios instead of hazard ratios. So again we can employ Table 9.3 for illustration purposes and we define the ratio of odds ratios (ROR) as τ.

	Treatment = 0	**Treatment = 1**
Covariate=0	1	λ
Covariate=1	v	$\lambda v \tau$

Table 9.3: 2*2 table for interaction effects

The table and graphics output by the program based on the Cox proportional hazards model are computed using the Cox model as implemented in Stata. Let A denote the treatment and Z a covariate of interest. The overall hazard is calculated using

$$h(t|A) = h_0(t)\exp(\alpha_1 A) \qquad (9.2)$$

where α_1 is defined as the coefficient for the treatment variable, while the hazards in the two groups as well as the hazard for the RHR are based on the model

$$h(t|A,Z) = h_0(t)\exp(\beta_1 A + \beta_2 Z + \beta_{12} AZ) \qquad (9.3)$$

We can estimate λ by β_1 and v by β_2. The interaction term is given by β_{12}.

The logistic option employs logistic regression, again as implemented in Stata. The overall treatment odds ratio is estimated using

$$\pi(A) = \frac{\exp(\alpha_0 + \alpha_1 A)}{1 + \exp(\alpha_0 + \alpha_1 A)} \qquad (9.4)$$

The odds ratio in both levels of the covariate and the ROR are based on the following model

$$\pi(A,Z) = \frac{\exp(g(A,Z))}{1 + \exp(g(A,Z))} \qquad (9.5)$$

for

$$g(A,Z) = \beta_0 + \beta_1 A + \beta_2 Z + \beta_{12} AZ \qquad (9.6)$$

where β_0 is the coefficient on the constant term, β_i, $i = 1, 2$, are the coefficients on the independent variables and β_{12} denotes the coefficient for the interaction term.

The graphical output of this program is based on forest plots. A "forest plot" is a pictorial presentation of the hazard or odds ratio with corresponding confidence intervals. A more detailed description of forest plots and their history may be found in Lewis & Clarke [75].

9.5 Design of the dialog

The program may be invoked using the `Fintplot` menu and its associated dialog. On completion of the calculations Stata displays a table of output containing the overall treatment hazard ratio, the hazard ratio in both groups of the prognostic factor chosen and an estimate of the RHR or ROR for interaction. Furthermore, a forest plot is displayed using Stata 8 graphics. The program has an `Overview` dialog option which provides a forest plot of the overall treatment hazard or odds ratio and RHRs or RORs for up to five covariates with treatment. Calculations are performed in the ado files `fintplot` and `fintplotk`. The default method of analysis is the Cox proportional hazards model.

`fintmenu` can be executed by typing `fintmenu on` and a new item `Fintplot` will appear on the system menu-bar under `User`. This menu can be turned off again by typing `fintmenu off`.

9.5.1 Forest plot and table for interaction

The following description will concentrate on the `Fintplot - Detail` dialog, however, the `Fintplot - Overview` dialog may be used in a similar manner. The dataset employed in the analysis needs to have been `stset` prior to using this menu if the Cox proportional hazards model is to be used and the covariate levels need to be binary. The `User` may decide on sensible binary levels for the covariates which are of further interest by first employing the `Overview` dialog.

The `Fintplot - Detail` dialog allows both the `by` and `if` options to be executed separately or at the same time. Variables used for the `by` option of the program need to be discrete and can be entered under `Separate by observations`. If the Cox proportional hazards model is chosen, the program also allows for stratification. The variable to be used for stratification needs to be entered under `Stratify by observations` which is also located in the by option part of the dialog window. In the case of the if option, the Create button allows the easier construction of the logical argument. In addition, the confidence level may be set prior to running the program in the usual way using set level #. Lastly, if the log scale is preferred for the forest plot, one needs to tick the box for Log scale in the main dialog window. The table will remain unchanged by this option.

9.6 Examples

The examples given below illustrate the program by employing data from the Glioma and Low Infant Birth Weight studies. Caution needs to be observed in looking at the results as these were not tests for interactions predefined in the protocol.

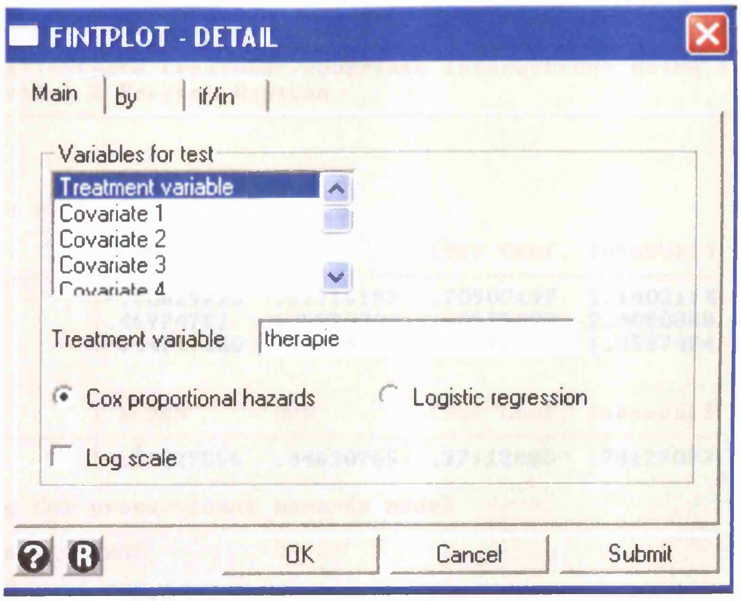
9.6.1 Forest plot for an interaction of two different covariates with treatment

The first example was run using the Glioma2 study described above. Further information on this study is available in an article by Ulm et al. [139].

The data was stset prior to running the main analysis. For this first run, we have decided to look at the possible interaction between treatment (Trt) and two different binary categorisations of the Karnofsky-index (x06 and x07). Figure 9-7 illustrates how we enter the information into the dialog window. The treatment variable should always be entered first. Upon pressing OK or Submit we obtain the output given in Figure 9-8.

The log hazard ratios and hazard ratios in both levels of the factor and the overall hazard ratio are given as well as confidence intervals. This output is split into both categorisations of the Karnofsky index (x06 and x07). Most importantly the second table for each categorisation gives the log RHR and RHR for the interaction between treatment and the Karnofsky index.

Figure 9-9 illustrates the forest plot output by the program for these interactions. Here the diamond shape gives the overall hazard ratio for treatment without differentiating by factor. The square shapes then display an estimate of the hazard ratios in the two groups. Lastly, the circle shape gives us the RHR for the interaction. When looking at the plot of therapie and x06 we can see that the confidence interval for the first level of x06 is



Variables for test

Treatment variable
Covariate 1
Covariate 2
Covariate 3
Covariate 4

Treatment variable: therapie

● Cox proportional hazards ○ Logistic regression

□ Log scale

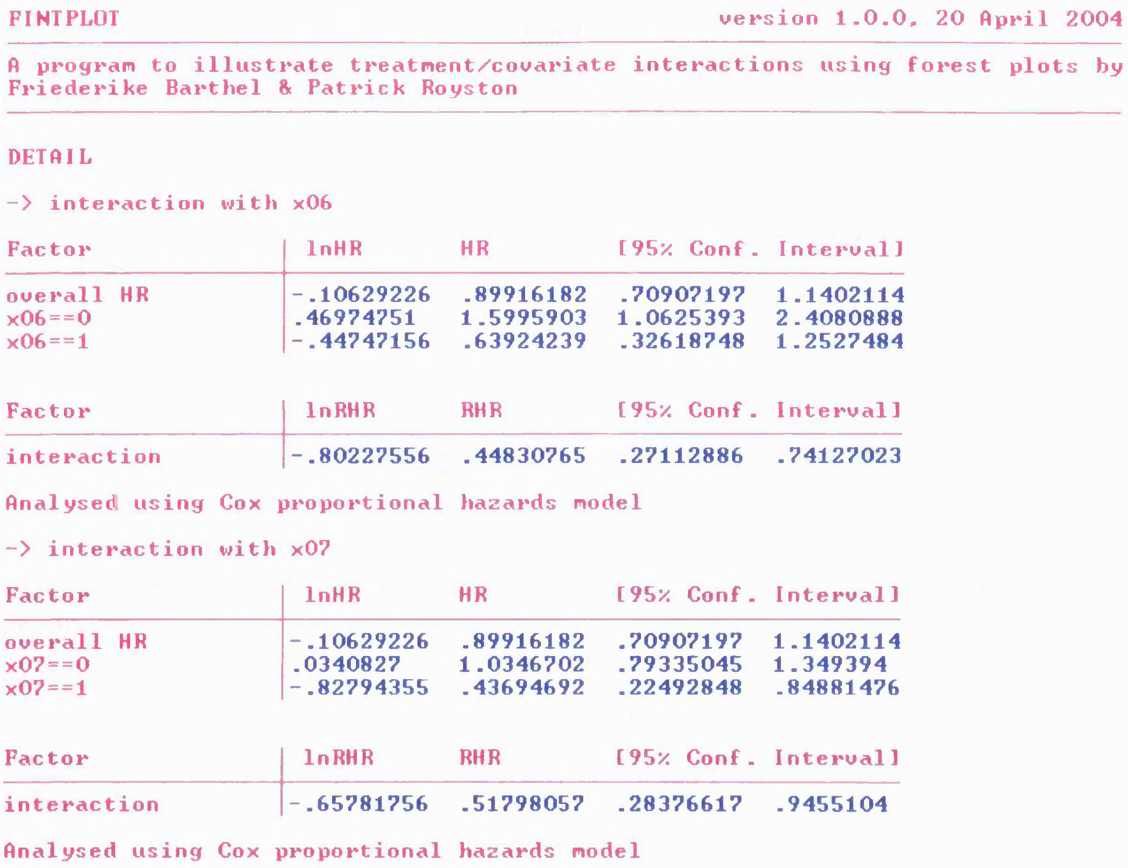
Figure 9-7: Dialog window illustrating analysis of two interactions under Cox model

too wide for the table. It has hence been truncated at a value of 2.5. Both the tables and forest plots show that there is evidence of an interaction between treatment and the Karnofsky index with an RHR of 0.45 or 0.52 depending on the specification.

