Novel Methods for

Early Phase Clinical Trials

Amy Louise Cotterill, B.Sc., M.Sc.

Submitted for the degree of Doctor of Philosophy

at Lancaster University.

August 21, 2015

Novel Methods for Early Phase Clinical Trials

Amy Louise Cotterill, B.Sc., M.Sc.

Submitted for the degree of Doctor of Philosophy at Lancaster University.

August 21, 2015

Abstract

Early phase clinical trials are conducted with limited time and patient resources. Despite design restrictions, patient safety must be prioritised and trial conclusions must be accurate; maximising a promising treatment's chance of success in later large-scale, long-term trials. Increasing the efficiency of early phase clinical trials, through utilising available data more effectively, can lead to improved decision making during, and as a result of, the trial. This thesis contains three distinct pieces of research; each of which proposes a novel, early phase clinical trial design with this overall objective.

The initial focus of the thesis is on dose-escalation. In the single-agent setting, subgroups of the population, between which the reaction to treatment may differ, are accounted for in dose-escalation. This is achieved using a Bayesian model-based approach to dose-escalation with spike and slab priors in order to identify a recommended dose of the treatment (for use in later trials) in each subgroup. Accounting for a potential subgroup effect in a dose-escalation trial can yield safety benefits for

patients within, and post- trial due to subgorup-specific dosing which should improve the benefit-risk ratio of the treatment.

Dual-agent dose-escalation is considered next. In the dual-agent setting, single-agent data, including toxicity and pharmacokinetic exposure information, is available. This information is used to define escalation rules that combine the outputs of independent dose-toxicity and dose-exposure models which are fitted to emerging trial data. This solution is practical to implement and reduces the subjectivity that currently surrounds the use of exposure data in dose-escalation. In addition, escalation decisions and consistency of the final recommended dose-pair are improved.

The focus of the third piece of research changes. In this work, Bayesian sample size calculations for single-arm and randomised phase II trials with time-to-event endpoints are considered. Calculation of the sample size required for a trial is based on a proportional hazards assumption and utilises historical data on the control (and experimental) treatments. The sample sizes obtained are consistent with those currently used in practice while better accounting for available information and uncertainty in parameter estimates of the time-to-event distribution. Investigating allocation ratio's in the randomised setting provides a basis for deciding whether a control arm is indeed necessary. That is, in a randomised trial, whether it is necessary for any patients to be randomised to the control treatment arm.

Acknowledgements

I would like to begin by thanking Thomas Jaki for his supervision, advice and continued patience throughout my PhD.

I would also like to thank John Whitehead who even after retirement continued to work with me to complete the paper we had started work on.

I would also like to thank Daniel Lorand and Jixian Wang for accepting me on an internship with the Early Clinical Biostatistics group at Novartis Oncology in Basel. I learnt a lot through working with the group and thoroughly enjoyed the experience. Thanks also to everyone that I worked with and met while in Basel for making my time there so memorable.

I would also like to thank others who helped along the way including Matt Sperin who provided helpful input before moving to Manchester. Thanks also to others in the department who helped me out during the PhD.

I gratefully acknowledge the funding received towards my PhD as part of Thomas Jaki's NIHR Career Development fellowship (CDF-2010-03-016).

Lastly, I would like to thank my friends and family for their support and encouragement throughout my PhD, it has been greatly appreciated. Thank you.

Declaration

I declare that the work in this thesis has been done by myself and has not been submitted elsewhere for the award of any other degree.

Amy Louise Cotterill

List of Papers

This thesis contains three chapters which are based on manuscripts which have been submitted for publication. I led the work in each manuscript, carrying out the methodological work and simulation studies. This was done under the helpful guidance of the other authors in each case. The relevant appendices can be found at the end of each chapter while the bibliography is listed at the end of the thesis;

- Chapter 3 has been submitted for publication as; A. Cotterill and T. Jaki. Dose-escalation strategies which utilise subgroup information, February 2015;

- Chapter 4 has been accepted for publication as; A. Cotterill, D. Lorand, J. Wang and T. Jaki. A practical design for a dual-agent dose-escalation trial that incorporates pharmacokinetic data. *Statistics in Medicine*, 2015; Early view DOI: 10.1002/sim.6482;

- Chapter 6 has been accepted for publication as; A. Cotterill and J. Whitehead. Bayesian methods for setting sample sizes and choosing allocation ratios in phase II clinical trials with time-to-event endpoints. *Statistics in Medicine*, 2015; Early view DOI: 10.1002/sim.6426.

Contents

List of Figures

List of Tables

List of Abbreviations

AUC: Area Under the Curve.

$C_{\mathbf{max}}$**:** Maximum Concentration after treatment administration.

CRM: Continual Reassessment Method.

DDI: Drug-Drug Interaction.

DLT: Dose Limiting Toxicity.

EWOC: Escalation With Overdose Control.

ICH: International Conference on Harmonisation of technical requirements for registration of pharmaceuticals for human use.

MCMC: Markov Chain Monte Carlo.

MTD: Maximum Tolerated Dose.

MVN: Multi-variate Normal distribution.

N: Normal distribution.

PK: Pharmacokinetic.

sd: Standard Deviation.

TD100θ**:** Dose with probability θ of causing a DLT in a patient.

$\widehat{\mathbf{TD}}100\theta$**:** Estimated TD$100\theta$, equivalently the recommended dose.

TDC100θ**:** Dose combination with probability θ of causing a DLT in a patient.

$t_{\mathbf{max}}$**:** Time after treatment administration at which C_{max} occurs.

Chapter 1

Introduction

1.1 Early Phase Clinical Trials

It can take up to 15 years for a novel treatment to progress through the research and development process before finally being made available to patients. As well as being lengthy, this process is expensive with current estimates exceeding one billion dollars (Mestre-Ferrandiz et al., 2013; Paul et al., 2010). This figure accounts not only for the research and development costs of the successful treatment, but also for the costs incurred from evaluating treatments that were subsequently not pursued. In this thesis, we propose methods which could help to reduce the cost and duration of the clinical trials stage in the research and development process of a novel treatment with minimal negative impact on the trial outcomes.

Clinical trials follow pre-clinical (in vitro and in vivo) studies in the drug development process. A clinical trial is defined in ICH E6 (CDER/CBER, 1996) as "Any investigation in human subjects intended to discover the ... effects [beneficial or

harmful] of an investigational product(s) ... with the object of ascertaining its safety and/or efficacy"; where efficacy is the treatment's "true biological effect" (Piantadosi, 1997). Of interest are experimental/investigational treatments, which are considered throughout this thesis to be one or more treatments administered to patients in a novel application or combination.

Ethical guidelines based on the Declaration of Helsinki (World Medical Association et al., 2013) govern the coduct of clinical trials. The guidelines highlight that the welfare of patients, treated within and outside of clinical trials, is the main priority. Relevant ethical considerations will be mentioned, but are not discussed at length, in this thesis. An overview of the aspects of clinical trial design pertinent to the novel methods proposed in this thesis are given in Chapters 2 and 5. Further practical information on clinical trial design can be found in Pocock (2004) and industry guidelines are provided by the International Conference on Harmonisation of technical requirements for registration of pharmaceuticals for human use.

In this thesis, the term 'early phase clinical trials' refers to non-confirmatory trials, often described as phase I and II. That is, the clinical trials investigating an experimental treatment's relatively short-term safety and/or efficacy in a controlled population. If suitable evidence of safety and efficacy is observed in these early phase clinical trials, then the experimental treatment proceeds to large-scale phase III trials. Phase III clinical trials are confirmatory trials of the treatment's effectiveness, or beneficial effect, in the general patient population under standard use. The treatment must have a sufficiently positive benefit-risk ratio to be considered suitable for use in the general population. If successful, the phase III trial can lead to the treatment

being licensed. Post-licencing research is then carried out to continually monitor the treatment and its use once it is made available for use outside of clinical trials.

Arrowsmith and Millar (2013) show that lack of efficacy is the main reason for a treatment failing in phase III trials - the most costly stage of drug research and development (Paul et al., 2010). Reducing the number of treatments which fail for lack of efficacy in phase III trials could therefore have a big impact on the overall cost of drug development. Two ways of approaching this issue are:

i. To enable phase III clinical trials to stop for futility before the calculated number of patients have been treated in the trial. In this way, less patients are treated with a sub-optimum treatment. This option is becoming increasingly common in phase II trials;

ii. Improving early phase clinical trial designs. Obtaining more accurate inferences from the early phase trials could lead to better informed decisions being made concerning whether or not the experimental treatment should progress to phase III trials.

The latter approach is considered in this thesis because it is the preferred option, due to the resource savings in terms of the design, start up and conduct of part of a phase III clinical trial which would be incurred in carrying out the first option. In each of the novel designs proposed in this thesis, data which are often already available at the design stage of the trial or collected during the clinical trial are utilised more effectively than in current practice. In doing this, the operating characteristics of the trial are improved in some way.

There are many potential benefits of improving the design of early phase clinical trials. As a consequence of improved trial designs, better decisions can be made, decreasing the time and cost of the drug development process. This can be achieved by ensuring that a promising treatment is pursued, and conversely that an ineffacious treatment is dropped early on in clinical trials. A promising treatment should therefore get to market quicker and with a reduced ecomonic burden. This is directly beneficial for trial sponsors. In addition, decreasing the cost of drug development of the treatment increases the chance of it being made available to potential patients, possibly sooner and at a lower cost.

As well as benefits for the potential patient market, patients involved in clinical trials could benefit from improved clinical trial designs. More efficient trials, through making better use of available data, can improve the safety of trials for participants and reduce the number treated with sub-optimal doses/treatments. This in turn may increase participation in trials, possibly further reducing trial durations.

In this work, simulation studies (based on published data from previous clinical trials), are used to compare the properties of standard early phase clinical trial designs to the proposed designs. The proposed designs all use Bayesian methodology and are intended to be practical to implement. The proposed designs are not intended as a sequence of clinical trials, but rather they relate to different settings in which the additional data of interest is likely to be available. Specifically, the use of biomarker, pharmacokinetic and historical survival data are considered.

1.2 Focus of Thesis

The initial focus of this thesis is on phase I dose-escalation trials in oncology which aim to recommend a dose of the experimental treatment for administration to patients in future trials of the treatment. An assumption, that both toxicity and efficacy increase monotonically with dose of the treatment, means that the recommended dose must provide a compromise between being highly toxic yet efficacious and being non-toxic yet inefficacious. Failure to identify the optimum dose of treatment can therefore lead to its failure in later trials investigating its safety and efficacy. More information on the design and conduct of dose-escalation trials is given in Chapter 2.

Dose-escalation of a single experimental treatment is considered in Chapter 3. The proposed dose-escalation method allows for a potential difference in reaction (explicityly toxicity) to the treatment between two subgroups of the patient population. These subgroups can be identified using a biomarker which has been selected based on historical data and/or pre-clinical information which is indicative of potential differences in tolerance to the treatment between the subgroups that it defines. The methodology developed aims to recommend different doses of the experimental treatment for use in each of the pre-defined subgroups, when this is necessary.

In Chapter 4, we go on to consider dose-escalation of a combination of two experimental treatments. In this setting, a dose of each treatment must be recommeded for use in future trials, and hence, a recommended dose-pair is identified from the trial. When designing the combination trial, some data from the single-agent trials of each treatment will be available. Single-agent pharmacokinetic information, which

provides a measure of the exposure of the body to the drug, can be used to obtain desirable exposure intervals for each treatment. The proposed dose-escalation method for a dual-agent treatment uses this information to improve the consistency in estimation of the recommended dose-pair and to reduce the risk of patients experiencing undesirably high exposures, especially in the presence of drug-drug interactions.

The focus of the thesis then turns to phase II clinical trials which collect preliminary evidence of a treatment's efficacy. Restrictions on the size and duration of phase II clinical trials lead to inferences on a treatment's efficacy often being based on a short-term, often binary, endpoint. Inferences based on a binary endpoint generally require observation of fewer patients to reach a conclusion, and hence, they tend to have lower trial costs than those based on a time-to-event endpoint. The short-term (binary or time-to-event) endpoint often used in phase II trials in place the actual time-to-event endpoint of interest (often time to mortality or disease progression) which is to be considered in the phase III trial. The reason for this is that the selected short-term endpoint is expected to be available much sooner than the actual endpoint of interest, and hence, the trial duration and therefore cost is reduced.

If the short-term endpoint is not highly correlated and causally linked with the actual time-to-event endpoint of interest, then data from the phase II trial will poorly predict phase III efficacy. Another instance, which can arise in both oncology and translational medicine, where little is to be gained from using a short-term endpoint is that in which no feasible or worthwhile short-term endpoint is available. An introduction to phase II clinical trials and the common design restrictions is given in Chapter 5 along with an introduction to sample size calculation in this setting.

A novel method of sample size calculation, for a single-arm and a randomised phase II trial, is presented in Chapter 6. The calculation uses Bayesian methodology and is based on a proportional hazards assumption between the (short-term or actual) time-to-event endpoint of interest on the experimental and control treatments. The proposed method of sample size calculation is relevant for an experimental treatment being investigated in an application where a relevant time-to-event endpoint can feasibly be collected, historical data on the control treatment is available and a confirmatory phase III trial will follow. The resulting sample sizes aree greater than the corresponding calculation based on a binary endpoint and the trial duration greater than that based on a shorter-term time-to-event endpoint. However, in the applications discussed, where these are not feasible or worthwhile alternatives, the use of Bayesian methodology to incorporate historical data enables the number of events required in the trial to be reduced when compared to the frequentist counterpart.

The main results, applications, limitations and future work for these topics are discussed in Chapter 7.

Chapter 2

Bayesian Model Based Methods in Dose-escalation

2.1 Dose-escalation Trials in Oncology

Dose-escalation trials are usually first-in-man trials of an experimental treatment in a given application. Despite the necessary focus of these trials on safety, due to the relatively untested nature of the treatment, their main objective is to identify a dose (or doses) of the treatment for exploration in a greater number of patients in trials of its efficacy. In order to maximise the treatment's chance of success in later trials, the dose(s) recommended for use in future trials must be accurately selected. In this thesis we consider identification of a single recommended dose of treatment but the extension of the definition to identify mulitple recommended doses is straight-forward. The recommended dose is an estimate of the optimal dose of the treatment, that is, a dose which is efficacious and has an acceptable level of toxicity.