9.6.2 Forest plot of an interaction of one covariate with treatment using both by and log scale options

The data used in this example originates from a study of the Risk Factors Associated with Low Infant Birth Weight. Data collection took place at Baystate Medical Center in Springfield, Massachusetts during 1986. Information was gathered on the birth weight in grams (bwt), the age of the mother (age), the mothers weight in pounds at the last menstrual period (lwt), race (race), smoking status during pregnancy (smoke), history of premature labour (ptl), history of hypertension (ht), presence of uterine irritability (ui) and the number of physician visits during the first trimester (ftv). Birth weight in grams was the further split into a low birth weight (low) categorisation whereby 1=birth weight < 2500 g. Further information on the analysis of this dataset is given in Hosmer & Lemeshow [60]. At the planning stage an interaction analysis was specified.

Hosmer & Lemeshow suggest splitting lwt into two categories (lwd) whereby 1 denotes a weight of under 110 pounds. Furthermore they have investigated a possible interaction



```
FINTPLOT                                          version 1.0.0, 20 April 2004
------------------------------------------------------------------------------
A program to illustrate treatment/covariate interactions using forest plots by
Friederike Barthel & Patrick Royston
------------------------------------------------------------------------------

DETAIL

-> interaction with x06

Factor               |  lnHR        HR          [95% Conf.  Interval]
---------------------+-------------------------------------------------
overall HR           | -.10629226  .89916182  .70907197  1.1402114
x06==0               | .46974751   1.5995903  1.0625393  2.4080888
x06==1               | -.44747156  .63924239  .32618748  1.2527484

Factor               |  lnRHR       RHR         [95% Conf.  Interval]
---------------------+-------------------------------------------------
interaction          | -.80227556  .44830765  .27112886   .74127023

Analysed using Cox proportional hazards model

-> interaction with x07

Factor               |  lnHR        HR          [95% Conf.  Interval]
---------------------+-------------------------------------------------
overall HR           | -.10629226  .89916182  .70907197  1.1402114
x07==0               | .0340827    1.0346702  .79335045  1.349394
x07==1               | -.82794355  .43694692  .22492848  .84881476

Factor               |  lnRHR       RHR         [95% Conf.  Interval]
---------------------+-------------------------------------------------
interaction          | -.65781756  .51798057  .28376617  .9455104

Analysed using Cox proportional hazards model
```

Figure 9-8: Fintplot table output for Glioma study

between smoke and lwd split by age. Hence we have decided to create a new variable age5 which takes on the value 2 for age>25 and 1 otherwise. We will be using logistic regression in this example. The dialog window is invoked as before; however, we now need to enter an outcome variable for the events, which is low in this dataset. Furthermore we tick the box for **Logistic** and **Log scale**. To split the data by age5 we need to switch to the **by** option menu and enter age5 as a variable under **Separate by observations**. Figures 9-10 and 9-11 illustrate this. Once we press the **OK** or **Submit** buttons we obtain the output given in Figure 9-12.

This output can be read as in the first example, however, in this case we have a split by AGE5. The forest plot is illustrated in Figure 9-13. All symbols have the same meaning as defined in Section 9.6.1. We can hence illustrate the potential influence of other variables. The output from both the table and the forestplot suggest no evidence of an interaction between smoking and weight at the last menstrual period when we separate

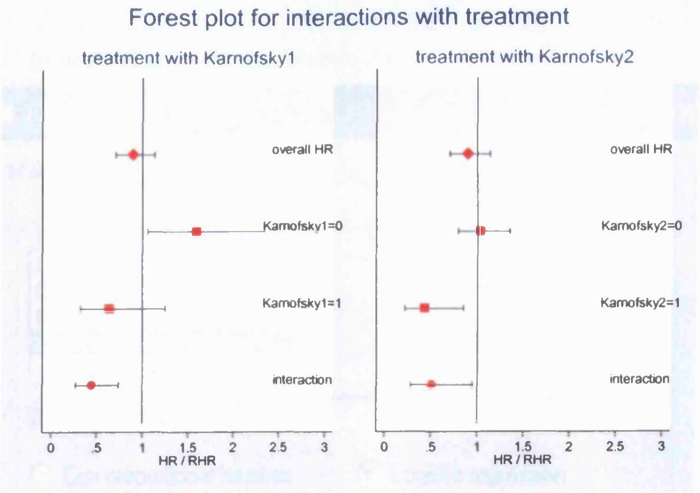
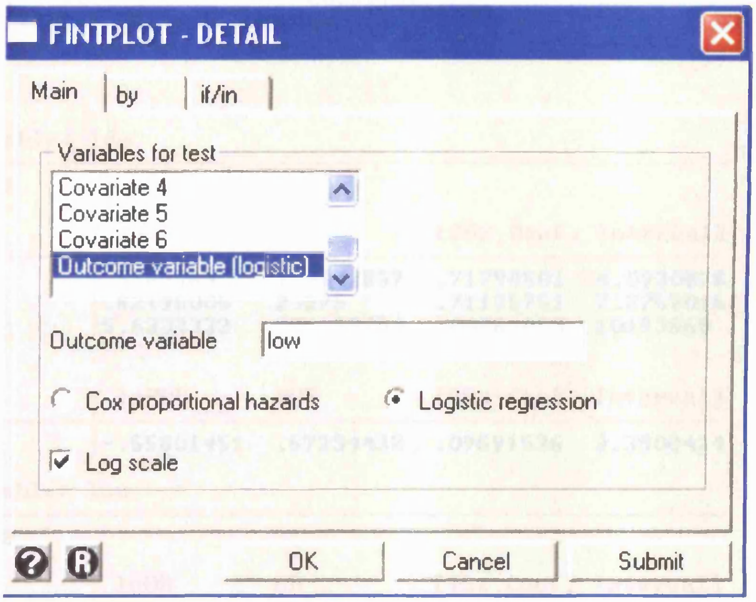


Figure 9-9: Forest plot for interaction of treatment with two categorisations of the Karnofsky index

the data by age5. However, the analysis is not very conclusive due to wide confidence intervals which stem from the fact that there is only a small amount of data available in each group.

9.7 Conclusions

It is becoming increasingly important to analyse the effect an intervention has across different levels of a covariate in order to allow for more individual patient care. Hence we have developed a Stata tool to express such interactions both quantitatively and visually within a 2*2 table framework. It is flexible in the options it provides and operates under either the Cox proportional hazards or the logistic regression model.



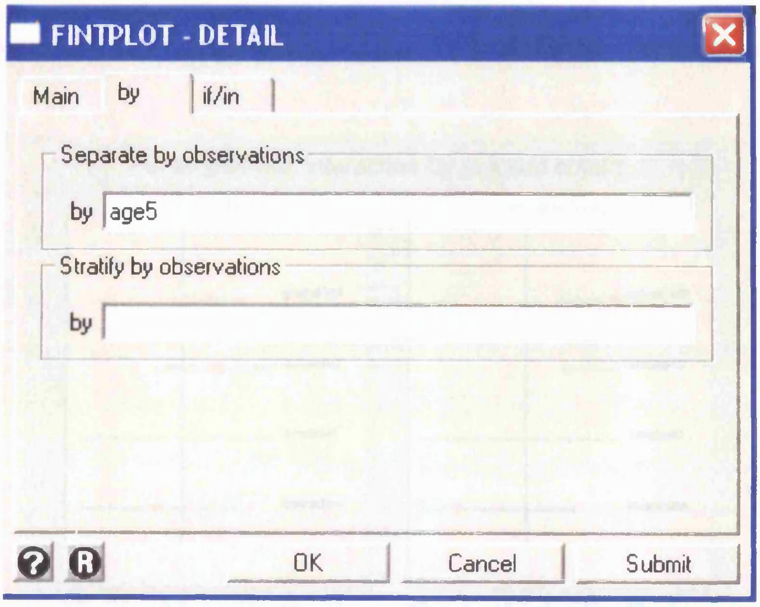
Variables for test
Covariate 4
Covariate 5
Covariate 6
Outcome variable (logistic)

Outcome variable: low

○ Cox proportional hazards ● Logistic regression

☑ Log scale

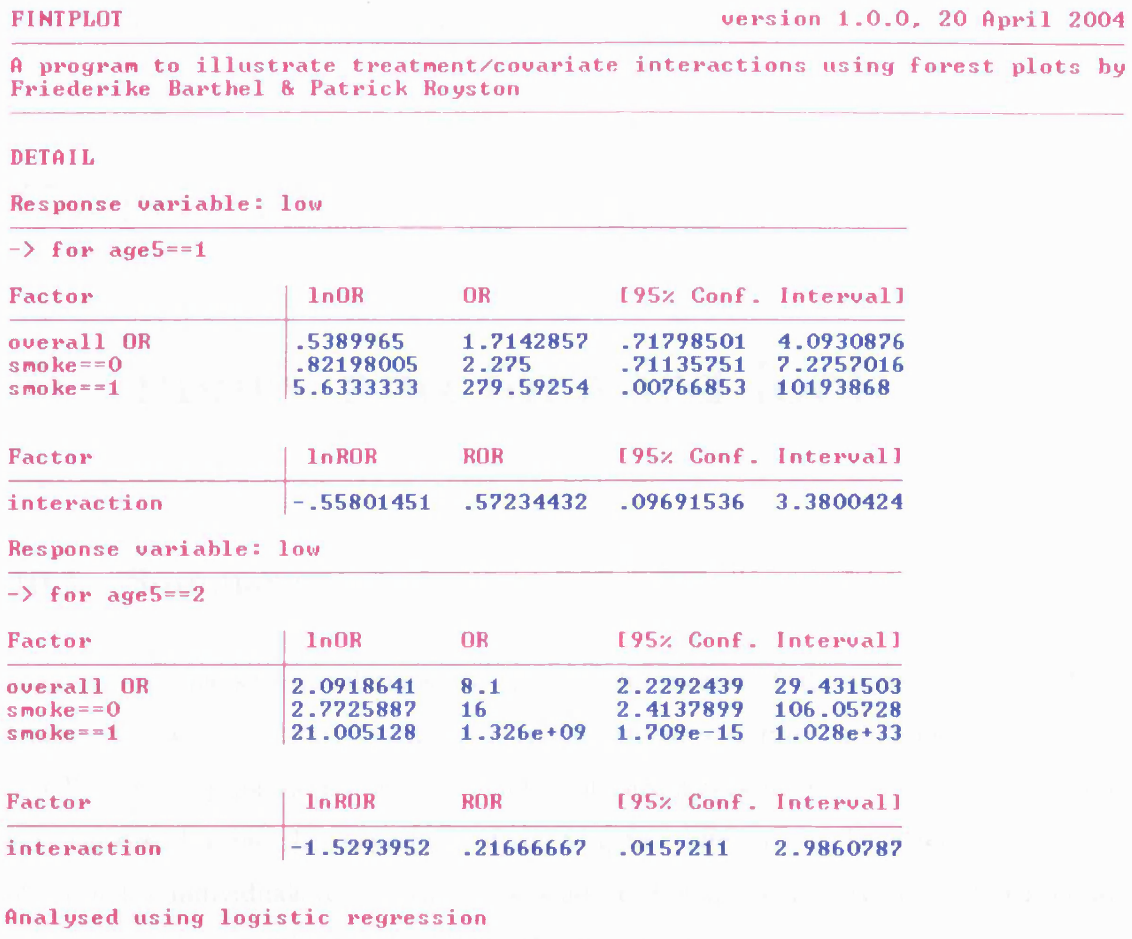
Figure 9-10: Dialog window illustrating input of outcome variable for logistic regression



Separate by observations
by: age5

Stratify by observations
by:

Figure 9-11: Dialog window illustrating use of by option



```
FINTPLOT                                        version 1.0.0, 20 April 2004
----------------------------------------------------------------------------
A program to illustrate treatment/covariate interactions using forest plots by
Friederike Barthel & Patrick Royston
----------------------------------------------------------------------------

DETAIL

Response variable: low
-------------------------------------------------------------
-> for age5==1

Factor               | lnOR         OR          [95% Conf.  Interval]
---------------------+-----------------------------------------------
overall OR           |.5389965     1.7142857   .71798501   4.0930876
smoke==0             |.82198005    2.275       .71135751   7.2757016
smoke==1             |5.6333333    279.59254   .00766853   10193868

Factor               | lnROR        ROR         [95% Conf.  Interval]
---------------------+-----------------------------------------------
interaction          |-.55801451   .57234432   .09691536   3.3800424

Response variable: low
-------------------------------------------------------------
-> for age5==2

Factor               | lnOR         OR          [95% Conf.  Interval]
---------------------+-----------------------------------------------
overall OR           |2.0918641    8.1         2.2292439   29.431503
smoke==0             |2.7725887    16          2.4137899   106.05728
smoke==1             |21.005128    1.326e+09   1.709e-15   1.028e+33

Factor               | lnROR        ROR         [95% Conf.  Interval]
---------------------+-----------------------------------------------
interaction          |-1.5293952   .21666667   .0157211    2.9860787

Analysed using logistic regression
```

Figure 9-12: Fintplot table output for Low Birth Weight study

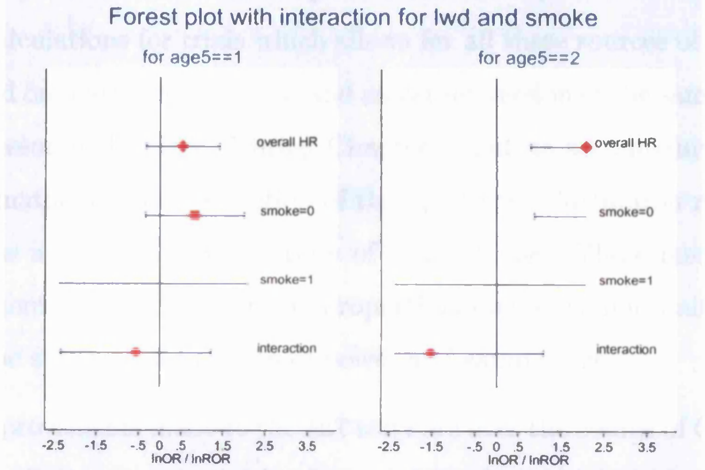


Figure 9-13: Forest plot using logistic regression, log scale and by options for Low Infant Birth weight data set

Chapter 10

Summary and forward look

10.1 Summary

Adequate sample size calculations are vital for the success of all randomised controlled trials. They are particularly complex for trials with survival-type endpoints because they usually involve prior estimates of a number of parameters including the control group survival distribution, the magnitude of the targeted difference to be detected, the rate of accrual of individuals to the study, the length of follow-up of individuals after accrual closure and the potential for (time-related) dilution of any effect through, for example, loss to follow-up or change of treatment. All of these parameters can have an important impact on the trial size needed.

In the first part of the thesis in Chapters 3 and 4 we presented a general approach to sample size calculations for trials which allows for all these sources of variability. This approach is based on mathematical ideas and an earlier version of the sample size program derived by Professor A. Babiker. During Chapter 3 and its accompanying paper [7] we formulated the mathematical description of the approach. Simulation results show that these calculations are accurate in a variety of trial settings. These results also indicate that the adjustments particularly for non-proportional hazards, non-uniform accrual and cross-over may be substantial in terms of power and sample size.

The main improvements made to the ART software over the course of Chapter 4 are the new design of the dialog menu exploiting features introduced in Stata 8 and more detailed output. In addition, the sample size calculations may now be performed for non-inferiority designs. Thus users should find the new version easier to use and more informative than the first release. Furthermore, a comparison with other software packages has shown that this is the only widely available program to take into account all of the above mentioned complexities.

In trials which are aimed at comparing treatments to treat diseases that are serious and life-threatening such as cancer or HIV, surrogate endpoints are attractive since they can be measured sooner and more easily than those which are considered the most valuable clinical endpoints in such diseases: morbidity and mortality. We have presented statistical methodology from the literature in Chapter 5 that aims to assess the strength of surrogate markers for clinical endpoints, both within individual clinical studies and across clinical studies. To be useful to investigators, surrogate endpoints should also result in a reduction of either sample size or the duration of the study. The acceptance of surrogate endpoints in clinical trials as the basis for drug approvals is recognised as carrying risks. These include the risk that a treatment-induced effect on a surrogate endpoint will not correlate with a clinical effect, resulting in an ineffective product proceeding to market if the analysis at the end of the trial was based only on the surrogate endpoint. Therefore we propose to use a primary outcome such as mortality for the final analysis and an intermediate outcome for the analysis at intermediate stages.

The multi-stage, multi-arm methodology presented in the second part of the thesis, specifically in Chapters 6 and 7, aims to address the pressing need to speed the process of the evaluation of new therapies, particularly in cancer. This approach has two distinguishing characteristics: many new therapies are compared at once against a control treatment and ineffective therapies are rejected on the basis of an intermediate outcome measure, by a randomised comparison of each new arm against the control. This intermediate outcome measure is not required to be a perfect surrogate for the final outcome in the Prentice sense [99] but rather it is essential that the effect sizes of the new treatment on the intermediate and final outcome measures are related. In general, the main advantage of this approach is the ability to reject one or all of the experimental treatments early. This means are that fewer patients need to be recruited, the trial takes less time to run, there is increased flexibility in the design and costs are reduced.

In Chapter 6 the multi-arm, multi-stage design first introduced by Royston et al. [103] was extended to more than two stages. This included the development of Stata software for sample size calculations for this type of design. The underlying assumptions of the design were examined and further improvements to the methodology were suggested, particularly in the area of non-exponential survival.

Chapter 7 provided a review of literature around bivariate exponential distributions as well as an addition to the methodology in the form of the $NBVE$ (Normal Bivariate Exponential) distribution. This was necessary in order to facilitate the simulation of trials with a progression free intermediate outcome and overall survival outcome. Simulation studies illustrate the accuracy and robustness of the sample size calculations for the multi-stage, multi-arm design in this chapter.

In the third part of the thesis in Chapter 8 we explored strategies to analyse the inherent variability in trial time and / or number of events and provided a Stata tool to assess the variability at the beginning of the trial as well as update these estimates throughout patient and event accrual. We have shown that it may be beneficial to take into account the variability in trial duration at the planning stage since it may have significant impact on the total cost and practicality of the trial. This knowledge may be of particular importance in multi-stage trials where the stages itself are often relatively short and have a smaller sample size. It is necessary to realise that this degree of variability already arises even if we have estimated all other parameters correctly when calculating sample sizes.

It is becoming increasingly important to analyse potential treatment-covariate interactions in order to allow for more targeted patient care. Thus it is of great interest to observe whether the treatment effect is consistent across some demographic factors such as age, gender, baseline disease severity, some prognostic factors, or previous medical conditions and concomitant medications. Hence we have developed a Stata program with which such interactions can be expressed and displayed both quantitatively and visually within a 2*2 table framework. It is flexible in the options it provides and operates under either the Cox proportional hazards or logistic regression model. This programme as well as the underlying methods were described in the fourth part of the thesis in Chapter 9.

10.2 Extensions

10.2.1 Chapters 5 to 7

We have identified a number of areas in which the multi-stage, multi-arm designs may be extended. These include the following:

1. Assessing potential gains from earlier trials: A number of conventionally designed recent trials have collected information both on an intermediate outcome and a final outcome measure. We propose to reanalyse these studies to assess whether using the multi-stage methodology we could have identified and 'rejected' ineffective therapies, without inappropriately rejecting effective therapies. Hence we would like to assess whether the use of emerging data on the intermediate outcome would have allowed us to stop early in trials in which little or no effect on overall survival was observed. Similarly, we wish to ascertain whether in studies with a positive outcome on overall survival the trial would have been stopped early inappropriately if data on the intermediate outcome measure had been used. Specifically, we could assess whether employing our methodology in these studies would have reduced the number of patients needed and saved time.