The ethics of patients involved in dose-escalation trials require that they are exposed to no more risk than is absolutely necessary (World Medical Association et al., 2013). In estimating the optimal dose of a relatively untested treatment it is therefore not plausible to immediately administer patients with a dose of the experimental treatment which has a high chance of being toxic. Erring on the side of caution, and under the assumption that the toxicity of a treatment increases monotonically with dose, a low dose of the treatment is administered to the first cohort of patients. A cohort is a group of patients enrolled at the same stage in dose-escalation and treated with the same dose of the experimental treatment. Only once this dose is found to be tolerated, in terms of its toxic side-effects, can a higher dose be given to the proceeding cohort. This process, along with possible de-escalation or treatment of additional patients at a dose, continues until it is decided that the optimal dose has been estimated within a desired level of accuracy (Pocock, 2004).

The sequential nature of dose-escalation trials means that, in order to control the duration of the trial, the decision over whether to proceed to a higher dose must be based on an endpoint which is available relatively soon after administration of the treatment. For this reason, estimation of the optimal dose of a treatment is generally based on a short-term safety endpoint. An implicit assumption is that as the toxicity of a treatment increases, so does its efficacy. The optimal dose can then be defined as the TD100θ; the dose which has probability θ of causing an unacceptable toxicity in a patient. The TD100θ is hoped to be efficacious enough to be beneficial to patients. The recommended dose from a dose-escalation trial can then be defined as $\widehat{\text{TD}100\theta}$, the estimated TD100$\theta$.

Standard, 'acceptable' values of θ, such as 0.16, (though difficult to justify) exist and are often used. An unacceptable toxicity on the other hand is disease, patient group and treatment target dependent and must be defined on a trial by trial basis. This is usually done in terms of dose limiting toxicities (DLTs). DLTs are toxic side-effects which are felt to be caused by unacceptably high levels of the treatment. This means that even if the treatment was considered efficacious at this dose, its benefit-risk ratio would not be suitable to warrant its administration to patients, hence, limiting the dose of the experimental treatment administered to patients (NCI, 2014). For example: Nausea and vomiting are not acceptable side-effects of an asthma treatment so they would both be included in the list of DLT's. However, in an oncology trial, the seriousness of the condition may lead to vomiting still being classified as a DLT but nausea, though not desirable, may not be.

The work in Chapters 3 and 4 focusses on dose-escalation trials of cytotoxic drugs used to treat cancer patients. Since cytotoxic drugs aim to kill cells, the assumption that toxicity increases monotonically with dose is commonly used. For other cancer treatments, such as protein inhibitors, this may not be the case and a different approach is required (see the design of Zhang et al., 2006, for example).

In oncology trials, treatments are usually administered to cancer patients in cycles (typically of length 21 or 28 days) until the patient's disease progresses or the experimental treatment is withdrawn due to safety concerns. Outside of oncology, phase I trial participants are often healthy volunteers. The patient is therefore expected to return to a 'healthy' status upon stopping the treatment. As a result, it is common for participants to take a break from the treatment before being re-dosed at a

higher dose level. A single patient can therefore contribute data to multiple dose levels (Whitehead et al., 2001). This is rare in oncology trials where participants are usually cancer patients, potentially with no alternative treatment options. Intra-patient dose-escalation can occur in cancer patients but the outcome would be conditional on the previous outcome due to the patient's deteriorating state. Although this dependence is also present in healthy volunteer studies, the trial designs used in such a situation are designed to minimise the impact of this effect. Although this dependence is also present in healthy volunteer studies, the trial designs used in such a situation are designed to minimise the impact of this effect. For this reason, aswell as for ease and trial duration considerations, only binary, cycle 1 DLT information is typically used in dose-escalation decisions in oncology trials. Dose-escalation methods do exist which utilise toxicity data from later cycles (Sinclair and Whitehead, 2014) but this option is not considered in this thesis.

Most of the dose-escalation trial designs discussed in this thesis are transferable to applications outside of oncology, providing that the assumptions underlying the designs are relevant. However, the designs may need to be adapted to suit the specific needs of the patients involved in the trial. Regardless of the patient-group, there are some issues and practicalities that must be addressed when designing a dose-escalation trial which are, more or less, unique from later trials. According to Rosenberger and Haines (2002) and Storer (1989), some of these are:

Non-hypothesis driven: The objective of dose-escalation is to identify the recommended dose of the experimental treatment, not to test specific hypotheses. The

definition of the recommended dose, although fixed for a single trial, can vary between trials.

No control group: Only a small number of patients are treated at each dose of treatment and so comparison between the experimental treatment and a control treatment would be difficult to do reliably. Due to the lack of a concurrent control group, no reliable estimate of treatment effect can be made at this stage.

Ethics: Trial patients must not be exposed to unnecessary risk and so the trial must be efficient, obtaining the maximum possible evidence to accurately estimate the TD100θ using as few patients as possible. If the recommended dose is too high, then patients in later trials will be exposed to unnecessary levels of risk, it is therefore better to be conservative in the estimate of the TD100θ. However, if the estimate is too low then future patients will receive a sub-optimal treatment, possibly resulting in incorrectly abandoning the experimental treatment. A balance between the ethics of trial and future patients must be found.

Small samples: The number of patients exposed to a possibly non-beneficial treatment with unknown toxicity must be minimised. In addition, treating more patients than is required to identify the $\widehat{\text{TD}100\theta}$ with suitable accuracy would lead to an increase in trial duration, delaying progression to efficacy trials of the treatment. As a consequence of the limited number of patients treated in dose-escalation trials, the resulting estimate of the $\widehat{\text{TD}100\theta}$ is highly dependent on the patients selected for the trial.

Sequential, often long observation times are required: Dose-escalation is sequential so that all trial data, including the most recently observed, is used to select the dose for administration to the next cohort of patients. This is done to make the trial as safe as possible for each participating patient. The long trial duration can lead to possible non-treatment related drop-out.

Toxicity grading can be subjective: Clear definitions of what constitutes a DLT must be pre-specied to reduce classification errors.

Patients treated at sub-optimal doses: By the nature of dose-escalation, some patients will be treated below the TD100θ and others above it, though this number must be minimised.

2.2 Bayesian Model Based Methods

During dose-escalation, decisions must be made over when to escalate the dose of the experimental treatment, and by how much. Further, a decision as to when to stop the trial, having identified the $\widehat{\text{TD}100\theta}$ with suitable accuracy, is required. As mentioned in Section 2.1, these decisions will be made based on binary cycle 1 DLT information. In oncology, a treatment administered at the optimum dosage has the potential to make a drastic difference to the lives of the late-stage cancer patients involved in the trial. As a result, rapid escalation is plausible so that fewer patients are treated with inefficacious doses. In trials using cancer patients, there is also an increased tolerance for overdosing than there is in healthy volunteer trials. However, overdosing is still classed as being more dangerous than underdosing and so a balance must be struck;

the rate of escalation must be controlled while not treating unnecessary numbers of patients at overly low doses.

An intuitive dose-escalation trial design invoves specifying simple rules which define when escalation, de-escalation or expansion of a dose should occur, and under what conditions to stop the trial. Such designs are algorithmic, the most widely known and used being the $3 + 3$ design (Carter, 1973). An example of potential desision rules for such a design are given in Section 4.1. Such designs are simple to implement but have many short-comings (e.g. Chen and Beckman, 2009; Goodman et al., 1995; Reiner et al., 1999; Rogatko et al., 2007; Thall and Lee, 2003), not least their non-quantitive definition of the recommended dose as a "dose which, if exceeded, would put patients at unacceptable risk for toxicity" (Rosenberger and Haines, 2002). The recommended dose from these trials is not really an estimate of the TD100θ; it is usually referred to as the maximum tolerated dose.

An alternative class of designs are model-based. Such designs do not require sequential administration of each dose of treatment pre-specified for use in the trial and, hence, allow faster escalation (when required) than algorithmic designs. Model-based designs also have the ability to include safety constraints on escalation. In addition, these designs allow much more flexibility in the design and running of the trial, as well as enabling qualitative definition of the recommended dose as the $\widehat{\text{TD}}100\theta$. This is achieved through assuming some model for the dose-toxicity relationship and updating the model estimates as data arises. At the end of the trial, the dose with expected posterior probability of DLT closest to θ is selected as the $\widehat{\text{TD}}100\theta$.

In the Continual reassessment method (CRM) (O'Quigley et al., 1990), the dose-

toxicity model is described by some function. O'Quigley et al. state that a one-parameter model is suitable to accurately estimate the $\widehat{\text{TD}100\theta}$. Other authors (e.g. Neuenschwander et al., 2008; Whitehead and Williamson, 1998) instead use a two-parameter dose-toxicity model which is better suited than a one-parameter model to model the entire dose-toxicity curve (O'Quigley et al., 1990). This can be advantageous as it allows straight-forward inference (compared to that from the CRM) about doses aside from the $\widehat{\text{TD}100\theta}$ to be drawn. This may be required in practice if updated clinical opinion leads to the target toxicity level θ being changed. Another example where knowledge of the entire dose-toxicity curve can be useful is when multiple doses of the treatment are to be taken to phase II efficacy trials to determine which of the selected doses has the best benefit to risk ratio. Taking forward a dose with a toxicity rate too far below θ would not be beneficial, given the assumption that this corresponds to low efficacy. Knowledge of the entire dose-toxicity curve is required to sensibly deduce this.

When it is not felt that a reasonable assumption can be made concerning the form of the dose-toxicity curve, then a curve-free design may be preferable. Curve-free methods, such as those proposed by Gasparini and Eisele (2000) and Whitehead et al. (2010), have been suggested as non-parametric alternatives to algorithmic and model-based designs. Curve-free designs require specification of the prior expected probability of DLT at each dose of the treatment which is to be made available for administration to patients in the trial. These probabilities are updated during the trial as data arises. As with model-based methods, curve-free designs enable quantitative definition of the recommended dose as the $\widehat{\text{TD}100\theta}$. The operating characteristics of

model-based and curve-free designs are comparable, with clear differences only really arising in the correct/mis- specification of the underlying model for the dose-toxicity relationship in model-based designs (Jaki et al., 2013). A comparison of the properties of algorithmic, model-based and curve-free dose-escalation trial designs is presented by Jaki et al. (2013).

For the work in this thesis, we assume that it is reasonable to assume a model for the dose-toxicity relationship and, therefore, model-based designs are the focus of further discussions on dose-escalation methods. Frequentist model-based designs have been proposed (e.g. O'Quigley and Shen, 1996) but in early phase trials (including dose-escalation) where there is belief in the treatment but little observed data, Bayesian designs can be beneficial. The use of a Bayesian design enables intuitive incorporation of relevant historical data along with available trial data. Increasing the amount of information upon which trial decisions are based can improve the safety of the trial for patients. For this reason, the dose-escalation designs proposed in this thesis use Bayesian methods. Two existing Bayesian model-based designs are described in detail in Section 2.2.2. These are the Bayesian (two-parameter) logistic regression approaches of Whitehead and Williamson (1998) and Neuenschwander et al. (2008). Alternative Bayesian model-based approaches exist (e.g. Babb et al., 1998; Thall and Lee, 2003), all of which are based on the same basic principles.

A literature review of model-based dose-escalation trial designs for a single-agent treatment is given in Section 3.1.1. In Section 3.2, the design of Whitehead and Williamson (1998) is extended to allow different recommended doses to be selected in each of two pre-defined subgroups, if this is deemed necessary. In Section 4.3,

a dual-agent dose-escalation trial design which incorporates pharmacokinetic data is considered. The design underlying this method is based on the dose-escalation method described by Neuenschwander et al. (2008). Through the comparison of the Whitehead & Williamson design and that of Neuenschwander et al., in Section 2.2.2, and demonstration of extensions to these designs in Chapters 3 and 4, it should be clear that most alternative, model-based dose-escalation methods could be extended to similar end.

2.2.1 Bayesian Methods

Bayesian methods are endorsed for use in small clinical trials within the pharmaceutical industry (CHMP et al., 2006). Adamina et al. (2009) discuss the potential benefits of using Bayesian statistics in oncology. Bayesian methods can be especially useful in early phase clinical trials where belief in the experimental treatment heavily outweighs knowledge of it in practice (given that the treatment has progressed to clinical trials). Bayesian trial designs enable prior belief about the treatment to be incorporated into the trial along with trial data. This means that trials are subjective, and consequently the use of Bayesian statistics can be controversial. However, when prior information is wisely incorporated, such designs should be more efficient than their frequentist counterparts because they make better use of available information. Bayesian designs can be seen as a means of formalising learning and, in many cases, the resulting inferences are more natural than those obtained from a frequentist analysis.

In the trial designs presented in this thesis, an understanding of the Bayesian paradigm is assumed (otherwise see Hoff, 2009, for an overview). Take, for example,

Y_{jx} to be a binary indicator of whether patient j, treated with dose x from the set of available doses $\mathbf{d}$, experienced a DLT in the first cycle of treatment. The probability of a DLT at dose x is then given by $\pi(x)$ and the responses $Y_{jx} \sim$ Bernoulli$(\pi(x))$ such that:

$$Y_{jx} = \begin{cases} 1 & \text{if patient } j \text{ experienced a DLT at dose } x, \\ 0 & \text{otherwise.} \end{cases}$$

Say that from n_x patients treated at dose x, $t_x = \sum_{j=1}^{n_x} y_{jx}$ is the number who experienced a DLT in the first cycle of treatment. Set $u_x = n_x - t_x$ as the number of patients who did not experience a DLT at dose x in this time. Now, the random variable of which t_x is a realisation is T_x with $T_x \sim$ Binomial$(n_x, \pi(x))$.The likelihood of the data for dose x is (Chow and Liu, 2014) then:

$$\begin{aligned} f(t_x|n_x, \pi(x)) &= \binom{n_x}{t_x} \{\pi(x)\}^{t_x} \{1-\pi(x)\}^{u_x}, \\ &\propto \{\pi(x)\}^{t_x} \{1-\pi(x)\}^{u_x}. \end{aligned}$$

In the Bayesian setting, a prior distribution $f_0(\pi(x))$ is specified for the probability that a patient treated with dose x experiences a DLT. The Beta distribution is a natural choice of prior distribution for binomial data because it is the conjugate prior to the binomial likelihood. This means that the prior and posterior distributions have the same form and upon observing more data, only the parameters of the posterior distribution are updated and not the distributional shape (Hoff, 2009). In addition, the parameters of the beta distribution have an interpretation which is relevant to

this setting. For the prior $\pi(x) \sim \text{Beta}(a_x, b_x)$, a_x can be interpreted as the number of patients who experienced a DLT and b_x as the number who did not experience a DLT, at dose x. The mean of the distribution is then $a/(a+b)$, the proportion of patients who experienced a DLT at dose x. The posterior distribution of the probability of DLT at dose x, given t_x DLTs observed in n_x patients treated at dose x, is then obtained from the prior and likelihood of the observed data using Bayes Theorem (Bayes and Price, 1763):

$$\begin{aligned} f(\pi(x)|t_x) &\propto f(t_x|\pi(x))f_0(\pi(x)), \\ \Rightarrow \quad \pi(x)|t_x &\sim \text{Beta}(a_x + t_x, b_x + u_x). \end{aligned}$$

Take a simple example: Prior to dosing any patients, the prior belief is that if six patients were treated at dose x, then two of them would experience a DLT in the first cycle of treatment. This can be represented by the prior distribution, $f_0(x) \sim \text{Beta}(2, 4)$, shown by the dashed black curve in Figure 2.2.1. Now, say that in a cohort of three patients treated at dose x, two DLTs were observed. The prior distribution can be updated to give the posterior distribution, $f(\pi(x)|t_x) \sim \text{Beta}(2 + t_x, 4 + u_x) \equiv \text{Beta}(4, 5)$ which is shown by the solid grey curve in Figure 2.2.1.