2. Extension to other outcomes: Hitherto, the methodology has been developed for two correlated survival-type outcome measures. We propose to extend the methodology to cases in which the intermediate outcome is a binary or ordered categorical endpoint, such as tumour response.

3. Assessing correlation among treatment effects: The calculation of overall Type I and II errors in multi-stage, multi-arm trials depends upon adequate specification of the correlation of treatment effects on the intermediate and primary outcomes at different time-points over the course of the trial. While some progress has been made in this area using studies on ovarian cancer with two survival-type outcomes, other cancer types and other outcome measures require further work.

4. Operating characteristics: So far work on the operating characteristics as outlined in Chapter 6 has concentrated on the type I and II errors for comparing one experimental arm with a control only. However, in order to determine the overall type I and II errors for the design two main correlations need to be taken into account: a) two or more analyses are conducted over time and b) two or more simultaneous comparisons with control. Some preliminary work has been carried out in this area, using the multivariate normal distribution of the treatment effects, which addresses the first issue of these issues.

5. Bias in treatment-effect estimates: We need to consider whether ceasing further randomisations to a research arm in a multi-stage, multi-arm trial and continuing allocation to other arms may introduce bias in estimated treatment effects. The

original publication [103] states that such bias is avoided by reporting treatment effects for all treatments, irrespective of whether they were dropped early or not. We will investigate this issue further from a methodological perspective and by computer simulation.

6. Practical issues in multi-stage, multi-arm trials: The first trial (ICON5, in ovarian cancer) employing the two-stage design as outlined by Royston et al. [103] has just been completed at the MRC Clinical Trials Unit. A further trial (STAMPEDE, in prostate cancer) has been designed with an extension to more than two stages, and is under way. To assess the practical impact of the approach and provide guidance on undertaking these trials, we propose to examine the issues raised in these multi-arm, multi-stage trials. These include:

    (a) Patient information: How much information needs to be given in the patient information sheet? What information should be given to the patients allocated to an arm which is stopped early?

    (b) Protocol / Statistical Analysis Plan: We propose to write a plan on 'How to describe a multi-stage multi-arm trial?' in a protocol or trial proposal. Furthermore, guidance on the statistical analysis plan is needed.

    (c) End of Stage 1: There is currently a grey area between the end of Stage 1 in terms of the number of events needed and the time of the actual analysis. A similar discrepancy in times occurs at the end of the trial. Currently, recruitment is stopped when the needed number of events have been accrued. This may lead to overpowered trials. Hence the question arises of whether there is an optimal time of stopping recruitment?

10.2.2 Chapter 8

We furthermore propose that the work described in Chapter 8 on the variability in total trial time be extended to incorporate those trial design options provided in ART as outlined in Chapter 4. The second part of the analysis of variability program currently utilises trial data from the control group only. However, due to issues of unblinding trial data before the final analysis, only the data from all groups as a whole may be available. Thus it may be beneficial to allow for the input of the trial dataset as a whole in order to obtain an overall median survival. By assuming the hazard ratio used for sample size calculations at the outset we may then calculate median survival in the control group only and employ the tool as described above. A third area of extensions to this methodology may be the use of spline functions as proposed by Royston et al. [102] to predict the rest of the distribution from trial data available so far and subsequently calculate median survival based on that information.

Appendix A

Publications and prizes

The research in this thesis won the Fisher Memorial Trust bursary to attend and present at the International Biometric Conference in Cairns in 2004. A poster based on research in this thesis was awarded the Best PhD Poster Prize at the GSK BDS UK Statisticians' and Programmers' Conference, 2004. Furthermore, I was awarded the University College London Momber Scholarship, 2003/2004, and the Costas Goutis Prize, 2004. Below is a list of publications and conference presentations that have resulted from the work described in this thesis.

A.1 Papers

1. F. M.-S. Barthel, P. Royston, A. Babiker. 2005. *A menu-driven facility for complex sample size calculation in randomized controlled trials with a survival or a binary outcome: update.* The Stata Journal: 5, 123-129.
2. F. M.-S. Barthel, A. Babiker, P. Royston, M. K. B. Parmar. 2006. *Evaluation of sample size and power for multi-arm survival trials allowing for non-uniform accrual, non-proportional hazards, loss to follow-up and cross-over.* Statistics in Medicine: accepted.
3. F. M.-S. Barthel, P. Royston. *Graphical representation of interactions.* The Stata Journal: submitted.
4. F. M.-S. Barthel, P. Royston. *multinorm: Multivariate normal probabilities.* to be submitted to The Stata Journal.

A.2 Presentations

1. F. M.-S. Barthel, P. Royston, M. K. B. Parmar. 2005. *Sample sizes for time-to-event outcomes: implications of the variability in events and time.* 26th Anniversary Meeting of the SCT. Portland. Clinical Trials: 2 (Supp. 1), S31.

2. F. M.-S Barthel, P. Royston, M. K. B. Parmar. 2005. *Designs for multi-stage multi-arm clinical trials with survival outcomes – assessing robustness and practicality.* MANDEC Seminar. Manchester. invited.

3. F. M.-S. Barthel, P. Royston, M. K. B. Parmar. 2004. *Designs for two-stage multi-arm clinical trials with survival outcomes - assessing robustness and practicality.* GSK BDS UK Statistician's and Programmers' Conference 2004. Ware.

4. F. M.-S. Barthel, A. Babiker, P. Royston, M. K. B. Parmar. 2004. *Evaluation of sample size and power for multi-arm survival trials allowing for non-proportional hazards, loss to follow-up and cross-over.* ISCB 2004. Leiden. Abstract book: 87.

5. F. M.-S. Barthel, P. Royston, M. K. B. Parmar. 2004. *Designs for multi-arm clinical trials with survival outcomes - assessing robustness and practicality.* ISCB 2004. Leiden. Abstract book: 141.

6. F. M.-S. Barthel, P. Royston, M. K. B. Parmar. 2004. *Sample sizes for time-to-event outcomes: implications of the variability in events and time.* IBC 2004. Cairns. Proceedings of the XXIInd International Biometric Conference.

7. F. M.-S. Barthel, P. Royston, M. K. B. Parmar. 2004. *Designs for multi-arm clinical trials with survival outcomes - assessing robustness and practicality.* 6. Workshop Adaptiv-sequentielle Verfahren. Mainz. Abstractheft.

8. F. M.-S. Barthel. 2004. *Simulation results for two-stage multi-arm trials. Workshop on the analysis of clinical trials incorporating treatment selection.* Reading. invited.

9. F. M.-S. Barthel, A. Babiker, P. Royston, M. K. B. Parmar. 2004. *Evaluation of sample size and power for multi-arm survival trials allowing for non-proportional hazards, loss to follow-up and cross-over.* Karlsruher Stochastik-Tage 2004. 6th German Open Conference on Probability and Statistics. Abstracts and list of participants: 153 - 154.

Appendix B

Derivation of the non-centrality parameter τ

This appendix provides further details on the calculation of the sample size N under loss to follow-up, non-proportional hazards and cross-over as referred to in Section 3.2. All variables are defined as described in Sections 3.2 and 3.3. This work is based on fundamentals derived by Professor A. Babiker.

Let the observed numbers of events $(O_2^j, ..., O_K^j)'$ have a multinomial distribution with probability $[e_2(t_j;\Delta), \; ... \; , e_k(t_j;\Delta)]'$. Define

$$M_k(\Delta) = \frac{1}{\sqrt{N}} E(U_k|H_1)$$

where the expectation of the logrank statistic U_k under the alternative hypothesis is given by

$$E(U_k|H_1) = \sum_{j=1}^{m} W(t_j)[e_k(t_j;\Delta) - e_k(t_j;0)]$$

and let $M(\Delta) = (M_2(\Delta), ..., M_K(\Delta))'$. The covariance of U is structured as a $(K-1) \times (K-1)$ matrix $V(\Delta) = (v_{kl})$ where

$$v_{kl}(\Delta) = \sum_{j=1}^{m} [W(t_j)]^2 e_k(t_j;\Delta)[\delta_{kl} - e_l(t_j;\Delta)] \tag{B.1}$$

for $k, l = 2, ..., K$; and $\delta_{kl} = 1$ if $k = l, 0$ otherwise. According to the Central Limit theorem [51], as $N \longrightarrow \infty$, $((U/\sqrt{N}) - M(\Delta))$ is asymptotically distributed as multivariate normal $\mathbf{N}(0, V(\Delta)/N)$, i.e. for large N, $U/\sqrt{N}$ is approximately distributed as

$\mathbf{N}(M(\Delta), V(\Delta))$. It follows that

$$\overline{Q} = U'(V(\Delta))^{-1}U$$

is distributed as non-central χ^2_{K-1} with non-centrality parameter

$$\tau = M(\Delta)'(\frac{V(\Delta)}{N})^{-1}M(\Delta) = NM(\Delta)'(V(\Delta))^{-1}M(\Delta) \tag{B.2}$$

and

$$Q = U'(V(0))^{-1}U$$

is distributed as central χ^2_{K-1} under H_0 [118]. Under local alternatives H_1, we can replace $V(\Delta)$ by $V(0)$ in the expressions for $\overline{Q}$ and τ and so the logrank statistic Q is approximately non-central χ^2 with the non-centrality parameter given by Equation 3.1 in Section 3.2.