From the resulting posterior distribution, posterior probabilities can be calculated. For example, the posterior probability that $\pi(x)$ is greater than 0.35 is 0.71 (represented by the shaded area in Figure 2.2.1). So, with observation of two out of three patients in a cohort experiencing a DLT at dose x, the probability that $\pi(x)$ is greater than 0.35 has increased from 0.43, based only upon prior belief, to 0.71. This process

can be repeated after responses are observed from each cohort of patients to obtain the updated distribution of $\pi(x)$. Similarly for other values of x, where x is an element of the available dose set $\mathbf{d}$. This is done under an assumption of independence of the doses making up dose set $\mathbf{d}$.

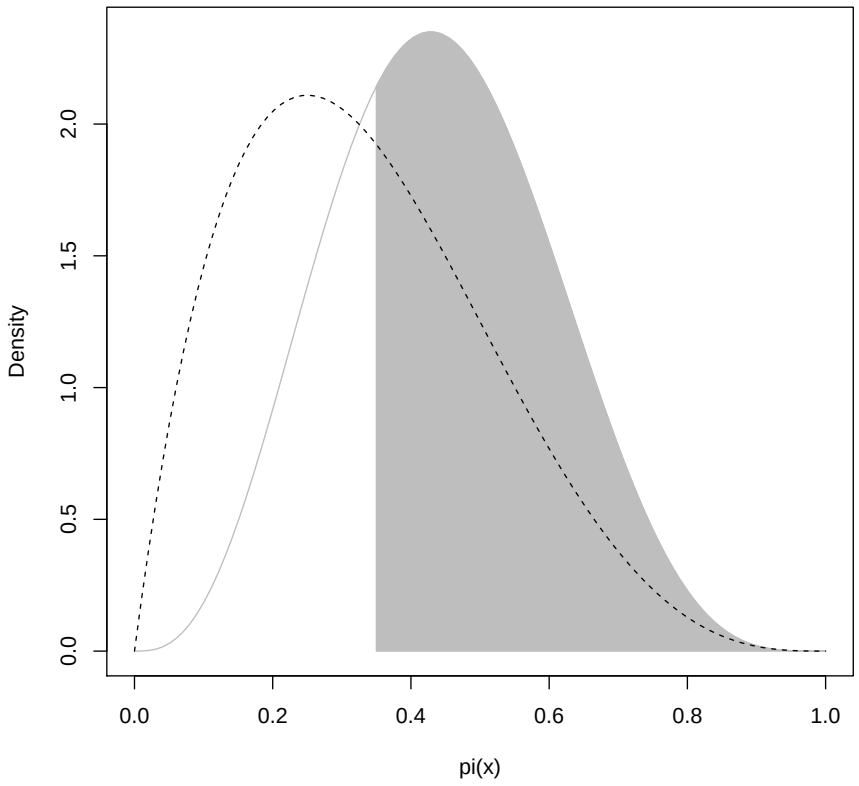

Figure 2.2.1: Prior Beta$(2, 4)$ distribution (shown by the dashed black curve) and posterior Beta$(4, 5)$ distribution (shown by the solid grey curve) of $\pi(x)$ with the shaded area being equal to the posterior probability of $\pi(x)$ being greater than 0.35.

In the example presented, the posterior distribution is tractible due to model conjugacy and the posterior probabilities can be calculated analytically. This will not always be the case. For example, using a logistic regression model to describe the relationship between dose and the probability that a patient experiences a DLT results

in a complex posterior distribution for any given prior. Markov Chain Monte Carlo (MCMC) can be used to obtain inferences from complex distributions of this kind. A short overview of MCMC methods is given here but more details can be found in Robert and Casella (2005). As an alternative to MCMC, for situations involving only a few unknown parameters, numerical methods may be better suited. These methods are less computationally intensive than MCMC and hence, in low dimensions are quicker to obtain inferences from than MCMC.

A Markov chain of length H is a sequence of random variables Z_h, for $h = 1, 2, ..., H$, for which the distribution of Z_h is conditional only on the value of $Z_{(h-1)}$. So, for a given starting value, z_0, the value of z_1 is dependent only upon the value of z_0. Similarly, the value of z_2 is dependent only upon the value of z_1, and so on. The values z_0, z_1, ..., z_H form the Markov Chain.

Markov Chain Monte Carlo methods define a distribution $p(.)$ which has the same limiting distribution as the posterior distribution of interest. A sample from $p(.)$, with the initial section of the chain removed as 'burn-in' (the part of the chain before it converged to a sample of the limiting distribution), is then in effect a sample from the posterior distribution of interest, $f(.)$.

Inferences on the posterior distribution can be calculated from the sample using Monte Carlo methods. Monte Carlo methods take generated samples from the posterior distribution and uses the strong law of large numbers to replace complex integrations with sums. Take $z_1, z_2, ..., z_H$, to be a sample from the limiting distribution $p(z)$ and the posterior distribution of interest to be $f(.)$. The mean of the distribution

and an example of the calculation of an excedence probability are:

$$\begin{aligned}\mathbb{E}[f(z)] = \int_z f(z)p(z).dz &\approx \frac{1}{H}\sum_{h=1}^{H} p(z_h),\\ \mathbb{P}[f(z) > 3] &\approx \frac{1}{H}\sum_{h=1}^{H} \mathbb{I}_{[z_h>3]}.\end{aligned}$$

Computer programs are available which perform MCMC for a range of problems, removing the need to personally program an algorithm to perform the required calculations. Different programs and packages use different algorithms but for the relatively simple problems tackled in this work, the specific choice of algorithm is fairly irrelevant. Simulations were carried out using *R* (R Core Team, 2014) and the MCMC package selected for the work in Chapter 3 was `BoomSpikeSlab` (Scott, 2014) and that in Chapter 4 was `Rstan` (Stan Development Team, 2013). *BoomSpikeSlab* uses variable selection based on the work of George and McCulloch (1997) and Tüchler (2008). *Rstan* is based on a no U-turn sampler which is described in detail in the manual (Stan Development Team, 2012). The packages are discussed in further detail in the relevant chapter.

2.2.2 Conduct of a Bayesian Logistic Regression Approach in Dose-escalation

A range of Bayesian model-based dose-escalation designs have been proposed by different authors. Although the method of escalation and/or selection of the $\widehat{\text{TD}100\theta}$ differs between these methods, the underlying methodology is very similar in most

cases. In these approaches, risk is defined as the probability of experiencing a DLT and the TD100θ is treated as an unknown parameter requiring estimation. Under this definition, the $\widehat{\text{TD}100\theta}$ can be identified from a continuous range of doses, even if it is not one of the discrete doses administered to patients during the trial.

In finding the recommended dose by dose-escalation, it is assumed (as stated in Whitehead and Williamson, 1998) that: i) The probability that a patient experiences a DLT increases monotonically with dose of the treatment, ii) The probability that the treatment is efficacious increases monotonically with dose of the treatment, iii) Information on whether a patient experienced a DLT is available relatively soon after administration of the treatment. Under these standard assumptions, most Bayesian model-based approaches to dose-escalation follow the same basic method:

1. Specify the general trial set-up;

2. Specify the decision rules: For escalation and stopping;

3. Specify the model(s) for the dose-response relationship(s) utilised in the decision rules (defined in Step 2);

4. Specify priors on the parameters of the selected dose-response model(s) (defined in Step 3);

5. Identify a start dose for the trial;

6. Administer a cohort of patients with the dose considered "optimal" for them at their time of entry to the trial, based on the decision rules defined in Step 2;

7. Update the dose-response model(s) (defined in Step 3) with the observed patient responses;

8. Dose-escalation continues by repeating Steps 6 and 7 until one of the stopping criteria (defined in Step 2) is met;

9. Identify the recommended dose for use in future trials of the treatment.

Each of the steps in dose-escalation are described in turn in the remainder of this section. The discussion follows two Bayesian logistic regression approaches and highlights the differences between the methods. The two designs are those of Whitehead and Williamson (1998) and Neuenschwander et al. (2008) which are the underlying designs for the methods of dose-escalation proposed in Chapters 3 and 4, respectively.

Step 1: Specify the general trial set-up

Pre-clinical and historical trial data are extrapolated in order to identify an expected therapeutic dose range of the experimental treatment. Together with practical considerations, which may constrain the dose levels of a treatment available for administration from a continuous range to a set of doses (for a drug in tablet form, say), this information can be used to identify a set of doses to be made avaiable for administration to patients in the trial. Although in practice dose-escalation (using a model-based design) is not constrained to these pre-specified doses, pre-specification is necessary for simulation purposes. Simulation is encouraged by regulatory agencies in clinical trial design (Manolis et al., 2013) to confirm that operating characteristics of the proposed trial are reasonable under a range of potential scenarios.

Operating characteristics (or tradition in some cases) of the design are also likely to influence the choice of cohort size used in the trial. Cohorts consisting of a single patient provide optimal escalation decisions by selecting the next dose for administration using all available data under a Bayesian approach which explores the available dose range (Gerke and Siedentop, 2008). However, larger cohorts of size 3-6 are common. Using larger cohorts means that, in general, more information is obtained at each dose, removing the risk of escalating after a single/couple of observations at a dose. This can slow escalation but comes from algorithmic methods in which de-escalation and re-escalation is not generally considered. Practical reasons such as timings of dose-escalation meetings (in which clinical and statistical experts meet to discuss the next escalation step) can also be motivators for inflated cohort sizes. Under Bayesian methodology, the cohort size can differ between cohorts. Small cohorts could be used at the start of the trial when the probability of a patient experiencing a DLT is expected to be low. At higher doses, where there is greater uncertainty over the expected toxicity, larger cohorts can be used.

Knowledge of the treatment area is used to draw up a list of toxicities which are considered to be dose-limiting for the treatment of interest. The time-frame in which these toxicities will be considered to impact dose-escalation decisions (commonly the first cycle of treatment in oncology) is decided. Another practical consideration is the maximum number of patients available for the trial. This might be based on the availability of resources, prevelance of the condition and expected recruitment rates.

Another design consideration is the definition of the recommended dose. Whitehead and Williamson (1998) select a single value of θ, as a toxicity level which, under

the assumption that toxicity increases monotonically with efficacy, implies a suitable level of efficiacy of the treatment without unnecessary toxicity. Based upon this, the TD100θ is clearly defined as the dose with probability θ of causing a DLT in a patient. Ideally, the recommended dose would then be defined as the estimate of the TD100θ resulting from the trial. Practical additions to this definition are often necessary. For example, these could restrict selection of the recommended dose to those administered in the trial, or to doses with probability of a patient experiencing a DLT less than some value, δ (for $\delta > \theta$). If this were the case, then the TD100θ corresponds to the dose with the targeted toxicity level θ which satisfies an additional safety criterion based on toxicity rate δ. As well as its used in the definition of the recommended dose, δ can be useful in escalation to control the rate of escalation and reduce the chance of undesirably large escalation steps being taken in the presence of uncertainty. The use of δ in escalation is discussed further in the explanation of Step 2.

Instead of defining a point probabilities of toxicity, Neuenschwander et al. (2008) classify the probability of DLT in relation to its expected efficacy. For example, a dose x with probability $\pi(x)$ of causing a DLT in a patient is classified:

- for $\pi(x) \in [0.00, 0.16]$ as an underdose;
- for $\pi(x) \in (0.16, 0.35]$ as being in the target toxicity interval; and
- for $\pi(x) \in (0.35, 1.00]$ as an overdose.

This is simply an alternative method of defining the target toxicity. Although it is a more general criteria, the toxicity classifications clearly define the boundaries between acceptable and unacceptable levels of toxicity of the treatment. The use of

the toxicity interval instead of using a point estimate accounts for the lack of power to detect a single value of θ in relatively small dose-escalation trials. The recommended dose in this case is defined as the dose which maximises the posterior probability of being in the target toxicity interval. Using the TD100θ notation, the recommended dose by this method has $\theta \in (0.16, 0.35]$. As before, the recommended dose could be restricted to administered doses and a safety criterion involving δ can be incorporated.

So, in this case (as with the method of Whitehead & Williamson) the recommended dose is not truly $\widehat{\text{TD}}100\theta$ due to the safety criterion and potentially the restricted dose set available for administration to patients. In this thesis, $\widehat{\text{TD}}100\theta$ refers to the recommended dose definition, including safety and/or other constraints on the selection of the recommended dose.

Step 2: Specify the decision rules

In estimating the TD100θ, two kinds of decision rules need to be specified: The escalation rule determines when to escalate and by how much, and the stopping rule which determines when to stop the trial either for safety concerns or having estimated the TD100θ with a suitable level of accuracy.

A fully Bayesian procedure administers patients with the dose of treatment which maximises a specified gain function. Difficulties of such an approach can arise in defining the gain function. Several gain functions are defined in Whitehead and Williamson (1998). The gain function of interest to us is the patient gain;

$$\frac{1}{\{\hat{\pi}(x) - \theta\}^2},$$

where $\hat{\pi}(x)$ is the posterior estimate of the probability that a patient administered dose x of the experimental treatment experiences a DLT. The model used to estimate $\hat{\pi}(x)$ is discussed in Step 3. Based upon this gain function, patients are administered the dose which, based on the posterior modal estimates of the model parameters, has posterior probability of toxicity closest to the (single) target toxicity level θ. So, use of this gain function leads to patients being administered the dose (from those available in the trial) which is optimal for them based on all currently available data. Basing the estimate of $\hat{\pi}(x)$ on the posterior modal estimates of the model parameters, as opposed to the full posterior distribution, can be considered a waste of information. However, for the small amounts of data available in dose-escalation, it can be argued that the choice of estimate has little effect. The posterior modal estimates are used here instead of alternative point estimates such as the mean and median. This is because, for the model and priors defined by Whitehead & Williamson, the modal estimate is derived through conjugate analysis where the other inferences are not. This make the modal estimate less computationally intensive to derive.

The patient gain function can be employed as the only escalation rule in a dose-escalation trial (as in Whitehead and Williamson, 1998). However, doing so can lead to undesirably rapid escalation in some situations. For example, trial data overcoming prior data early on in the trial leading to the skipping of multiple pre-specified doses. One method of controlling escalation is to specify maximum increases for escalation steps. Alternatively, or as well as incorporating this restriction, safety criteria based on the model estimates can be introduced (in a similar way to the Escalation With Overdose Control criteria used by Babb et al., 1998). This can be achieved by

extending the definition of the escalation rule to:

- Administer patients the dose which, based on the posterior modal estimates of the model parameters, maximises the patient gain $1/\{\hat{\pi}(x) - \theta\}$ for estimate $\hat{\pi}(x)$ based on the assumed model, **within doses which satisfy** $\hat{\pi}(\mathbf{x}) < \delta$.

Whitehead and Williamson (1998) suggest alternative formulations of the patient gain which are more conservative and also propose gain functions based around information gain. Use of the information gain can lead to quicker and potentially improved identification of the TD100θ over the patient gain. However, it is not the preferred gain function because it is not as beneficial for patients involved in the trial as the patient gain. This is because the dose administered to patients is that which maximises the information which can be obtained; this is beneficial to investigators but may lead to patients being dosed sub-optimally based on current information.

Neuenschwander et al. (2008) define a target toxicity range, as opposed to a single point value, and utilise the entire posterior distribution in making inferences from the model. This is done to allow for uncertainty in $\hat{\pi}(x)$ and that which surround the choice of a single target toxicity, θ. The counter-part escalation rule to that of Whitehead and Williamson (1998) which accounts for these design differences is;

- Administer patients the dose which, based on the full posterior distribution of $\hat{\pi}(x)$, has **maximum posterior probability based on the assumed model of being in the target toxicity interval** within doses with posterior probability of being classified as an overdose $< \delta$.

The other decision rules that we consider are stopping rules. Ideally these will

come into force when the TD100θ has been estimated with suitable accuracy. This decision could be based upon the width of credible intervals around the estimate of the TD100θ (Whitehead and Williamson, 1998). Equivalently, for the Neuenschwander et al. (2008) set-up, having a posterior probability of being in the target toxicity interval greater than some boundary could warrant stopping the trial for accuracy. Under both methods, it is difficult to specify the boundary defining accurate estimation of the TD100θ. To reduce the chance of prematurely stopping the trial for accuracy, checks based on the definition of the TD100θ identified from the trial can also be incorporated. For example, ensuring that at least 9 patients have been treated at the estimated TD100θ or that doses above the estimated TD100θ do not satisfy the safety criteria.

Practical and ethical reasons warrant the use of two additional stopping rules. A rule of practicality may be: once a given number of patients have been treated in the trial, escalation ceases (if it has not already done so for accuracy). A rule that ensures that the trial is ethical enables escalation to stop if no dose from those available for the trial satisfies the safety constraint on escalation. Implicitly, this stopping rule implies that doses below the pre-specified dose range are expected to be too low to be efficacious regardless of their toxicity. If this is not the case then de-escalation could occur within an extended dose range.

Step 3: Specify the model(s) for the dose-response relationship(s)

The designs of Whitehead and Williamson (1998) and Neuenschwander et al. (2008) both utilise only one dose-response model in dose-escalation. That is, they both model

the dose-toxicity relationship. They consider the probability of a patient experiencing a DLT at dose x, $\pi(x)$, to be suitably modelled by a two-parameter logistic regression model with logit link function such that, for linear predictor η and reference dose d^*;

$$\begin{aligned} \log\left\{\frac{\pi(x)}{1-\pi(x)}\right\} &= \beta_0 + \beta_1 \log\left(\frac{x}{d^*}+1\right), \\ \Rightarrow \pi(x) &= \frac{e^{\eta}}{1+e^{\eta}} \text{ for } \eta = \beta_0 + \beta_1 \log\left(\frac{x}{d^*}+1\right). \end{aligned} \tag{2.2.1}$$

Instead of using the dose of treatment directly in this model, the transformation $\log(x/d^* + 1)$ was used. One is added to the standardised dose to enable the model to handle a zero dose of treatment. Although this transformation seems unnecessary for the single-agent trials discussed so far, it becomes more relevant later on when combination trials are discussed.

The reference dose d^* is used to standardise the actual dose in the model. This transformation, as well as taking the log of the standardised dose, changes the scale that doses are considered on. In this thesis, this is the transformation used to demonstrate the methods.Alternatively, the untransformed dose or another one-to-one transformation of dose could be used with no negative impact on inferences or interpretation. The transformation selected for use in the model should be one which produces an suitable increase in the probability of toxicity for a one unit increase in transformed dose. This is therefore more of a clinical than statistical consideration with experience showing that relationships based on untransformed dose usually lead to escalation decisionswhich are considered to be too cautious.

The proposed dose-escalation design in Section 3.2 extends this dose-toxicity model

to a four-parameter model with the additional terms relating to subgroup membership. In Section 4.2, the two-parameter dose-toxicity model is extended to a five-parameter model for a dual-agent treatment and in addition, a dose-exposure model is specified. For ease of notation in Chapter 4, the two-parameter dose-toxicity model in Equation 2.2.1 is re-parameterised with $\log(\alpha)$ and β in place of β_0 and β_1 to make the notation clearer in the dual-agent setting. These extended/additional dose-response models and their corresponding use in dose-escalation are discussed in the Chapters 3 and 4, respectively, alongside examples of prior specification for each. Prior specification on the model parameters is also discussed in Step 4.

Step 4: Specify priors on the model parameters

It is possible to incorporate available, relevant historical information on the experimental treatment into the prior(s) on the dose-response model(s). Whitehead and Williamson (1998) use the intuitive interpretation of the Beta distribution with prior data (as described in Section 2.2.1) to specify Beta distributions on the probability of toxicity at two independent doses. In order to avoid the prior over-riding trial data, which may contradict the prior, the prior data are down-weighted compared to the trial data. The result is that down-weighted prior data are incorporated into the model as if it were trial data under this design. See Tsutakawa (1975) for full details of this prior derivation and Whitehead and Williamson (1998) for details on eliciting such a prior from expert opinion.

As an alternative method of prior elicitation, Neuenschwander et al. (2008) propose utilising relevant historical data to specify a bi-variate normal distribution on the

model parameters β_0 and $\log(\beta_1)$. This can be achieved using a meta-analytic type approach similar in derivation to a power prior (Ibrahim and Chen, 2000). The parameters of the bi-variate normal distribution can be found as follows:

1. Obtain a relatively non-informative prior on β_0 and $\log(\beta_1)$:

    i. Assume a median probability of DLT at the reference dose d^*;

    ii. Assume that doubling the dose of treatment would double the odds of a patient experiencing a DLT;

    iii. Assume large standard deviations for each parameter and that the correlation between the parameters is equal to zero;

    iv. The resulting confidence intervals for the probability of toxicity will cover most of the probability space.

2. Update the non-informative prior derived in Step 1 with the historical toxicity data;

3. Assume some level of between trial heterogeneity for the historical trial and that to be performed;

4. Use the assumed between trial heterogeneity to increase the variance of β_0 and $\log(\beta_1)$ using a meta-analytic type approach to obtain a weakly-informative prior distribution for use in the trial.

In the prior elicitation methods used by Whitehead and Williamson (1998) and Neuenschwander et al. (2008), the prior data is effectively down-weighted compared

to the trial data. This is done to account for heterogeneity between data which will be collected in the current trial and the historical data used in prior elicitation. We therefore expect that, the more similar the current and historical trials, the less the historical data is down-weighted. Whitehead and Williamson (1998) select the weight of the historical data in terms of the total number of patients the data will represent in the posterior distribution. Neuenschwander et al. (2008) instead take all available historical data and increase the variance of the resulting distribution to account for heterogeneity. Both of these methods have the same general effect of decreasing the information provided by the prior.

A lack of relevant prior information does not render Bayesian designs useless. If no suitable prior information is available, then a weakly informative prior can be used. This could be derived from Step 1 of the prior derivation for the design of Neuenschwander et al. (2008). Alternatively, and arguably more useful, is to specify a weakly informative prior which helps to control the operating characteristics of the trial. For example, setting the prior such that the desired start dose for the trial is that which optimises the specified gain function. Also, specifying the prior such that escalation resulting from its use is suitably cautious, under a range of likely scenario's. Specifying the prior in this way leads it, in some sense, to take the role of escalation rules which restrict the rate of escalation. Such a prior can be obtained for the Whitehead and Williamson (1998) method with hypothetical data, as opposed to elicited data, being used and a range of likely trial scenarios being investigated.

When the prior is selected to control the operating characteristics of the trial, the prior data is heavily down-weighted, to say $1/10^{\text{th}}$ of the planned sample size of the

trial (as in Whitehead and Williamson, 1998). This means that the prior help to will control escalation at the beginning of the trial but will quite easily be over-powered by trial data. Later escalation decisions are therefore expected to be more heavily data driven. An example of such a prior is demonstrated in the work in Chapter 3. On the other hand, the work in Chapter 4 demonstrates the setting where relevant data is available from historical trials and so this is not as heavily down-weighted.

Step 5: Identify a start dose for the trial

Ideally, the start dose for the trial would be selected as the dose which is optimal based on the specified prior and gain function. This may be the case when the prior is chosen to control the operating characteristics of the trial, with consideration of the toxicity of the desired start dose, this may be the case. Alternatively, the start dose can be forced to be a desired dose (from those classified as safe by the safety criterion). When this is the case, escalation rules which constrain the rate of escalation will probably over-ride model decisions in the initial few cohorts treated in dose-escalation. This is because the data acquired in these initial cohorts are gaining evidence on low doses which, based on prior data and the model specified, are already believed to be safe.

Step 6: Administer patients with the "optimal" dose

The first cohort of patients treated in the trial will be administered with the start dose for the trial. This could be a forced dose or optimal based on prior knowledge, as discussed in Step 5. Later cohorts of patients are administered the dose which is optimal based on the escalation criteria defined for the trial (as discussed in Step 2).

The dose advised for escalation by the model is the optimal dose based on the

model specified, escalation criteria defined and information inputted to the model. In practice, the dose advised by the model is not always administered to patients since experts review additional historical or current trial data (including safety data not used in the model, efficacy and pharmacokinetic data), along with the model recommendation, in selecting the actual dose for administration. Data used in escalation decisions which is not formally incorporated in the escalation criteria is being used subjectively.

Step 7: Update the dose-response model(s)

The specified dose-response model(s) are updated with the observed patient responses and inferences required for the trial decision rules are obtained from them.

Step 8: Dose-escalation continues until a stopping criterion is met

The trial continues through escalation, including possible stationary or even de-escalation steps by repeating Steps 6 and 7 until one of the stopping rules specified for the trial in Step 2 is met. At this point, dose-escalation stops and the model is updated with all information obtained in the trial.

Step 9: Identify the recommended dose

When one of the stopping rules is met, dose-escalation stops. By this point, the estimate of the TD100θ should have converged to the true value if a suitable number of patients have been treated in the trial. In practice it is common to confirm safety of the recommended dose by carrying out a dose-expansion trial but this stage in trials is not covered in this thesis.

If the trial was stopped for safety concerns, then there is no dose recommended for use in future trials. The binary nature of the toxicity endpoint used mean that chance toxicities at low doses can lead to this, even when it is not necessarily the case. If, instead, the trial stopped for non-safety related reasons, then a recommended dose for use in future trials can be identified. The most intuitive definition of the recommended dose is the $\widehat{\text{TD}100\theta}$, the dose which optimises the escalation criteria based on all available information. As discussed in Step 1 with regard to the definition of the recommended dose, in practice other criteria may also need to be considered in defining the recommended dose.

2.3 Dual-agent Dose-escalation

A single drug can make an effective treatment but sometimes cells develop resistance or need to be targeted through multiple pathways in order for the treatment to be efficacious (Greco et al., 1996). In an attempt to overcome these problems, or simply to increase the efficacy of a treatment, multiple drugs can be administered as a combination treatment.

In this thesis, any novel treatment which is under investigation is called an 'experimental treatment'. In a clinical trial of a mono-therapy/single-agent treatment, there is only one experimental treatment. When more than one experimental treatment is administered to a patient, the combination is referred to as the 'combination treatment'. For ease, when the context is clear, the combination treatment will simply be referred to as the treatment. Standard treatments which cannot ethically be

withheld from patients are not classified as experimental treatments. So, an experimental treatment administered alongside standard treatment is still referred to as a single-agent. Similarly, when two experimental treatments are administered on top of standard treatment, the dual-agent combination treatment refers only to the two experimental treatments. The dual-agent setting is the one discussed in the remainder of this thesis when referring to a combination, unless specifically stated otherwise.

The objective of a dose-escalation trial of a combination treatment is to identify the toxic dose combination with probability θ of causing a DLT in a patient (TDC100θ). An obvious complication over the single-agent setting is that there are now multiple drugs to escalate. The number of available dose levels of each drug can make it implausible to test each possible combination in sequence. When this is the case, only a range of the possible dose combinations can be tested.

Another complication in the dual-agent setting is that drug-drug interactions (DDIs) can occur. In this work, DDIs are assumed only to occur between the experimental treatments and not with standard treatment (alternatively for the effect to be consistent across the trial population under an assumption of homogeneity). When a DDI occurs, there will either be a synergistic or an antagonistic reaction compared to the event of no interaction. Synergy is when the experimental treatments work together to produce a beneficial effect greater than that expected from either single-agent in the case of no interaction. Antagonism on the other hand is when the experimental treatments work against each other, resulting in an overall beneficial effect which is less than that expected in the case of no interaction. Difficulties arise in defining 'no interaction' since multiple models have been defined to

describe it and with none being distinctly better than the other in many situations (Sühnel, 1998). This issue is addressed in Section 4.2 in relation to the dose-response models used for toxicity and exposure data, where exposure is a measure of the concentration of a drug in the body (more details on exposure data are given in Section 2.4.2).

DDIs can affect toxicity and/or exposure (Rodrigues, 2008). Dose-toxicity and dose-exposure interactions are not always aligned (proportionally, say) and can not be reliably predicted from pre-clinical data. Greco et al. (1996) give the example of an in vitro test of a combination treatment which results in an antagonistic DDI being expected in the clinical setting. This may be the observed interaction. On the other hand, if the combination is more specific to tumour cells than normal cells, then the effect may be decreased toxicity, despite increased exposure being observed due to the antagonistic reaction occurring. Greco et al. (1996) suggest that until there is better knowledge of drug pathways and modes of action, pre-clinical DDI information should be used to obtain a better idea of the mechanism of the drugs' actions rather than as a predictor of clinical outcome.