In order to calculate the sample size N we need to find $M(\Delta)$ and $V(0)$ asymptotically as $N \rightarrow \infty$. To do this we incorporate our knowledge about patient accrual, loss to follow-up and cross-over into τ. Let $F^R(t)$, $S^E_k(t)$ and $S^L_k(t)$ follow the notation in Section 3.3.1. Then the probability that a randomly selected patient is in group k and is still at risk of failure at time t is $F^R(T-t)H_k(t)$, where

$$H_k(t) = p_k S^L_k(t) S^E_k(t) \tag{B.3}$$

The limit of $e_k(t;\Delta)$ as $N \rightarrow \infty$ is given by

$$\rho_k(t;\Delta) = \frac{H_k(t)\Delta_k(t)}{\sum_{l=1}^{K} H_l(t)\Delta_l(t)} \tag{B.4}$$

Let

$$\psi^E_k(t) = F^R(T-t)p_k S^L_k(t) f^E_k(t) \tag{B.5}$$

and

$$\psi^E_{\cdot}(t) = \sum_{l=1}^{K} \psi^E_l(t) \tag{B.6}$$

If we let $g(t) = \lim_{N\to\infty} W(t)$ then

$$M(\Delta) \to \lim_{N\to\infty} \frac{1}{\sqrt{N}} \int_0^\infty g(t)[\rho_k(t;\Delta) - \rho_k(t;0)]\psi_{\cdot}^{E}(t)dt \tag{B.7}$$

as $N \to \infty$. Furthermore

$$\frac{V(\Delta)}{N} \to \lim_{N\to\infty} \frac{1}{N} \int_0^\infty [g(t)]^2 \rho_k(t;\Delta)[\delta_{kl} - \rho_l(t;\Delta])\psi_{\cdot}^{E}(t)dt \tag{B.8}$$

Under the unweighted logrank test the weights are given by

$$g(t) = 1 \tag{B.9}$$

while Tarone & Ware weighting [130] has the form

$$g(t) = \{F^R(T-t)[\sum_{k=1}^{K} p_k S_k^L(t) S_k^E(t)]\}^{1/2} \tag{B.10}$$

and Harrington & Fleming weighting [56] can be calculated as

$$g(t) = \left\{ \frac{\sum_{k=1}^{K} p_k S_k^L(t) S_k^E(t)}{\sum_{k=1}^{K} p_k S_k^L(t)} \right\}^p \tag{B.11}$$

Appendix C

An approximation to the distribution of the logrank test statistic Q under more distant alternatives

This work is based on fundamentals derived by Professor A. Babiker. In Appendix B, the distribution of the logrank statistic, under local alternatives H_1, was approximated by that of $\overline{Q}$ replacing $V(\Delta)$ by $V(0)$ leading to a non-central χ^2 with $K-1$ degrees of freedom and non-centrality parameter

$$\tau = NM(\Delta)'V(0)^{-1}M(\Delta)$$

More generally, the distribution of Q under H_1 can be approximated by one of two methods:

The first is based on approximating the distribution of Q under H_1 by that of a constant multiple of non-central $\chi^2_{K-1}(\tau)$ using the first two moments [118]

$$E(Q|H_1) = tr(V(0)^{-1}V(\Delta)) + NM(\Delta)'V(0)^{-1}M(\Delta) \qquad \text{(C.1)}$$

and

$$V(Q|H_1) = 2\{tr(V(0)^{-1}V(\Delta))^2 + 2NM(\Delta)'V(0)^{-1}M(\Delta)V(0)^{-1}M(\Delta)\} \qquad \text{(C.2)}$$

to solve for the multiplying factor and the non-centrality parameter τ. This means that we approximate the distribution of Q by that of cX where $X \sim \chi^2_{K-1}(\tau)$. Setting $L = K - 1$ and equating the right hand side of equations C.1 and C.2 with $cE(X)$ and $c^2Var(X)$ respectively, where $X \sim \chi^2_{K-1}(\tau)$ [33], we obtain

$$tr(V(0)^{-1}V(\Delta)) + NM(\Delta)'V(0)^{-1}M(\Delta) = c(L + \tau)$$

and

$$tr(V(0)^{-1}V(\Delta))^2 + 2NM(\Delta)'V(0)^{-1}M(\Delta)V(0)^{-1}M(\Delta) = c^2(L + 2\tau)$$

Hence the non-centrality parameter of Q is then given by

$$\tau = \frac{[(b_0^2 - Lb_1) + \sqrt{(b_0^2 - Lb_1)^2 + b_1L(b_0^2 - Lb_1)}}{b_1}$$

where we define $a_0 = tr(V(0)^{-1}V(\Delta))$, $q_0 = M(\Delta)'V(0)^{-1}M(\Delta)$, $a_1 = tr(V(0)^{-1}V(\Delta))^2$,

$$q_1 = M(\Delta)'V(0)^{-1}M(\Delta)V(0)^{-1}M(\Delta)$$

$b_0 = a_0 + Nq_0$ and $b_1 = a_1 + 2Nq_1$. Using N as calculated under local alternatives as a starting value we can then find N under distant alternatives iteratively such that it satisfies the following equation

$$1 - \beta = nchi(K - 1, \tau, [invchi(K - 1, \alpha)]\frac{(K - 1 + \tau)}{b_0})$$

where $nchi(L, \tau, z)$ is the value of the cumulative distribution function of a non-central χ^2 with L degrees of freedom and non-centrality parameter τ, at z. $invchi(L, \alpha)$ gives the inverse of the cumulative distribution function of the central χ^2 distribution with L degrees of freedom at α.

The second method is based on sampling from Q under H_1 by using the knowledge that $U/\sqrt{N}$ is asymptotically multivariate normal with mean $M(\Delta)$ and covariance matrix $V(\Delta)$. Under this method we obtain 10,000 replications of the vector U. For each of these Q is calculated which gives us the empirical distribution for the quadratic form under H_1. The power is then calculated by counting how many times Q is greater than $invchi(K - 1, \alpha)$. This second method is implemented in the sample size program ART.

Appendix D

Overall power and significance level using the multivariate normal distribution

As outlined in Chapter 6 the overall power and significance level of a multi-stage trial with s stages will follow a multivariate normal distribution Φ_s. In order to arrive at the required sample size and other design characteristics for the s-stage trial the sample size calculations described in that chapter were programmed in Stata 8. However, Stata 8 currently does not provide for a multivariate normal with $s > 2$. Hence we had to provide such a program. For this purpose, the literature was surveyed, in particular methods provided by Genz [47] [48] [49]. His methods were previously programmed and evaluated in Fortran 77.

His first paper (1992) [47] on the subject provides a method for evaluating the general multivariate normal cumulative distribution function as defined by

$$F(\underline{a},\underline{b}) = \frac{1}{\sqrt{|\underline{\Sigma}|(2\pi)^s}} \int_{a_1}^{b_1}\int_{a_2}^{b_2}...\int_{a_s}^{b_s} \exp\{-\frac{1}{2}\underline{\theta}^t\Sigma^{-1}\underline{\theta}\}d\underline{\theta}$$

where $\underline{\theta} = (\theta_1, \theta_2, ..., \theta_s)^t$ and $\underline{\Sigma}$ is a $s \times s$ symmetric positive definite covariance matrix. However, for our purposes we can set $a_1 = a_2 = ... = a_s = -\infty$. The algorithm suggested by Genz for the evaluation of the integral operates as follows:

i) Input $\Sigma, \underline{a}, \underline{b}, \varepsilon, \alpha$ and $N_{\max}$ where ε is defined as the error tolerance, α as the Monte-Carlo confidence factor for the standard error e.g. 2.5 and $N_{\max}$ limits the maximum number of repetitions allowed for the algorithm.

ii) Compute the lower triangular Cholesky factor C for $\underline{\Sigma}$.

iii) Initialise $Intsum = 0$, $N = 0$ and $Varsum = 0$. Furthermore define $d_1 = \Phi(a_1/c_{1,1})$, $e_1 = \Phi(b_1/c_{1,1})$ and $f_1 = e_1 - d_1$.

iv) Repeat:

a) Generate random uniform $w_1, w_2, ..., w_{s-1} \epsilon[0,1]$

b) For $i = 2, 3, ..., s$

$$y_{i-1} = \Phi^{-1}(d_{i-1} + w_{i-1}(e_{i-1} - d_{i-1}))$$

$$d_i = \Phi\left(\frac{a_i - \sum\limits_{j=1}^{i-1} c_{i,j}y_j}{c_{i,i}}\right)$$

$$e_i = \Phi\left(\frac{b_i - \sum\limits_{j=1}^{i-1} c_{i,j}y_j}{c_{i,i}}\right)$$

and

$$f_i = (e_i - d_i)f_{i-1}$$

c) Set $N = N + 1$,

$$\delta = \frac{f_s - Intsum}{N}$$

$$Intsum = Intsum + \delta$$

$$Varsum = \frac{(N-2)Varsum}{N + \delta^2}$$

$$Error = \alpha\sqrt{Varsum}$$

Until $Error < \varepsilon$ or $N = N_{\max}$

v) Output $F = Intsum/N$, $Error$ and N.

This algorithm can be simplified for our purpose by setting $d_i = 0$ since we assume $a_i = -\infty$, $i = 1, ..., s$.