We maintain the assumption that toxicity increases monotonically with dose for a single-agent treatment. In the combination setting, this leads to marginal monotonicity. However, the monotonicity assumption is not used in the two-dimensional space. Since we wish to identify the TDC100θ, the dose-toxicity surface (which combines the dose-toxicity curves of each of the two drugs when administered in combination) is of interest. The plot in Figure 2.3.1 shows an example dose-toxicity surface. It can be seen that in the combination, multiple dose-pairs have probability θ of causing a

DLT in a patient. Information other than DLT data (such as exposure data) is useful in selecting which of the dose-pairs, with probability θ of causing a DLT in a patient, should be recommended for use in future trials of the treatment. More information about dose-response surfaces in relation to the choice of drug-drug interaction models can be found in Greco et al. (1995).

Dose-escalation of a combination treatment follows the same general steps as described in Section 2.2.2. At the time of design of a combination dose-escalation trial, single-agent trials of both of the relevant treatments will have been completed and pre-clinical trials on the combination will have been carried out. This means that some safety, pharmacokinetic and possibly efficacy data available will be available on the single-agent treatments. This information can be used in defining the decision rules and prior distribution used in the combination trial. Although pre-clinical information on DDIs is highly speculative, this data can used in specifying priors on the model parameters relating to the possible DDI. The high level of uncertainty over the interaction is encompassed through a high prior variance.

A review of Bayesian, dual-agent dose-escalation methods is given in Section 4.1. In the remainder of Chapter 4, a design for utilising pharmacokinetic exposure data in dual-agent dose-escalation is considered.

2.4 Utilising Additional Data in Dose-escalation

Some of the benefits of using Bayesian dose-escalation trial designs, such as those described in Section 2.1, are being recognised and model-based designs are becoming

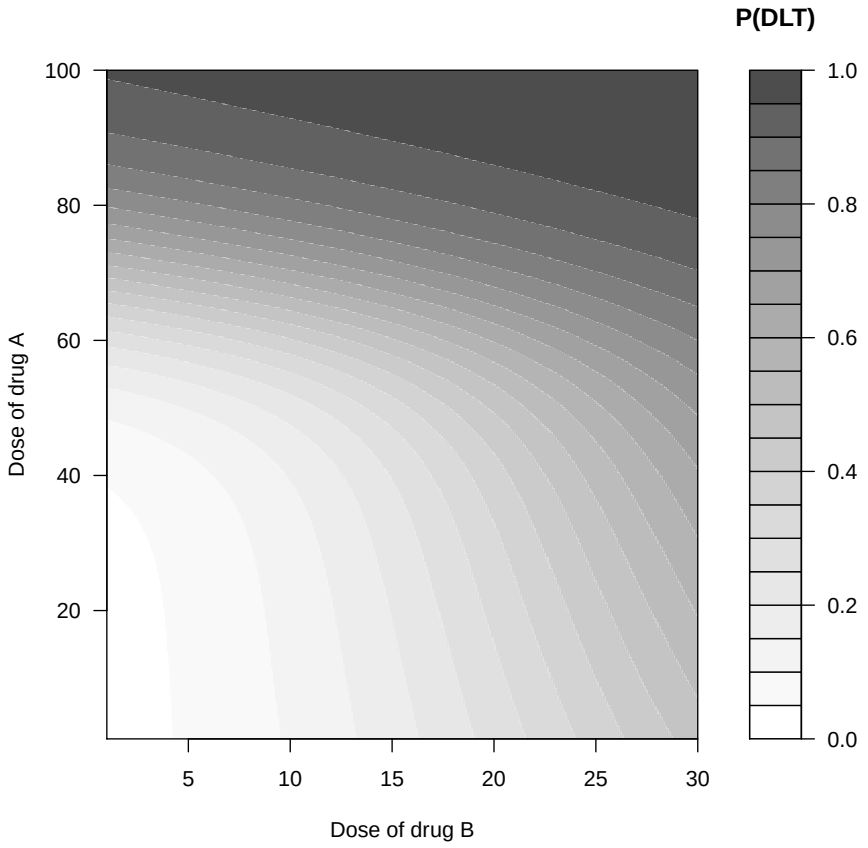

Figure 2.3.1: Example dose-toxicity surface for a combination treatment of drugs A and B.

more common in practice (Biswas et al., 2009). A recent review by Dahlberg et al. (2014) concluded that sample sizes of phase I trials are increasing. This reflects recognition from trial sponsors of the importance of obtaining accurate results from phase I trials. Despite these advances, there is still work to be done in this area. Paul et al. (2010) recognise that spending more on phase I trials could decrease drug development costs in the long-term. Considering data other than DLT information during dose-escalation can be beneficial to long-term drug development by improving decisions made in, and resulting from, dose-escalation trials.

In order for the incorporation of additional data in trial designs to be feasible in practice, the extra information needed to set up and run the trial must be easily obtained within the constraints of the trial. The fact that of the 98 drugs approved in 2000, only 27 were chemically new (Schmid and Smith, 2004), shows that often, instead of chemically new treatments being developed, new applications (e.g. populations, diseases) or useful combinations of already accepted treatments are being found. In these situations, the treatment still needs to be shown to be efficacious but initial research costs involved in drug discovery are minimal (Schmid and Smith, 2004) and speed of development compared to a chemically new treatment is increased. It also means that data are likely to be available which can be used to aid the design of future trials of a treatment in the new application or combination.

In this thesis, we consider using biomarker and pharmacokinetic data to improve escalation decisions. These data are discussed in detail in Sections 2.4.1 and 2.4.2 before being incorporated into dose-escalation designs in Chapters 3 and 4, respectively. The proposed dose-escalation designs are practical to implement since the additional data they require during the trial are available within reasonable time constraints.

2.4.1 Biomarker Data to Identify Patient Subgroups

The reaction to a certain treatment may differ between subgroups of a patient population. This reaction could cause a change in the way the body processes the drug, affecting the safety and/or efficacy of the treatment. For example, the presence of KRAS mutations in colorectal cancer is a reliable predictor of poor response to specific common treatments for this disease (Lièvre et al., 2006). Indicators of differences

in reaction to treatment, such as presence of KRAS mutation, can be referred to as biomarkers of susceptibility (WHO, 1993). These biomarkers include physical characteristics such as age, gender or ethnicity, as well as genetic differences. From here on, the term 'biomarker' is used to refer to a biomarker of susceptibility.

Currently, the over-riding use of biomarkers in early phase clinical trials is to exclude certain patient subgroups from treatment (in order to justify an assumption of a homogeneous trial population). If, before the trial, it is known that no dose of the treatment has a suitable benefit-risk ratio in the subgroup, then it may be just as well to exclude members of that subgroup from the trial. It is possible though that in the potentially excluded subgroup, a lower dose (that identified for the remaining patient population) could have suitable efficacy gains for use as a secondary line of treatment in this subgroup, say. Conversely, ignoring potential subgroup effects is also not ideal because it can lead to a diluted treatment effect. When this is suspected to be the case, it can be investigated through phase II/III enrichment designs (see Temple, 2005, for a short overview of such designs). Often these designs drop subgroups with the lowest efficacy, resulting in the use of a sub-optimal dose in the remaining population. A literature review of clinical trial designs which account for potential subgroup effects is given in Section 3.1.2.

Due to a limited number of patients being involved in early phase clinical trials, reliable identification of relevant biomarkers within the trial is unrealistic. Potential subgroups of interest can, however, be identified before the trial begins in some cases. For example, differences between patient reactions in historical trials of the same treatment in another application, or of a treatment with similar action in the same

application, can be used to identify a biomarker of interest. Texts (e.g. Jain, 2010) which report the findings of exploratory trials to produce lists of biomarkers for specific diseases which are likely to be influential are also available.

There is therefore scope for using subgroup data, based on a pre-defined biomarker, to advise dose-escalation (see ICH E6 CDER/CBER, 2011). The trial population could then be wider, increasing the population that the treatment could be found to be efficacious in. Allowing different optimal doses to be estimated in each subgroup can improve the benefit-risk ratios for patients. The potential loss of accounting for a subgroup effect when in fact there is not one, is considerably less detrimental than not accounting for a subgroup effect when in fact there is one.

2.4.2 Pharmacokinetic Data in the Combination Setting

Pharmacokinetic (PK) data measures the exposure, or level, of the drug in the body. It is concerned with the absorption, distribution, metabolism and excretion of a drug (Källén, 2008). The PK properties of a drug are thoroughly investigated as part of pharmacology trials which monitor patients dosed at a specific range of concentrations of the experimental treatment. Although an in-depth analysis of the PK properties of a drug is possible in dose-escalation trials, some PK data are routinely obtained from some, or all, trial patients to form part of the treatment's safety profile. Despite this, its use in dose-escalation is at present often only informal, if at all. It is therefore plausible that benefit could be gained by formally incorporating this already available PK information into dose-escalation trial designs.

Jambhekar et al. (2009) give an introduction to PK data, the basics of which are described here. The PK exposure parameters of interest to us can be calculated from a patient's concentration-time curve. This is obtained by taking blood samples from the patient at regular intervals (for example, at times $t = 0.5, 1, 2, 4, 8, 16, 24$ and 48 hours) after administration of the treatment at time $t = 0$. Each of the samples is analysed to find the plasma concentration of the drug in the patient's blood at the sampled time. This information is then plotted to obtain a patient's concentration-time curve, such as that shown in Figure 2.4.1.

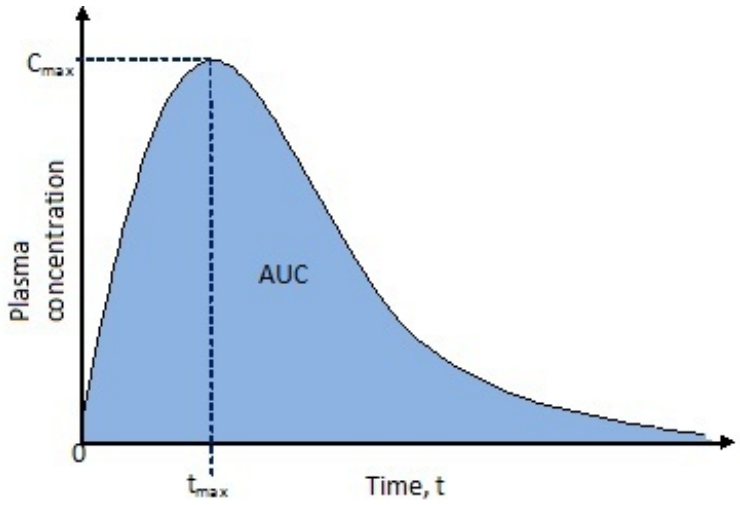

Figure 2.4.1: Example concentration-time plot showing the pharmacokinetic exposure parameters, $C_{\max}$ and AUC.

The PK parameters most likely to be relevant in dose-escalation are two measures of exposure, $C_{\max}$ and AUC_J, which are marked on Figure 2.4.1. These exposure parameters are continuous variables which, by definition, are restricted to being greater than zero. The first is the maximum exposure (after administration) to the treatment, $C_{\max}$, which occurs at time $t_{\max}$. The second is the area under the concentration-time

curve, AUC_J. This is a measure of the total exposure to the trial drug over time interval J. Some common values of J are:

- $0-24$: The AUC between $t=0$ and $t=24$ hours, which can be calculated by fitting a curve to the observed exposures,
- τ: The AUC between $t=0$ and $t=$ time at the end of the cycle, which (assuming that the plasma concentration of the drug at the end of the cycle was obtained) can also be calculated by fitting a curve to the observed exposures,
- ∞: The AUC between $t=0$ and $t=\infty$, which is obtained by extrapolating the curve fitted to the observed exposures.

In a dose-escalation trial of a treatment administered once weekly with a cycle of length 28 days, full PK profiles of patients may be taken on days 1, 8 and 22 of treatment, for example. The PK profile taken at day 1 is used to obtain an idea of the characteristics of the treatment after a single dose. Profiles taken at days 8 and 22 are used to monitor the multiple-dosing characteristics of the treatment. This is to check that multiple-dosing is not leading to excessive build up of the treatment in patients, resulting in excessively high exposure. The final PK profile, at day 22, is hoped to be late enough in the treatment cycle that the system has reached a steady state. This means that upon continued dosing at the same regimen, the exposure pattern is expected to remain constant.

Calculation of exposure parameters by fitting a curve to each patient's data individually (as described above) is common in dose-escalation trials when only a few

patients data are treated at each dose. However, in specific pharmacology trials, population PK models are fitted which use a single model for the data from all patients. Population PK models require less intensive sampling and give an idea of the variability in the parameters between patients. They can also be used to obtain additional inferences about the exposure to the drug. More information about the population PK methods can be found in Källén (2008).

We have chosen to utilise PK data in dose-escalation because it can be an early indicator of efficacy or long-term toxicity (Clark et al., 1994). Although it is unlikely for a dose-escalation trial to be of sufficient duration to observe efficacy or long-term toxicity outcomes directly, PK data obtained during the dose-escalation trial can be used to obtain an idea of the likelihood/risk of these outcomes. This means that the number of patients dosed at unnecessarily high levels can be reduced; based on an assumption that a suitable benefit-risk ratio has already been met at a dose with lower toxicity. The chance of long-term toxicity can also potentially be reduced. Application of the PK data in this way is more pertinent in the combination setting because of the availability of historical single-agent data to advise on the use of exposure data in escalation and the possibility of DDIs.

The intensive sampling routine planned for each patient, to estimate the exposure parameters, mean that it is common to have missing values within a patient's PK profile or for an entire profile to be missing. There can also be delays in the PK data being processed so that the exposure parameters of a cohort are not ready for the dose-escalation meeting to decide the dose for the next cohort of patients. Currently, this does not delay escalation because PK data are not essential for making such

a decision. This situation is not ideal. Hopefully, in proposing a practical dose-escalation design with improved operating characteristics over standard designs, the PK data will be prioritised.

Chapter 3

Dose-escalation Strategies which Utilise Subgroup Information

Abstract

Dose-escalation trials commonly assume a homogeneous trial population to identify a single recommended dose of the experimental treatment for use in future trials. Incorrectly assuming a homogeneous population can lead to a diluted treatment effect in a heterogeneous population. Equally, exclusion of a subgroup that could in fact benefit from the treatment can cause a beneficial treatment effect to be missed. Accounting for a potential subgroup effect (i.e. difference in reaction to the treatment between subgroups) in dose-escalation can increase the chance of finding the treatment to be efficacious in a larger patient population.

The case of two pre-defined subgroups is considered. Biomarker information from historical trials investigating the same treatment in an alternative application, for

example, can be used to identify subgroups of interest.

A standard Bayesian model-based method of dose-escalation is extended to account for a subgroup effect by including covariates for subgroup membership in the dose-toxicity model. A stratified design performs well but uses available data inefficiently and makes no inferences concerning presence of a subgroup effect. A hypothesis test could potentially rectify this problem but the small sample sizes result in a low powered test. As an alternative, we propose the use of spike and slab priors for identifying presence of a subgroup effect. This method assesses the presence of a subgroup effect at each escalation step and at the end of the trial. This enables efficient use of the available trial data throughout escalation and in identifying the recommended dose(s). A simulation study, based on real trial data, was carried out and this design was found to be both promising and feasible.

Keywords: Dose-escalation, subgroup effect, biomarker, Bayesian model-based method, spike and slab.

3.1 Introduction

The aim of a dose-escalation trial is to identify the recommended dose of an experimental treatment to be used in later phase trials investigating the treatment's efficacy. To maximise the treatment's chance of success in efficacy trials, it is important that the recommended dose is optimal for the patient population. Despite this, time restrictions mean that selection of the recommended dose is often based purely on toxicity data which are available relatively soon after treatment. The toxicity data upon which decisions are based is usually a binary indicator of whether a patient

experienced a dose-limiting toxicity (DLT) in their first cycle of treatment.

A common assumption in dose-escalation trials is that toxicity increases monotonically with the dose of the treatment. Since the recommended dose is chosen based only on toxicity data, an implicit assumption is that increasing toxicity leads to increased efficacy of the treatment. Using a Bayesian model-based design for dose escalation, the optimal dose can be referred to as the TD100θ (Whitehead and Williamson, 1998). That is, the dose of treatment with probability θ of causing a dose-limiting toxicity in a patient within their first cycle of treatment. Bayesian model-based designs require a model to be assumed for the dose-toxicity relationship. These designs can utilise available trial data and prior knowledge to advise escalation and estimate the TD100θ.

In standard dose-escalation trial, the trial population is assumed to be homogeneous (Rosenberger and Haines, 2002) and a single TD100θ is identified for the entire population. However, in a general patient population this is unlikely to be the case. Variability between subgroups of patients in a population can lead to differences in tolerance or efficacy of the treatment. Consequently, the benefit-risk ratio of the treatment is impacted for subgroup members. When there is notable variability between subgroups of a population, we refer to the presence of a subgroup effect. Often, the underlying cause of variability is unknown but there can be visible or measurable indicators, referred to as biomarkers, which can be used as intermediate markers of subgroup membership. Examples include ethnicity, pre-treatment or a genetic mutation. For example, presence of a KRAS mutation in patients with non-small cell lung cancer indicates lower survival when treated with Erlotinib and chemotherapy, than

is usual for patients without the mutation (Lièvre et al., 2006).

The limited number of patients available for treatment in dose-escalation trials makes in-trial identification of relevant biomarkers unrealistic. Instead, we consider cases where historical information is used to pre-define potential biomarkers of interest. For example, historical trials of the same treatment in another application, or of a treatment with similar action being tested in the same application, can be used to identify a biomarker of interest.

Currently, historical data on potential subgroup effects is largely utilised in the specification of trial inclusion criteria. These can be used to reduce the variability in the trial population in order to justify an assumption of a homogeneous trial population. In doing this, the population to whom the treatment could be made available is restricted. There is also a risk of excluding patients who could in fact benefit from the treatment. This was the case for Cetuximab, a treatment for colorectal cancer, which was initially tested in a restricted population. It was later noticed that patients excluded from the original trial could in fact benefit from the treatment (Chen and Beckman, 2009). As a consequence, further trials had to be carried out in the additional patient group.

On the other hand, inclusion of a subgroup (in the trial population) in which the treatment is inefficacious could mask a treatment effect in the remaining population. Gefitinib for the treatment of non-small cell lung cancer is an example where this was the case. On further investigation, the subgroup effect was identified and a reduced population who could benefit from Gefitinib found (Chen and Beckman, 2009). In both the Cetuximab and Gefitinib examples, the error was highlighted and adjusted

for. Unfortunately there are potentialy many similar cases for which the error has not been realised. In addition, had a potential subgroup effect been accounted for at the design stage of these trials, then more efficient trials which utilised less resources could have been implemented.

It is becoming more common for potential subgroup effects (aside from in exploratory analyses) to be considered in phase II and III trials. In these so called enrichment trials, subgroup effects are investigated in order to identify a subgroup of the population who appear most likely to benefit from the treatment (see Temple, 2005, for a short overview of such designs). This can lead to exclusion of a subgroup of the patient population from the trial. In such a case, the dose being used in the trial was selected based on patients from the initial population. The recommended dose may therefore be sub-optimal for the final population. In addition, administering different doses of the treatment between subgroups might suffice, removing the need to completely exclude subgroups from the trial. Ideally, through accounting for a potential subgroup effect in dose-escalation, we will estimate a TD100θ in each subgroup when this is necessary. This can increase the chance of finding the treatment to be efficacious in a larger patient population and is a step towards patient-specific dosing.

In Section 3.1.1, a Bayesian model-based method of dose-escalation which is currently used and assumes a homogeneous population is described and the general notation used in the remainder of the chapter is introduced. This continues into a brief review of alternative model-based dose-escalation designs, and in Section 3.1.2 current methods of accounting for a subgroup effect in clinical trials are discussed. The standard dose-escalation trial design described in Section 3.1.1 is used as the under-

lying design for the proposed methods of accounting for a potential subgroup effect in dose-escalation. The proposed methods are presented in Section 3.2 and compared through a simulation study in Section 3.3. The chapter concludes with a discussion of the methods, their limitations and possible extensions in Section 3.4.

3.1.1 A Standard Bayesian Model-based Method of Dose-escalation

Bayesian model-based designs enable available prior and trial information to be utilised in dose-escalation decisions. Using all of this available information in dose-escalation makes escalation decisions more efficient and also safer for patients involved in the trial. The approach of Whitehead and Williamson (1998) is a standard Bayesian model-based method of dose-escalation which assumes a homogeneous trial population; their method is described here. It is used as the baseline for comparison of the proposed methods and also as the design underlying the proposed methods for accounting for a potential subgroup effect in dose-escalation which are described in Sections 3.2.1, 3.2.2 and 3.2.3. Alternative approaches such as the continual reassessment method (O'Quigley et al., 1990) could, however, also be used as the basis for the extensions discussed below.

Dose set $\mathbf{d}$ of the experimental treatment is to be made available for administration to patients in the dose-escalation trial. In reality, escalation using a model-based design is not constrained to this dose set but it is required for the purpose of simulation. Define the dose of treatment administered to a patient as $x \in \mathbf{d}$ and d^* as

some fixed reference dose used to standardise dose in the dose-toxicity model. We are interested in $\pi(x)$, the probability that a patient experiences a DLT given dose x of the experimental treatment. Specifically, the value of x for which $\pi(x) = \theta$. Escalation under the standard design, assuming a homogeneous trial population, proceeds as follows:

1. Model the dose-toxicity relationship in the entire population by:

$$\log\left\{\frac{\pi(x)}{1-\pi(x)}\right\} = \beta_0 + \beta_1 \log\left(\frac{x}{d^*} + 1\right) \text{ where } \pi(x) = \mathbb{P}(\text{DLT}|x). \qquad (3.1.1)$$

   We consider the transformed, standardised dose, $\log(x/d^* + 1)$ in the assumed dose-toxicity model but an alternative one-to-one transformation could be used. Choice of the reference dose and transformation is given in Section 2.2.2 under Step 3;

2. Set a prior on the model parameters: This is achieved by specifying pseudo data relating to a prior proportion of DLTs occuring at two 'prior' doses. This prior data is weighted to total a fraction of the planned sample size of the trial. A value of $1/10^{\text{th}}$, as used by Whitehead and Williamson (1998), is used in this chapter; further discussion about the choice of weight is given in Section 2.2.2 under Step 4. By incorporating the pseudo data into the dose-toxicity model in the same way as trial data, beta priors are effectively induced on the probability of toxicity at the two doses (Tsutakawa, 1975). The prior proportion of DLTs at the two doses can be elicited from clinical experts (as described in Whitehead and Williamson, 1998, for example). Alternatively, the prior can be selected to

control the operating characteristics of dose-escalation. For example, specifying:

- The desired start dose for the trial as the lower of the two doses selected for prior specification with a prior proportion of DLTs equal to θ;
- A dose at the top of the planned dose range for the other prior dose with a prior proportion of DLTs selected to control the rate of escalation under some likely trial scenarios.

3. Allocate patients the dose $x \in \mathbf{d}$ which, based on the prior and available trial data at their time of arrival into the trial:

   - Maximises the patient gain, $\frac{1}{\{\hat{\pi}(x)-\theta\}^2}$;
   - Within doses which satisfy the safety criterion, $\hat{\pi}(x) < \delta$,

   for an unacceptable level of toxicity δ and $\hat{\pi}(x) = 1/[1 + \mathrm{e}^{-\{\hat{\beta}_0+\hat{\beta}_1 \log(x/d^*+1)\}}]$ where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the modal a posteri (MAP) estimates of the model parameters. When prior knowledge is incorporated into the dose-toxicity model as pseudo data, the MAP estimates are equivalent to the maximum likelihood estimates of the parameters so that standard software can be used without the need for Markov Chain Monte Carlo.

4. Stop escalation:

   - For safety if at any point in the trial no available doses satisfy the safety criterion: No recommended dose is declared;
   - Once a maximum number of patients have been treated in the trial: The recommended dose is declared as the estimated TD100θ for the entire

population based on data collected in the trial (i.e. not including prior data). That is, the dose which maximises the patient gain and satisfies the safety criterion (based on the two-parameter dose-toxicity model of Equation 3.1.1), from the range of available doses which are less than or equal to the maximum dose administered during the trial.

Other authors, such as Neuenschwander et al. (2008), have assumed the same two-parameter dose-toxicity model for dose-escalation. Their approach differs in specification of escalation rules for the trial (Step 3). Whitehead and Williamson (1998) themselves suggest alternatives to those described here but we have chosen to use the patient gain as the most ethical option. Addition of the safety constraints in a similar manner to Babb et al. (1998) control the rate of esclation, improving the safety of the trial for the patients involved.

Alternative dose-toxicity models have been suggested; the continual reassessment method (CRM) of O'Quigley et al. (1990) uses a one-parameter power model which accurately estimates the TD100θ but does not effectively model the entire dose-toxicity relationship. Goodman et al. (1995), among others, have proposed modifications on the CRM to reduce the aggressiveness of escalation. A two-parameter model is more suitable than a one-parameter model for comparison of the dose-toxicity relationship between subgroups, as we are interested in. This is because, although the subgroup effect may not lead to different recommended doses in each subgroups, the shape of the dose-toxicity curves between subgroups may differ. This could indicate different reactions to the treatment across the dose range which may be pronounced in the

efficacy or longer-term toxicity outcomes which will be investigated in later trials.

Other Bayesian model-based designs have been proposed which aim to optimise escalation, although these are often considered unethical as they do not account for the needs of patients (Dette et al., 2008; Haines et al., 2003). Reviews of dose-toxicity models and available methods of dose-escalation are provided in Rosenberger and Haines (2002) and Jaki et al. (2013). Most Bayesian model-based dose-escalation trial designs have the same foundations and so the methods presented in this chapter could be altered for the use of an alternative dose-toxicity model or escalation rules.

3.1.2 Current Methods of Accounting for Subgroup Information in Clinical Trials

The most straight-forward way to account for a subgroup effect in dose-escalation is to stratify by subgroup membership and carry out independent dose-escalation in each subgroup. This has been done in practice (e.g. Nicholson et al., 1998) but is inefficient (in its use of information for identifying a dose for escalation and estimating the TD100θ), especially if there is in fact no underlying subgroup effect. Wijesinha and Piantadosi (1995) and O'Quigley et al. (1999) propose using additional terms in the dose-esclation model to account for subgroup membership. In this way, some information is shared between subgroups during escalation. Babb and Rogatko (2001) use a similar method but consider a continuous biomarker; their design is demonstrated in Cheng et al. (2004).

In current practice, it is more common for a subgroup effect to be investigated in later phase trials. Such designs use hypothesis testing at an interim point in the trial to identify subgroup(s) of the population that react favourably to treatment and hence it is felt worth pursuing the experimental treatment in (Brannath et al., 2009; Chen and Beckman, 2009; Jenkins et al., 2011).

3.2 Proposed Methods of Accounting for Subgroup Information in Dose-escalation

When the trial population is truly homogeneous, then a standard method of dose-escalation (such as that of Whitehead and Williamson, 1998, which was described in Section 3.1.2), which does not account for a potential subgroup effect, will be suitable. However, when there is uncertainty around the assumption of a homogeneous population, then this design is not appropriate. We compare the standard design to three alternative methods of dose-escalation which account for subgroup membership throughout escalation.

Say that patients entering the trial can be reliably classified as being in one of two distinct, clearly identifiable subgroups based on the presence or absence of a pre-defined biomarker. The treatment is expected to be more toxic in the biomarker positive patients than in the remaining biomarker negative patients. Let $\mathbb{I}_+$ be an indicator of subgroup membership which is equal to 1 for a biomarker positive patient and 0 for a biomarker negative patient.

Some of the benefits of Bayesian dose-escalation designs have discussed in Sections 2.2 and 3.1.1. The main reason being that using all available information in dose-escalation leads to more informed escalation decisions which should reduce the risk and increase the benefit of treatment for trial patients. As we have mentioned, the use of prior data in Bayesian trial designs can be intuitive and beneficial. However, it also makes the escalation decisions subjective. To remove the subjectivity in conclusions drawn from the trial, we have chosen not to use the prior data in selecting the recommended dose(s) for use in future trials. This is therefore found in a frequentist manner, by fitting the trial data to the logistic regression model. In addition, since the prior data used in demonstration of the methods given in this chapter is been specified to control the operating characteristics of the trial, it does not seem appropriate for the prior data to be accounted for when drawing conclusions from the trial.

3.2.1 Method 1: Include Terms for Subgroup Membership

In this method, the standard two-parameter dose-toxicity model from Equation 3.1.1 is extended to include terms for subgroup membership. This enables escalation decisions to be made which account for subgroup membership. Hence, making the dose administered to patients better suited to them. A consequence of allowing esclation to differ between subgroups is that the safety stopping criterion can come into play for one or both subgroups. Escalation under this method proceeds as follows:

1. Model the dose-toxicity relationship using the four-parameter logistic model:

$$\log\left\{\frac{\pi(x,\mathbb{I}_+)}{1-\pi(x,\mathbb{I}_+)}\right\} = \beta_0 + \beta_1 \log\left(\frac{x}{d^*}+1\right) + \mathbb{I}_+\left\{\beta_2 + \beta_3 \log\left(\frac{x}{d^*}+1\right)\right\}, \tag{3.2.1}$$

where $\pi(x,\mathbb{I}_+) = \mathbb{P}(\text{DLT}|x,\mathbb{I}_+)$.