The following example illustrates this approach using parameters which could be chosen for our multi-stage design. Let $n = 4$, $a = (-\infty, -\infty, -\infty, -\infty)$ and $b = (0.5, 0.25, 0.1, 0.05)$. Furthermore let

$$\Sigma = \begin{pmatrix} 1 & 1 & 1 & 0.6 \\ 1 & 1 & 1 & 0.6 \\ 1 & 1 & 1 & 0.6 \\ 0.6 & 0.6 & 0.6 & 1 \end{pmatrix}$$

This means that we want to solve

$$F(\underline{a}, \underline{b}) = \frac{1}{\sqrt{|\underline{\Sigma}|(2\pi)^4}} \int_{-\infty}^{0.5} \int_{-\infty}^{0.25} \int_{-\infty}^{0.1} \int_{-\infty}^{0.05} \exp\{-\frac{1}{2}\underline{\theta}^t \underline{\Sigma}^{-1} \underline{\theta}\} d\underline{\theta}$$

According to step ii) we first need to find the Cholesky decomposition for Σ. To derive $\Sigma = CC^T$ we simply equate coefficients on both sides of the equation

$$\begin{pmatrix} s_{11} & s_{12} & \ldots & s_{1n} \\ s_{21} & s_{22} & \ldots & s_{2n} \\ s_{31} & s_{32} & \ldots & s_{3n} \\ \ldots & \ldots & \ldots & \ldots \\ s_{n1} & s_{n2} & \ldots & s_{nn} \end{pmatrix} = \begin{pmatrix} c_{11} & 0 & \ldots & 0 \\ c_{21} & c_{22} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ c_{n1} & c_{n2} & \ldots & c_{nn} \end{pmatrix} \begin{pmatrix} c_{11} & c_{21} & \ldots & c_{n1} \\ 0 & c_{22} & \ldots & c_{n2} \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & c_{nn} \end{pmatrix}$$

When solving for the unknown parameters for $i = 1, ..., s$ and $j = i + 1, ..., s$ we get

$$c_{ii} = \sqrt{\left(s_{ii} - \sum_{k=1}^{i-1} c_{ik}^2\right)}$$

and

$$c_{ji} = \frac{\left(s_{ji} - \sum_{k=1}^{i-1} c_{jk} c_{ik}\right)}{c_{ii}}$$

Hence for our example

$$\begin{pmatrix} 1 & 1 & 1 & 0.6 \\ 1 & 1 & 1 & 0.6 \\ 1 & 1 & 1 & 0.6 \\ 0.6 & 0.6 & 0.6 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0.6 & 0 & 0 & 0.8 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 0.6 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.8 \end{pmatrix}$$

The first transformation of our integral then gives density

$$F(\underline{a},\underline{b}) = \frac{1}{\sqrt{(2\pi)^4}} \int\limits_{a'_1}^{b'_1} \exp\{-\frac{y_1^2}{2}\} \int\limits_{a'_2(y_1)}^{b'_2(y_1)} \exp\{-\frac{y_2^2}{2}\} \int\limits_{a'_3(y_1,y_2)}^{b'_3(y_1,y_2)} \exp\{-\frac{y_3^2}{2}\} \int\limits_{a'_4(y_1,y_2,y_3)}^{b'_4(y_1,y_2,y_3)} \exp\{-\frac{y_4^2}{2}\} d\underline{y}$$

where

$$a'_i(y_1, ..., y_{i-1}) = \frac{\left(a_i - \sum\limits_{j=1}^{i-1} c_{ij} y_j\right)}{c_{ii}}$$

and

$$b'_i(y_1, ..., y_{i-1}) = \frac{\left(b_i - \sum\limits_{j=1}^{i-1} c_{ij} y_j\right)}{c_{ii}}$$

Thus $a'_1 = a'_2(y_1) = a'_3(y_1, y_2) = a'_4(y_1, y_2, y_3) = -\infty$, $b'_1 = 0.5$, $b'_2(y_1) = b'_3(y_1, y_2) = \infty$ and $b'_4(y_1, y_2, y_3) = 0.0625 - 0.75y_1$. Following this step we can further transform the integral to give

$$F(\underline{a},\underline{b}) = \int\limits_0^{\Phi(0.5)} \int\limits_0^{\Phi(\infty)} \int\limits_0^{\Phi(\infty)} \int\limits_0^{\Phi(0.0625-0.75\Phi^{-1}(z_1))} d\underline{z}$$

Finally, a third transformation gives

$$\begin{aligned} F(\underline{a},\underline{b}) &= \int\limits_0^1 \Phi(0.5) \int\limits_0^1 \Phi(\infty) \int\limits_0^1 \Phi(\infty) \int\limits_0^1 \Phi(0.0625 - 0.75\Phi^{-1}(w_1\Phi(1))d\underline{w} \\ &= \Phi(0.5) \iiint\limits_0^1 \Phi(0.0625 - 0.75\Phi^{-1}(w_1\Phi(1)) \int\limits_0^1 d\underline{w} \end{aligned}$$

This integral may then be further evaluated using inbuilt Stata functions.

Instead of the Monte-Carlo algorithm method given above lattice rules may be used as a more elegant way of evaluating the integral [48]. A further method is suggested in Genz (2004) [49] for the special case of the trivariate normal density. This method is based on Owen (1956) [88] who wrote the standard trivariate normal integral in terms of a bivariate normal integral. Hence we could extend this method for $s > 3$, however, for large s this does not provide a reduction in computation time.

Bibliography

[1] S. Ahnn and S. J. Anderson. Sample size determination for comparing more than two survival distributions. *Statistics in Medicine*, 14(20):2273 - 2282, 1995.

[2] S. Ahnn and S. J. Anderson. Sample size determination in complex clinical trials comparing more than two groups for survival endpoints. *Statistics in Medicine*, 17(21):2525 - 2534, 1998.

[3] A. Almeida, M. Castel-Branco, and A. Falcao. Linear regression for calibration lines revisited: Weighting schemes for bioanalytical methods. *Journal of Chromatography B*, 774:215 - 222, 2002.

[4] D. G. Altman and J. N. S. Matthews. Statistics notes: Interaction 1: Heterogeneity of effects. *BMJ*, 313:486, 1996.

[5] E. Bagiella and D. Heitjan. Predicting analysis times in randomized clinical trials. *Statistics in Medicine*, 20:2005 - 2063, 2001.

[6] P. L. Barratt, M. T. Seymour, S. P. Stenning, I. Georgiades, C. Walker, K. Birbeck, and P. Quirke. Dna markers predicting benefit from adjuvant fluorouracil in patients with colon cancer: a molecular study. *The Lancet*, 360:1381 - 1391, 2002.

[7] F. Barthel, A. Babiker, P. Royston, and M. Parmar. Evaluation of sample size and power for multi-arm survival trials allowing for non-proportional hazards, loss to follow-up and cross-over. *Statistics in Medicine*, 2006.

[8] F. Barthel, P. Royston, and A. Babiker. A menu-driven facility for complex sample size calculation in randomized controlled trials with a survival or a binary outcome: update. *The Stata Journal*, 5(1):123 - 129, 2005.

[9] P. Bauer, J. Rohmel, W. Maurer, and L. Hothorn. Testing strategies in multi-dose experiments including active control. *Statistics in Medicine*, 17:2133 - 2146, 1998.

[10] C. B. Begg and D. H. Y. Leung. On the use of surrogate end points in randomized trials. *J. R. Statistical Society A*, 163, Part 1:15 – 28, 2000.

[11] M. Birkett and S. Day. Internal pilot studies for estimating sample size. *Statistics in Medicine*, 13:2455 – 2463, 1994.

[12] H. W. Block and A. P. Basu. A continuous bivariate exponential extension. *Journal of the American Statistical Association*, 69:1031 – 1037, 1974.

[13] M. Buyse and G. Molenberghs. Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics*, 54:1014 – 1029, 1998.

[14] D. P. Byar. Assessing apparent treatment - covariate interactions in randomized controlled clinical trials. *Statistics in Medicine*, 4:255 – 263, 1985.

[15] C. Chen, H. Wang, and S. Snapinn. Proportion of treatment effect (pte) explained by a surrogate marker. *Statistics in Medicine*, 22(22):3449 – 3459, 2003.

[16] T. T. Chen and T.-H. NG. Optimal flexible designs in phase II clinical trials. *Statistics in Medicine*, 17:2301 – 2312, 1998.

[17] Y. Chen, D. DeMets, and K. Lan. Monitoring mortality at interim analyses while testing a composite endpoint at the final analysis. *Controlled Clinical Trials*, 24(1):16 – 27, 2003.

[18] Medical Research Council. *A Randomised Trial of Two vs Five CT Scans in the Surveillance of Patients with Stage I Teratoma of the Testis.* Clinical Protocol, 1997.

[19] Medical Research Council. *CHARTWEL vs Conventional Radiotherapy in Post Operative Head and Neck Cancer Patients.* Clinical Protocol, 2000.

[20] Medical Research Council. *A Randomised Trial to Assess the Role of Irinotecan and Oxaliplatin in Advanced Colorectal Cancer.* Clinical Protocol, 2003.

[21] Medical Research Council and EORTC. *Radiation Therapy With and Without Combination Chemotherapy in Patients With Resected Anaplastic Oligodendroglioma.* Clinical Protocol, 1996.

[22] D. Cox. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34(2):187 – 220, 1972.

[23] D. R. Cox. A remark on censoring and surrogate response variables. *Journal of the Royal Statistical Society, Series B*, 45:391 - 393, 1983.

[24] D. R. Cox and D. Oakes. *Analysis of Survival Data.* Monographs on Statistics and Applied Probability. Chapman and Hall, London, 1984.

[25] M. J. Daniels and M. D. Hughes. Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine*, 16:1965 - 1982, 1997.

[26] N. E. Day and S. W. Duffy. Trial design based on surrogate end points - application to comparison of different breast screening frequencies. *J. R. Statist. Soc. A*, 159, Part 1:49–60, 1996.

[27] S. J. Day and D. F. Graham. Sample size estimation for comparing two or more treatment groups in clinical trials. *Statistics in Medicine*, 10:33 - 43, 1991.

[28] F. Downton. Bivariate exponential distributions in reliability theory. *Journal of Royal Statistical Society, Series B*, 33:408 - 417, 1970.

[29] J. G. Einspahr, D. S. Alberts, S. M. Gapstur, R. M. Bostick, S. S. Emerson, and E. W. Gerner. Surrogate end-point biomarkers as measures of colon cancer risk and their use in cancer chemoprevention trials. *Cancer Epidemiology Biomarkers and Prevention*, 6:37–48, 1997.

[30] S. Ellenberg. Discussion of 'surrogate markers in aids and cancer trials'. *Statistics in Medicine*, 13:1437 - 1440, 1994.

[31] S. S. Ellenberg and J. M. Hamilton. Surrogate end points in clinical trials: Cancer. *Statistics in Medicine*, pages 405 - 413, 1989.