If historical evidence of a subgroup effect led to strong belief of its impact on either the intercept or slope parameter of the dose-toxicity model, then one of the additional terms could be removed and the resulting three-parameter model used in place of the four-parameter model. However, with a lack of information on the expected impact of the subgroup effect on the dose-toxicity relationship, the four-parameter dose-toxicity model is able to capture potential variability in both parameters;

2. Set a prior on the model parameters: This can be achieved in a similar manner to that for the standard design by specifying pseudo data on two doses for the biomarker positive subgroup and two doses for the biomarker negative subgroup. The pseudo data for each subgroup is weighted to, say $1/10^{\text{th}}$, of the planned sample size in that subgroup.

3. Allocate patients the dose $x \in \mathbf{d}$ which, based on their subgroup membership, the prior and available trial data at their time of arrival into the trial:

   - Maximises the patient gain, $\frac{1}{\{\hat{\pi}(x,\mathbb{I}_+)-\theta\}^2}$;
   - Within doses which satisfy the safety criterion, $\hat{\pi}(x,\mathbb{I}_+) < \delta$,

for unacceptable level of toxicity δ and for MAP estimates of the model parameters $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$, $\hat{\pi}(x, \mathbb{I}_+) = 1/(1 + \mathrm{e}^{-[\hat{\beta}_0 + \hat{\beta}_1 \log(x/d^* + 1) + \mathbb{I}_+ \{\hat{\beta}_2 + \hat{\beta}_3 \log(x/d^* + 1)\}]})$.

4. Stop escalation:

   - For safety in a subgroup if at any point in the trial no available doses satisfy the safety criterion for that subgroup: No recommended dose is declared in that subgroup. Escalation continues in the other subgroup using the two-parameter model of Equation 3.1.1 fitted to data from patients in the remaining subgroup only;

   - Once a maximum number of patients have been treated in the trial:

     – If one subgroup stopped for safety: The recommended dose is declared in the remaining subgroup as the estimated TD100θ based on data collected in the trial (i.e. not including prior data). That is, the dose which maximises the patient gain and satisfies the safety criterion (based on the two-parameter dose-toxicity model of Equation 3.1.1 fitted to the data from patients in that subgroup only), from the range of available doses which are less than or equal to the maximum dose administered to patients in the respective subgroup during the trial;

     – If neither subgroup stopped for safety: A recommended dose is declared in each subgroup as the estimated TD100θ_s for $s = +, -$, representing that in the biomarker positive and negative subgroups respectively, based on data collected in the trial (i.e. not including prior data). That is, the dose which maximises the patient gain and satisfies

the safety criterion (based on the four-parameter dose-toxicity model of Equation 3.2.1), from the range of available doses which are less than or equal to the maximum dose administered to patients in the respective subgroup during the trial.

By including covariates for subgroup membership in the dose-toxicity model, this method of dose-escalation enables recommended doses to be subgroup specific. A TD100θ is estimated in each subgroup (unless one or both subgroups stop for safety). When these recommendations are different between subgroups, then we expect that a significant subgroup effect has been observed. When the recommendations are the same between subgroups, this could be down to there truly being no significant subgroup effect. On the other hand, it could be a result of the discrete dose set or insufficient size of the trial to detect a difference. Under this method we have no way of telling this, and indeed deciding whether it would be beneficial to pool the data or if it would be beneficial to continue investigation of the subgorup effect which become clear when longer-term toxicity or efficacy outcomes are investigated.

The next method incorporates a formal test of whether a significant subgroup effect was observed in an attempt to clarify and formalise the conclusion drawn from the dose-escalation trials over the presence of a subgroup effect. In this case, the result of the hypothesis test is interpreted as a decision over whether a subgroup effect was observed. If it is concluded that a subgroup effect is present and the dose recommendations from the two subgroups are still the same, then this is likely to be caused by the use of discrete dose set in the trial.

3.2.2 Method 2: Hypothesis Test Concerning Presence of a Subgroup Effect

This method forms an extension of Method 1. The escalation and stopping procedure is unchanged, the only difference comes in selecting the recommended dose(s) when neither subgroup stopped for safety during the trial. Under this eventuality, instead of automatically recommending a dose in each subgroup, a hypothesis test is performed in an attempt to determine whether a subgroup effect was observed in the trial. As mentioned in Section 3.2, to reduce the subjectivity in the selection of the recommended dose(s), a frequentist calculation of the recommended dose is used. Following from this logic, a likelihood based hypothesis test is considered in this method. If the test concludes that:

- No significant subgroup effect was observed: The data are pooled and a single recommended dose is declared for the entire population as the estimated TD100θ based on data collected in the trial (i.e. not including prior data). That is, the dose which maximises the patient gain and satisfies the safety criterion (based on the two-parameter dose-toxicity model of Equation 3.1.1), from the range of available doses which are less than or equal to the maximum dose administered during the trial;

- A significant subgroup effect was observed: As before, a recommended dose is declared in each subgroup as the estimated TD100θ_s for $s = +, -$ based on data collected in the trial (i.e. not including prior data). That is, the dose which maximises the patient gain and satisfies the safety criterion (based on

the four-parameter dose-toxicity model of Equation 3.2.1), from the range of available doses which are less than or equal to the maximum dose administered to patients in the respective subgroup during the trial.

In this method, it is still possible to obtain the same dose recommendation from both subgroups despite a decision that a subgroup effect is present. However, this does provide a formal test of whether a subgroup effect was observed. This information can be useful in planning future trials. When no subgroup effect is detected, then the recommended dose is found with more accuracy under this method than in Method 1, since the data are pooled and fitted to the two-parameter dose-toxicity model.

A range of hypothesis tests are possible. We consider a z-test on the difference between recommended doses from the two subgroups. The test is based on the asymptotic approximation, $\widehat{\text{TD}}100\theta \sim N(\mathbb{E}[\text{TD}100\theta], \text{Var}[\text{TD}100\theta])$ and for estimate of the recommended dose $\widehat{\text{TD}}100\theta_s$ in subgroup s, the null (H_0) and alternative (H_1) hypotheses are;

$$H_0 : \widehat{\text{TD}}100\theta_- - \widehat{\text{TD}}100\theta_+ = 0 \quad \text{versus} \quad H_1 : \widehat{\text{TD}}100\theta_- - \widehat{\text{TD}}100\theta_+ \neq 0.$$

We fix the significance level of the (two-sided) test, $\mathbb{P}(\text{reject } H_0 | H_0 \text{ true}) = \alpha$. The power of the test, $\mathbb{P}(\text{reject } H_0 | H_0 \text{ not true}) = 1 - \beta$, then depends on the specified α and the number of patients in the trial.

Although this test uses asymptotic results, it was selected because it avoids the need for corrections for multiple testing and the use of equivalence hypotheses, both of which further lower the power of an already low-powered test. Further details of

this hypothesis test, along with details concerning its choice over some alternative hypotheses, are given in Appendix 3.5.1. As an alternative method which does not encounter this problem, we propose a fully Bayesian approach using spike and slab priors for variable selection.

3.2.3 Method 3: Fully Bayesian Method Using Spike and Slab Priors for Variable Selection

This method is based on the four-parameter dose-toxicity model given in Equation 3.2.1. In Method 2, the four-parameter dose-toxicity model was used throughout escalation and a decision as to whether the two-parameter model (which does not account for subgroup membership) is sufficient only made at the end of the trial. So during escalation, Method 2 assumed a subgroup effect was present; it did not allow for the fact that a subgorup effect may not be present until the trial analysis. In addition, the hypothesis test described for Method 2 was based on the difference in dose recommendations in the two subgroups. Hence, only providing information concerning whether the subgroup effect affected the point estimate of the TD100θ and not on the entire dose-toxicity curve. Alternative frequentist hypothesis tests which can achieve this were investigated but were found to be too low-powered to be practical; details can be found in Appendix 3.5.1.

The Bayesian alternative that we propose overcomes these problems to some extent by using spike and slab priors on the model terms for subgroup membership (β_2 and β_3 in Equation 3.2.1). This allows more efficient use of emerging trial data during

escalation by allowing for presence or absence of a subgroup effect throughout the trial. This is achieved by deciding at each escalation step, based on data available at that time, whether the two-parameter or four-parameter model is more suitable.

A spike and slab prior is effectively a two-component mixture prior. One component is usually a normal prior with high variance which makes up the 'slab' part of the prior. The other part is the 'spike' component which is selected as a distribution which has a large mass at zero. We choose to use a Dirac delta function, δ_0 (a point mass at zero), which results in a sparsity inducing spike and slab mixture prior. Figure 3.2.1 gives an example of a potential mixture prior on β composed of a normal slab and Dirac delta function spike. The result of using these priors is that a positive probability is placed on the probability of the term being equal to zero. Based upon this, spike and slab priors can be used in variable selection.

Now, take γ_2 to be a latent indicator function which indicates inclusion (when equal to 1, and is zero otherwise) of the variable β_2 in the dose-toxicity model. Then the resulting spike and slab prior on β_2 can be written as:

$$\beta_2|\gamma_2 \sim \gamma_2 \mathrm{N}(0, \sigma_2^2) + (1 - \gamma_2)\delta_0.$$

The decision over whether β_2 is required in the model, based on available data, can be based on its probability of inclusion in the model, w_2. This can be estimated by placing a Bernoulli prior on γ_2 such that;

$$\gamma_2 \sim w_2^{\gamma_2}(1 - w_2)^{(1-\gamma_2)}.$$

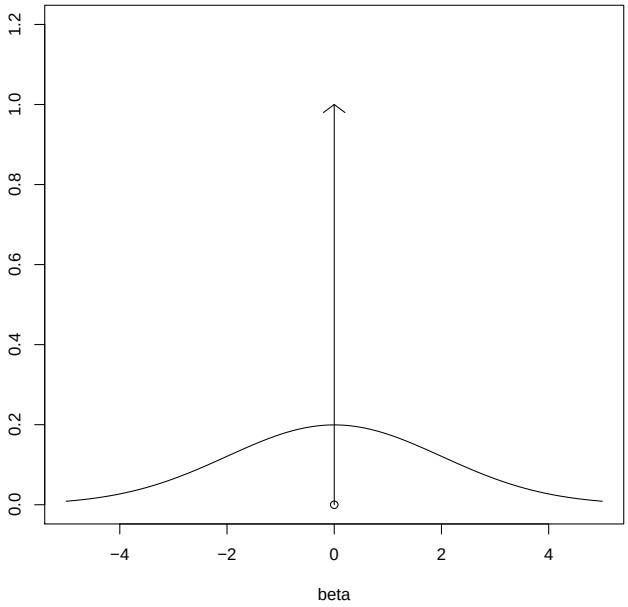

Figure 3.2.1: Example of a mixture prior on β composed of a normal slab and Dirac delta function spike.

Similarly we can consider a latent indicator function γ_3 and probability of inclusion w_3 on β_3. We assume that w_2 is independent of w_3 and as such, a prior setting of $w_2 = w_3 = 0.5$ implies a prior belief that one of the two predictors for subgroup effect are significant in the model (see Chapter 10 of Do et al., 2013). If instead, w_2 or w_3 is set equal to 1, then the corresponding term will be forced into the model with a normal prior (the slab component of the prior corresponding to that term) placed on it. This is effectively what is done for β_0 and β_1 which we require in the dose-toxicity model. We assume independence of w_2 and w_3 to increase the chance, over an alternative which assumes some dependence, that either β_2 or β_3 is selected in the model. This makes the model more flexible by allowing the model to capture heterogeneity in the form of a shift, slope difference or a combination of these which is useful since we have no information on the expected cause of the subgorup effect.

With prior information regarding this, dependence between these parameters could be defined.

A range of algorithms exist for implementing Bayesian variable selection using spike and slab priors in the linear regression setting (e.g. George and McCulloch, 1997; Ishwaran and Rao, 2005; Scheipl, 2011). Authors such as Wagner and Duller (2012) and Tüchler (2008) have extended these methods to the logistic regression setting. The applications of Bayesian variable selection for logistic regression models is wide-ranging; Wagner and Duller (2012) aim to identify relevant risk factors for bleeding while Genkin et al. (2005) is concerned with text categorisation. Methods which deal with multivariate regression and ANOVA are also available (e.g. Carvalho et al., 2008) which have application in selection of variables relating to gene expression.

We specified a dirac delta function for the 'spike' component of the prior on the terms for subgroup membership. Alternative choices include use of a normal distribution with large mass at zero and a double exponential model (or Lasso model, see Tibshirani, 1996, for details). Although a mixture of normal distributions results in a continuous prior, it is one which is not sparsity inducing. As a result, a straightforward decision concerning whether a term should be included in the model cannot be made. Bernardo et al. (2011) compare a range of prior settings, including those mentioned, and obtain no clear conclusion over the 'better' sparsity inducing prior.

A method related to Bayesian variable selection is Bayesian model averaging Hoeting et al. (1999). Although such methods would be feasible with the small number of parameters in our model, we wish to obtain a clear decision over whether the terms for subgroup membership should be included in the model. For this reason, we choose

to use variable selection.

When spike and slab priors are used, we have a form of in-built decision making process over whether these additional terms are required in the model. Once the relevant variables have been identified, the selected model is fitted to the data and escalation decisions can be made based upon this. Escalation decisions now occurs in two stages; variable selection and model fitting.

Escalation under this method follows the standard method described in Section 3.1.1 until a difference in outcomes between the pre-defined subgroups has been observed. Note that, if the specified pseudo data relates to a prior subgroup effect, then spike and slab will be implemented from the start of the trial. Once a difference in outcomes between subgroups is observed, escalation under this method proceeds as follows;

1. Model the dose-toxicity relationship using the four-parameter logistic model:

$$\begin{aligned} \log\left\{\frac{\pi(x,\mathbb{I}_+)}{1-\pi(x,\mathbb{I}_+)}\right\} &= \beta_0+\beta_1\log\left(\frac{x}{d^*}+1\right)+\mathbb{I}_+\left\{\beta_2+\beta_3\log\left(\frac{x}{d^*}+1\right)\right\} \\ \text{where } \pi(x,\mathbb{I}_+) &= \mathbb{P}(\text{DLT}|x,\mathbb{I}_+). \end{aligned}$$

   The terms β_0 and β_1 will always be included in the model used for escalation. However, spike and slab priors are specified on β_2 and β_3 and so one or both of these terms could be set equal to zero in the model for escalation.

2. Set a prior on the model parameters: Pseudo-data of the same form used in Method 1 (and 2) is used to define priors; for the variable selection and model

fitting stages in escalation.

Variable selection: Fit the pseudo data to the four-parameter logistic regression model of Equation 3.2.1. The resulting coefficient estimates are used to derive the slab component of the priors on the four parameters of the dose-toxicity model. A method of deriving the prior for variable selection is given in Appendix 3.5.2 with regard to the prior used in the simulation study (presented in Section 3.3).