[32] L. Ensign, E. Gehan, D. Kamen, and P. Thall. An optimal three-stage design for phase ii clinical trials. *Statistics in Medicine*, 13:1727 - 1736, 1994.

[33] M. Evans, N. Hastings, and B. Peacock. *Statistical Distributions - Third Edition.* Wiley Interscience, 2000.

[34] P. Flandre and J. O'Quigley. A two-stage procedure for survival studies with surrogate endpoints. *Biometrics 51*, pages 969–979, September 1995.

[35] T. R. Fleming. Surrogate markers in aids and cancer trials. *Statistics in Medicine*, 13:1423 - 1435, 1994.

[36] T. R. Fleming. Surrogate end points in cardiovascular disease trials. *American Heart Journal*, pages S193 - S196, 2000.

[37] T. R. Fleming, R. L. Prentice, M. S. Pepe, and D. Glidden. Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and aids research. *Statistics in Medicine*, 13:955 - 968, 1994.

[38] US Food and Drug Administration. Update on the tnf blocking agents. *www.fda.gov/ohrms/dockets/ac/03/briefing/3930B1*$_0 1_B - TNF.Briefing.htm$, 2001.

[39] US Food and Drug Administration. Innovation or stagnation: Challenge and opportunity on the critical path to new medical products. *US Dept of Health and Human Services*, 2004.

[40] M. Fréchet. Sur les tableaux de correlation dont les marges sont donnés. *Annales de l'Université de Lyon*, 14:53, 1951.

[41] L. S. Freedman. Tables of the number of patients required in clinical trials using the logrank test. *Statistics in Medicine*, 1(2):121 - 129, 1982.

[42] L. S. Freedman and B. I. Graubard. Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine*, 11:167 - 178, 1992.

[43] M. Gail, R. Pfeiffer, H. C. Van Houweligen, and R. J. Carrol. On meta-analytic assessment of surrogate outcomes. *Biostatistics*, 1:220 - 246, 2000.

[44] M. Gail and R. Simon. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*, 41:361 - 372, 1985.

[45] M. H. Gail. Applicability of sample size calculations based on a comparison of proportions for use with the logrank test. *Controlled Clinical Trials*, 6:112 - 119, 1985.

[46] A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian Data Analysis.* CRC Press, 2003.

[47] A. Genz. Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1:141 - 149, 1992.

[48] A. Genz. Comparison of methods for the computation of multivariate normal probabilities. *Computing Science and Statistics*, 25:400 - 405, 1993.

[49] A. Genz. Numerical computation of rectangular bivariate and trivariate normal and t probabilities. *Statistics and Computing*, 14:151 – 160, 2004.

[50] S. L. George and M. M. Desu. Planning the size and duration of a clinical trial studying the time to some critical event. *Journal of Chronic Diseases*, 27:15 – 24, 1974.

[51] B. Gnedenko and A. Kolmogorov. *Limit Distributions for Sums of Independent Random Variables.* Addison-Wesley, 1954.

[52] A. Gould. Sample size re-estimation: Recent developments and practical considerations. *Statistics in Medicine*, 20:2625 – 2643, 2001.

[53] E. J. Gumbel. Bivariate exponential distributions. *Journal of the American Statistical Association*, 55:698 – 707, 1960.

[54] S. Halabi and B. Singh. Sample size determination for comparing several survival curves with unequal allocations. *Statistics in Medicine*, 23:1793 – 1815, 2004.

[55] M. Halperin, E. Rogot, J. Gurian, and F. Ederer. Sample sizes for medical trials with special reference to long-term therapy. *Journal of Chronic Diseases*, 21:13 – 24, 1968.

[56] D. P. Harrington and T. R. Fleming. A class of rank test procedures for censored survival data. *Biometrika*, 69(3):553 – 566, 1982.

[57] G. Haynam, Z. Govindarajulu, and F. Leone. Tables of the cumulative non-central chi-square distribution. *Case Statistical Laboratory*, 104, 1962.

[58] M. Heo, M. S. Faith, and D. B. Allison. Power and sample size for survival analysis under the weibull distribution when the whole lifespan is of interest. *Mechanisms of Ageing and Development*, 102:45 – 53, 1998.

[59] J. Herson. The use of surrogate end points in clinical trials. *Statistics in Medicine*, 8:403 – 404, 1989.

[60] D. Hosmer and S. Lemeshow. *Applied Logistic Regression.* Series in Probability and Statistics. Wiley, 1989.

[61] S. Jung, T. Lee, K. Kim, and S. George. Admissible two-stage designs for phase ii cancer clinical trials. *Statistics in Medicine*, 23(4):561 – 569, 2004.

[62] G. J. Kelloff, C. C. Sigman, K. M. Johnson, C. W. Boone, P. Greenwald, J. A. Crowell, E. T. Hawk, and L. A. Doody. Perspectives on surrogate endpoints in the development of drugs that reduce the risk of cancer. *Cancer Epidemiology Biomarkers and Prevention*, 9:127–137, February 2000.

[63] W. K. Kelly. Novel trial designs: Which agents and how do we test them? *Urology*, 60 (Supplement 3A):109 – 114, 2002.

[64] D. P. Kelsen. Surrogate endpoints in assessment of new drugs in colorectal cancer. *The Lancet*, 356:353 – 354, July 29, 2000.

[65] W. F. Kibble. A two-variate gamma-type distribution. *Sankhya*, 5:137 – 150, 1941.

[66] G. Kilpatrick and P. Oldham. Sulphonamide prohylaxis in chronic bronchitis - a clinical trial. *BMJ*, 4884:385 – 388, 1954.

[67] A. Kirby, V. Gebski, and A. C. Keech. Determining the sample size in a clinical trial. *Medical Journal Australia*, 177:256 – 257, 2002.

[68] P. P. Koopmans. Clinical endpoints in trials of drugs for cancer: Time for a rethink? *British Medical Journal*, 324:1398–1391, 2002.

[69] E. Korn and R. Simon. Data monitoring committees and problems of lower-than-expected accrual or event rates. *Controlled Clinical Trials*, 17:526 – 535, 1996.

[70] J. Lachin. Introduction to sample size determination and power analysis for clinical trials. *Controlled Clinical Trials*, 2:93 – 193, 1981.

[71] J. M. Lachin and M. A. Foulkes. Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance, and stratification. *Biometrics*, 42(3):507 – 519, 1986.

[72] E. Lakatos. Sample size determination in clinical trials with time-dependent rates of losses and noncompliance. *Controlled Clinical Trials*, 7(3):189 – 199, 1986.

[73] E. Lakatos. Sample sizes based on the log-rank statistic in complex clinical trials. *Biometrics*, 44(1):229 – 241, 1988.

[74] N. D. C. Lewis. *Surrogate Markers in Clinical Trials.* PhD thesis, University of Cambridge, 2001.

[75] S. Lewis and M. Clarke. Forest plots: Trying to see the wood and the trees. *BMJ*, 322:1479 – 1480, 2001.

[76] D. Y. Lin, T. R. Fleming, and V. D. Gruttola. Estimating the proportion of treatment effect explained by a surrogate marker. *Statistics in Medicine*, 16:1515 – 1527, 1997.

[77] P.-Y. Liu and S. Dahlberg. Design and analysis of multiarm clinical trials with survival endpoints. *Controlled Clinical Trials*, 16:119 – 130, 1995.

[78] P.-Y. Liu, S. Dahlberg, and J. Crowley. Selection designs for pilot studies based on survival. *Biometrics*, 49:391–398, June 1993.

[79] P.-Y. Liu, W. Tsai, and M. Wolf. Design and analysis for survival data under order restrictions with a modified logrank test. *Statistics in Medicine*, 17:1469 – 1479, 1998.

[80] J. Lunceford and M. Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 23:2937 – 2960, 2004.

[81] D. Machin, M. J. Campbell, P. M. Fayers, and A. P. Y. Pinol. *Sample Size Tables for Clinical Studies.* Blackwell Science, 2nd edition, 1997.

[82] R. W. Makuch and R. M. Simon. Sample size requirements fo comparing time-to-failure among k treatment groups. *Journal of Chronic Diseases*, 35(11):861 – 867, 1982.

[83] A. W. Marshall and I. Olkin. A multivariate exponential distribution. *Journal of the American Statistical Association*, 1967.

[84] G. Molenberghs, T. Burzykowski, A. Alonso, and M. Buyse. A perspective on surrogate endpoints in controlled clinical trials. *Statistical Methods in Medical Research*, 13(3):177 – 206, 2004.

[85] S. Moolgavkar and A. Knudson. Mutation and cancer: a model for human carcinogenesis. *Journal of the National Cancer Institute*, 66:1037 – 1052, 1981.

[86] C. O'Connor, W. A. Gattis, and T. Ryan. The role of clinical nonfatal end points in cardiovascular phase II/III clinical trials. *American Heart Journal*, pages S143 – S154, 2000.

[87] International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. Ich harmonised tripartite guideline. statistical principles for clinical trials. *Federal Register*, 63(179):49583, 1998.

[88] D. Owen. Tables for computing bivariate normal probability. *Annals of Mathematical Statistics*, 27:1075 - 1090, 1956.

[89] M. Palta and S. B. Amini. Consideration of covariates and stratification in sample size determination for survival time studies. *Journal of Chronic Diseases*, 38(9):801 - 809, 1985.

[90] G. Pan and D. A. Wolfe. Test for qualitative interaction of clinical significance. *Statistics in Medicine*, 16:1645 - 1652, 1997.

[91] K. Pantel. Minimal residual disease - a perfect surrogate marker? *Clinical Cancer Research*, 6, November 2000 (Supplement).

[92] M. K. B. Parmar and D. Machin. *Survival Analysis - A Practical Approach.* John Wiley and Sons, England, 1995.

[93] G. Patil, M. Boswell, M. Ratnaparkhi, and J. Roux. *Dictionary and Classified Bibliography of Statistical Distributions in Scientific Work.* International Co-operative Publishing House, 1984.

[94] B. Peterson and S. L. George. Sample size requirements and length of study for testing interaction in a 1*k factorial design when time - to - failure is the outcome. *Controlled Clinical Trials*, 14:511 - 522, 1993.

[95] S. Piantadosi and M. H. Gail. A comparison of the power of two tests for qualitative interactions. *Statistics in Medicine*, 12:1239 - 1248, 1993.

[96] T. Pincus, G. Koch, H. Lei, B. Mangal, T. Sokka, R. Moskowitz, F. Wolfe, A. Gibofsky, L. Simon, S. Zlotnick, and F. Fort. Patient preference for placebo, acetaminophen (paracetamol) or celecoxib efficacy studies (paces): Two randomised, double blind, placebo controlled, crossover clinical trials in patients with knee or hip osteoarthritis. *Annals of the Rheumatic Diseases*, 63:931 - 939, 2004.

[97] M. Posch, P. Bauer, and W. Brannath. Issues in designing flexible trials. *Statistics in Medicine*, 22(6):953 - 969, 2003.

[98] J. Preissler, K. Lohman, and P. Rathouz. Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. *Statistics in Medicine*, 21:3035 - 3054, 2002.

[99] R. L. Prentice. Surrogate end points in clinical trials: Definition and operating criteria. *Statistics in Medicine*, 8:431 - 440, 1989.

[100] J. Rice. *Mathematical Statistics and Data Analysis - Second Edition.* Duxbury Press, 1995.

[101] P. Royston and A. Babiker. A menu-driven facility for complex sample size calculation in randomized controlled trials with a survival or a binary outcome. *The Stata Journal*, 2(2):151 - 163, 2002.

[102] P. Royston and M. K. B. Parmar. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, 21:2175–2197, 2002.

[103] P. Royston, M. K. B. Parmar, and W. Qian. Novel designs for multi-arm clinical trials with survival outcomes, with an application in ovarian cancer. *Statistics in Medicine*, 22(14):2239 - 2256, 2003.

[104] L. Rubinstein, M. Gail, and T. Santner. Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation. *Journal of Chronic Disease*, 34(9):469 - 479, 1981.

[105] G. Rueckler. A two-stage trial design for testing treatment, self-selection and treatment preference effects. *Statistics in Medicine*, 8:477 - 485, 1989.

[106] H. Sahai and A. Khurshid. Formulae and tables for the determination of sample sizes and power in clinical trials for testing differences in proportions for the two-sample design: A review. *Statistics in Medicine*, 15:1 - 21, 1996.

[107] A. Sankoh, M. Huque, and S. Dubey. Some comments on frequently used multiple endpoint adjustment methods in clinical trials. *Statistics in Medicine*, 16(22):2529 - 2542, 1997.

[108] S. K. Sarkar. A continuous bivariate exponential distribution. *Journal of the American Statistical Association*, 82:667 - 675, 1987.

[109] W. Sauerbrei. The use of resampling methods to simplify regression models in medical statistics. *Applied Statistician*, 48:313 – 329, 1999.

[110] D. J. Schaid, S. Wieand, and T. M. Therneau. Optimal two-stage screening designs for survival comparisons. *Biometrika*, 77:507 – 513, 1990.

[111] A. Schatzkin, L. S. Freedman, J. Dorgan, L. M. McShane, M. H. Schiffman, and S. M. Dawsey. Surrogate end points in cancer research: A critique. *Cancer Epidemiology Biomarkers and Prevention*, 5:947–953, 1996.

[112] C. Schmoor, W. Sauerbrei, and M. Schumacher. Sample size considerations for the evaluation of prognostic factors in survival analysis. *Statistics in Medicine*, 19:441 – 452, 2000.

[113] D. Schoenfeld. The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika*, 68(1):316 – 319, 1981.

[114] D. A. Schoenfeld. Sample-size formula for the proportional-hazards regression model. *Biometrics*, 39(2):499 – 503, 1983.

[115] D. A. Schoenfeld and J. R. Richter. Nomograms for calculating the number of patients needed for a clinical trial with survival as an endpoint. *Biometrics*, 38(1):163 – 170, 1982.

[116] M. A. Schork and R. D. Remington. The determination of sample size in treatment - control comparisons for chronic disease studies in which drop - out or non - adherence is a problem. *Journal of Chronic Diseases*, 20:233 – 239, 1967.

[117] J. R. Schultz, F. R. Nichol, G. L. Elfring, and S. D. Weed. Multiple-stage procedures for drug screening. *Biometrics*, 29:293 – 300, June 1973.

[118] S. Searle. *Matrix Algebra Useful for Statistics.* Wiley, 1982.

[119] V. Sebille and E. Bellisant. Comparison of four sequential methods allowing for early stopping of comparative clinical trials. *Clinical Science*, 5:569 – 578, May 2000.

[120] S. Self, R. Mauritsen, and J. Ohara. Power calculations of likelihood ratio tests in generalized linear models. *Biometrics*, 48(1):31 – 39, 1992.

[121] J. H. Shih. Sample size calculation for complex clinical trials with survival endpoints. *Controlled Clinical Trials*, 16(6):395 – 407, 1995.

[122] J. Shuster. Fixing the number of events in large comparative trials with low event rates: A binomial approach. *Controlled Clinical Trials*, 14:198 – 208, 1993.

[123] J. Shuster and J. Van Eys. Interaction between prognostic factors and treatment. *Controlled Clinical Trials*, 4:209 – 214, 1983.

[124] R. Simon. Optimal two stage designs for phase II clinical trials. *Controlled Clinical Trials*, 10:1–10, 1989.

[125] R. Simon. Bayesian subset analysis: Application to studying treatment - by - gender interactions. *Statistics in Medicine*, 21:2909 – 2916, 2002.

[126] R. Simon, P. F. Thall, and S. S. Ellenberg. New designs for the selection of treatments to be tested in randomized clinical trials. *Statistics in Medicine*, 13:417 – 429, 1994.

[127] E. Slud and L. Rubinstein. Dependent competing risks and summary survival curves. *Biometrika*, 70:643 – 649, 1983.

[128] N. Stallard and S. Todd. Sequential designs for phase iii clinical trials incorporating treatment selection. *Statistics in Medicine*, 22(5):689 – 703, 2003.

[129] S.-B. Tan, K. Dear, P. Bruzzi, and D. Machin. Strategy for randomised clinical trials in rare cancers. *British Medical Journal*, 327:47 – 49, 2003.

[130] R. E. Tarone and J. Ware. On distribution-free tests for equality of survival distributions. *Biometrika*, 64(1):156 – 160, 1977.

[131] P. F. Thall and J. M. Lachin. Assessment of stratum - covariate interactions in cox's proportional hazards regression model. *Statistics in Medicine*, 5:73 – 83, 1986.

[132] P. F. Thall, R. Simon, and S. S. Ellenberg. Two-stage selection and testing designs for comparative clinical trials. *Biometrika*, 75:303 – 310, 1988.

[133] P. F. Thall, R. Simon, and S. S. Ellenberg. A two-stage design for choosing among several experimental treatments and a control in clinical trials. *Biometrics*, 45:537 – 547, June 1989.

[134] P. F. Thall, R. Simon, S. S. Ellenberg, and R. Shrager. Optimal two stage designs for clinical trials with binary response. *Statistics in Medicine*, 7:571 – 579, 1988.

[135] S. Todd, A. Whitehead, N. Stallard, and J. Whitehead. Interim analyses and sequential designs in phase III studies. *Journal of Clinical Pharmacology*, 51:394 – 399, 2001.

[136] T. D. Tosteson and J. H. Ware. Designing a logistic regression study using surrogate measures for exposure and outcome. *Biometrika*, 77:11 – 21, 1990.

[137] A. A. Tsiatis. The asymptotic joint distribution of the efficient scores test for the proportional hazards model calculated over time. *Biometrika*, 68:311 – 315, 1981.

[138] H. Uesaka. Test for interaction between treatment and stratum with ordinal responses. *Biometrics*, 49:123 – 129, 1993.

[139] K. Ulm, C. Schmoor, W. Sauerbrei, G. Kemmler, U. Aydemir, B. Müller, and M. Schumacher. Strategien zur auswertung einer therapiestudie mit der Überlebenszeit als zielkriterium. *Biometrie und Informatik in Medizin und Biologie*, 20:171 – 205, 1989.

[140] J. Whitehead. *The Design and Analysis of Sequential Clinical Trials.* Wiley, 1992.

[141] S. Wieand and T. Therneau. Optimal two stage designs for clinical trials with binary responses. *Controlled Clinical Trials*, 8:20 – 28, 1987.

[142] A. Wienke. Frailty models. *Max Planck Institute for Demographic Research Working Paper WP 2003-032*, 2003.

[143] J. Wittes, E. Lakatos, and J. Probstfield. Surrogate end points in clinical trials: Cardiovascular diseases. *Statistics in Medicine*, 8:415 – 425, 1989.

[144] N. Wolmark, H. Rockette, E. Mamounas, J. Jones, S. Wieand, D. Wickerham, H. Bear, J. Atkins, N. Dimitrov, A. Glass, E. Fisher, and B. Fisher. Clinical trial to assess the relative efficacy of fluorouracil and leucovorin, fluorouracil and levamisole, and fluorouracil, leucovorin, and levamisole in patients with dukes' b and c carcinoma of the colon: Results from national surgical adjuvant breast and bowel project c-04. *Journal of Clinical Oncology*, 17:3553 – 3559, 1999.

[145] A. H. Xiang, H. N. Sather, and S. P. Azen. Power considerations for testing an interaction in a 2*k factorial design with a failure time outcome. *Controlled Clinical Trials*, 15:489 – 502, 1994.

[146] N. A. Yateman and A. M. Skene. Sample sizes for proportional hazards survival studies with arbitrary patient entry and loss to follow-up distributions. *Statistics in Medicine*, 11(8):1103 - 1113, 1992.