Model fitting: This can be achieved in the same way as for Method 1 (and 2).

3. Escalation follows the two-step process:

Variable selection: Fit the spike and slab model using Markov Chain Monte Carlo (MCMC). After removing burn-in iterations, find w_2 and w_3 (the probability that each term was included in the dose-toxicity model which is always 1 for β_0 and β_1 but varies for β_2 and β_3). If the inclusion probability of the parameter is greater than some pre-specified boundary, then that term will be non-zero in the fitted model. Otherwise it is equal to zero for this model fit.

Model fitting: Allocate patients the dose $x \in \mathbf{d}$ which, based on their subgroup membership (if relevant), the prior and available trial data at their time of arrival into the trial:

- Maximises the patient gain, $\frac{1}{\{\hat{\pi}(x,\mathbb{I}_+)-\theta\}^2}$,
- Within doses which satisfy the safety criterion, $\hat{\pi}(x,\mathbb{I}_+) < \delta$,

for unacceptable level of toxicity δ and $\hat{\pi}(x, \mathbb{I}_+) = 1/[1+\mathrm{e}^{-\{\hat{\beta}_0+\hat{\beta}_1 \log(x/d^*+1)+y\}}]$ where y is the terms for subgroup membership identified during variable selection for inclusion in the model. The estimates $\hat{\beta}_0$, $\hat{\beta}_1$, and potentially $\hat{\beta}_2$ and/or $\hat{\beta}_3$, are the MAP estimates of the dose-toxicity model parameters.

4. Stop escalation:

   - For safety in a subgroup if at any point in the trial no available doses satisfy the safety criterion for that subgroup: No recommended dose is declared in that subgroup. Escalation continues in the other subgroup using the two-parameter dose-toxicity model of Equation 3.1.1 fitted to data from patients in that subgroup only.
   - Once a maximum number of patients have been treated in the trial:
     - If one subgroup stopped for safety: The recommended dose is declared in the remaining subgroup as the estimated TD100θ based on data collected in the trial (i.e. not including prior data). That is, the dose which maximises the patient gain and satisfies the safety criterion (based on the two-parameter dose-toxicity model of Equation 3.1.1 fitted to the data from patients in that subgroup only), from the range of available doses which are less than or equal to the maximum dose administered to patients in the respective subgroup during the trial.
     - If neither subgroups stopped for safety: Carry out variable selection,
       * If β_2 and β_3 are equal to zero: The data are pooled and a single recommended dose is declared for the entire population as the

estimated TD100θ based on data collected in the trial (i.e. not including prior data). That is, the dose which maximises the patient gain and satisfies the safety criterion (based on the two-parameter dose-toxicity model of Equation 3.1.1), from the range of available doses which are less than or equal to the maximum dose administered during the trial.

* If β_2 and/or β_3 is non-zero: As in Method 1, a recommended dose is declared in each subgroup as the estimated TD100θ_s for $s = +, -$ based on data collected in the trial (i.e. not including prior data). That is, the dose which maximises the patient gain and satisfies the safety criterion (based on the four-parameter dose-toxicity model of Equation 3.2.1), from the range of available doses which are less than or equal to the maximum dose administered to patients in the respective subgroup during the trial.

The overall set-up of this method is relatively similar to the previous methods. However, before model fitting can occur in Step 3, variable selection must be carried out (and a relevant prior specified). The use of spike and slab priors mean that the model used in variable selection is not conjugate and so MCMC is required, making Method 3 more computationally complex than the previous methods.

The use of spike and slab priors on the terms for subgroup membership enable escalation decisions to be founded on the most relevant model, based on data available at that stage of the trial. This will make escalation more efficient and be beneficial

for patients. In addition, by considering whether each variable should be included in the model, the entire dose-toxicity curve is compared between groups as opposed to merely a point estimate of dose recommendation.

There is no formal test of whether a subgroup effect was observed in this method and so the decision over the presence or absence of a subgroup effect is exploratory. These exploratory conclusions, together with historical information and clinical expertise on the expected subgroup effect, may be suitable to decide whether a subgroup effect should be accounted for in later phase trials. Alternatively, a hypothesis test could be carried out on the final trial data with no adverse effect on escalation, although this has the aforementioned issues.

3.3 Simulation Study

Data from the single-agent paediatric dose-escalation trial reported by Nicholson et al. (1998) were used as the basis for the simulation study presented in this section. In the reported trial, Nicholson et al. (1998) used stratification to account for a potential subgroup effect and escalation proceeded in each subgroup under an 'up and down' design (see Storer, 1989, for an example of such a design). In this trial, biomarker positive patients had experienced a specific line of prior treatment which the biomarker negative patients had not. The decision to stratify by this prior treatment came from evidence obtained in adult trials of the treatment.

The data obtained in the trial is given in Table 3.3.1, both by subgroup membership and as the pooled data. Based upon the algorithmic design and definition of

the recommended dose specified in Nicholson et al. (1998), the maximum tolerated doses were identified as 215 and 180mg/m^2 in the biomarker negative and biomarker positive subgroups, respectively. Now, had the two-parameter dose-toxicity model in Equation 3.1.1 been fitted to these data, the results are likely to have been different. For example, under a model-based approach. the TD16 in the biomarker positive subgroup is very similar to the maximum tolerated dose identified under the algorithmic design at 181mg/m^2. However, in the biomarker negative subgroup, the TD16 is 244mg/m^2 under the model-based approach. It is the TD16 that we aim to identify in the simulation study in the remainder of this section.

	Number of DLTs observed by dose (mg/m^2)							Recommended dose (mg/m^2) based on	
	100	150	180	215	245	260	Total	algorithmic design	model-fit to data
$\mathbb{I}_+ = 0$ subgroup	0/5	0/4	0/4	0/6	2/7	1/1	3/27	215	244
$\mathbb{I}_+ = 1$ subgroup	1/6	0/4	0/8	2/4	-	-	3/22	180	181
Pooled data	1/11	0/8	0/12	2/10	2/7	1/1	6/49	-	206

Table 3.3.1: Toxicity data observed in the dose-escalation trial reported in Nicholson et al. (1998), given by subgroup membership and as the pooled data. Also given is the recommended dose declared from the trial; as a maximum tolerated dose based on escalation by an algorithmic design in each subgroup, and the TD16 (given a continuous range of doses) based on fitting the dose-toxicity model in Equation 3.1.1 to the data.

3.3.1 Simulation Study Design

The simulation study here is presented to illustrate the proposed dose-escalation methods described in Section 3.2. We compare them to the baseline method; the standard Bayesian model-based method of dose-escalation which was presented in Section 3.1.1.

We specify the dose set available for the trial as those used by Nicholson et al. (1998), $d = \{100, 150, 180, 215,\ 245, 260\}$mg/m^2. The recommended dose from adult

trials was 200mg/m^2, this dose is selected as the reference dose which is used to standardise doses in the dose-toxicity model. The starting dose for the trial was taken to be the lowest available dose of 100mg/m^2 and we specify $\theta = 0.16$ and set the unacceptable probability of toxicity, for use in the safety criterion, as $\delta = 0.35$. So, we aim to identify the dose, from those available which are less than the maximum administered in the trial, which has posterior probability of causing a DLT in a patient closest to 0.16 with estimated probability of toxicity less that 0.35.

We consider that upon entry to the trial, patients were reliably identified as being either biomarker positive or biomarker negative. Patients were recruited in cohorts of size 2 throughout the trial. Each cohort consists of one biomarker positive and one biomarker negative patient unless one subgroup has stopped escalation early, in which case both patients in the cohort will be from the remaining subgroup. The maximum number of patients to be treated in the trial is 60. If neither subgroup stops escalation early, then this will be made up of 30 patients from each subgroup. In the case of the baseline method, escalation will continue until 60 patients have been treated in the trial unless the trial stops early for safety. Although this might not be realistic, it is used here to enable us to compare the methods with a fixed amount of information.

The prior was specified such that it is worth $1/10^{\text{th}}$ of the planned sample size. That is, a total of 6 prior patients consisting of 3 on each subgroup. We specified the same prior data in both subgroups, this is done here to aid comparability of the methods but could of course be altered for use in a real trial. After running a range of potential pseudo-data specifications (details of these are given in Appendix 3.5.3) the prior data we used in the simulation study was selected as that in Table 3.3.2. Under

this prior specification, the dose-toxicity model advises a start dose of 100mg/m^2 (i.e. fitting only the pseudo data to the dose-toxicity model, the escalation rule advises a dose of 100mg/m^2 for escalation). In addition, under the scenario of no observed DLTs, the chosen prior leads to reasonable paced escalation with no skipped doses. Upon observation of a DLT at a low dose, it was felt likely for the model to re-escalate within the specified maximum trial size. Clearly these properties differ between the baseline approach and an approach which considers potential subgroup effect. For comparability between methods, our chosen prior is acceptable under both settings.

	Number of DLTs observed by dose (mg/m^2)	
	100	260
$\mathbb{I}_+ = 0$ subgroup	(1/3)/2	(1/2)/1
$\mathbb{I}_+ = 1$ subgroup	(1/3)/2	(1/2)/1
Pooled data	(2/3)/4	1/2

Table 3.3.2: Prior pseudo-data used for the simulation study, given by subgroup membership and the pooled data.

In the simulation study, toxicity data were generated from the four-parameter dose-toxicity model given in Equation 3.2.1. The parameter values of β_0 and β_1 used for data generation were the mean estimates obtained from a frequentist model fit to Equation 3.1.1 using the pooled trial data (given in Table 3.3.1). The parameter values for β_2 and β_3 were varied depending upon the simulated scenario. A 'true' probability of DLT refers to the probability of DLT based upon the dose-toxicity model and parameter values from which data were simulated. Similarly, a 'true' recommended dose refers to the dose, from the discrete set available for the trial, which has estimated probability of causing a DLT in a patient closest to the TD16

(from those estimates less than 0.35) based upon the model and parameter values from which data were simulated.

The type I error level for the hypothesis test in dose-escalation Method 2 was set at 0.30. This is likely to be too high to be accepted in practice but lower error levels were investigated but traditional type I error specifications of 0.05 and 0.10 turned out to be extremely low powered and, hence, not worth presenting here. Details of the power calculation which support this statement are given in Appendix 3.5.1.

Simulations for all methods were carried out using *R* (R Core Team, 2014). Method 3 required the addition of a variable selection step in the escalation procedure compared to the other methods. This step was carried out using the `BoomSpikeSlab` package (Scott, 2014) which is based on variable selection for logistic regression models as described by Tüchler (2008). Given that we have no outside information to suggest otherwise, the default settings were used for most parameters required by the functions called from `BoomSpikeSlab`. Details of these parameters are given in Appendix 3.5.2 along with details of prior specification for the variable selection steps. Running the Markov Chain for $20,000$ iterations and removing $5,000$ as burn-in was found to be suitable for convergence. We set the prior inclusion probability for β_2 and β_3 equal to 0.5; this is a relatively non-informative setting. We specify that a parameter is non-zero in the fitted model if it has posterior probability of inclusion in the model greater than 0.25. This relatively low value was chosen through simulation studies as a value which led to a reasonable chance of the parameters being included in the model, despite the small amount of data available. Investigations into the choice of the probability for inclusion of the terms in the model, along with ones

investigating the prior inclusion probability, are given in Appendix 3.5.4.

Results are presented for the following six scenarios based on estimates from $1,000$ simulated trials under the given scenario and method. The true probabilities of toxicity at each available dose for each of the scenarios are given in Table 3.3.3 and plots of the dose-toxicity curves they are generated from are given in Appendix 3.5.5:

1. No subgroup effect: This scenario is included for comparison of the methods when the 'true' recommended dose is the same for both subgroups. This could arise when the population is truly homogeneous, or when the biomarker considered in the trial is not the cause of the subgroup effect observed in the trial.

2. A small subgroup effect: Causing only one dose level difference in true recommended doses between subgroups. This scenario is included to investigate the sensitivity of the methods to small differences in tolerance to the treatment between the subgroups.

3. A medium subgroup effect: Causing two dose level difference in true recommended doses between subgroups. This scenario, and the next, is included to investigate the sensitivity of the methods to varying degrees of subgroup effect.

4. A medium subgroup effect: Causing three dose level difference in true recommended doses between subgroups.

5. A large subgroup effect: No safe dose in the biomarker positive subgroup and a true recommended dose in the biomarker positive subgroup in the middle of the available dose range.

6. No safe dose in either subgroup: This scenario is included to demonstrate the effectiveness of the safety criterion when there are no safe doses in either subgroup.

	$P(DLT\|d, I_+ = 0)$						$P(DLT\|d, I_+ = 1)$					
Scenario	100	150	180	215	245	260	100	150	180	215	245	260
1	0.02	0.06	0.10	0.18^X	0.28	0.33	0.02	0.06	0.10	0.18^X	0.28	0.33
2	0.02	0.06	0.10	0.18^X	0.28	0.33	0.02	0.08	0.14^X	0.26	0.38	0.45
3	0.02	0.06	0.10	0.18^X	0.28	0.33	0.03	0.13^X	0.24	0.42	0.58	0.65
4	0.02	0.06	0.10	0.18^X	0.28	0.33	0.09^X	0.36	0.60	0.81	0.90	0.93
5	0.02	0.06	0.10	0.18^X	0.28	0.33	0.42	0.90	0.97	0.99	1.00	1.00
6	0.38	0.67	0.79	0.88	0.93	0.94	0.38	0.67	0.79	0.88	0.93	0.94

Table 3.3.3: Simulated probability of DLT at each dose (in mg/m^2) to be tested in simulations, given for each subgroup. Grey cells highlight dose-pairs with probability of causing a DLT in a patient greater than 0.35. The 'X' marks the dose with probability of toxicity closest to 0.16, in cases where there is a tolerated dose.

3.3.2 Simulation Study Results

The standard Bayesian model-based dose-escalation trial design described in Section 3.1.1 (based on the assumption of a homogeneous trial population) is used as the baseline method for comparison of the three proposed dose-escalation methods described in Section 3.2, which account for a potential subgroup effect. When recommended dose(s) are referred to, these are the frequentist estimates; they are obtained by fitting the relevant logistic regression model to the trial data only i.e. not including prior data. The prior that we used for the simulation study was selected to control the operating characteristics of the trial; it was not based on real trial data. For this reason, it is not appropriate for the prior data to affect the final outcome of the trial. If, however, the prior was selected based on historical data, then it may be desirable to consider this data in identifying the recommended dose(s) from the trial.

Even in this setting, a frequentist estimate might be used to reduce the subjectivity of decisions made from the dose-escalation trial that could impact on future trials of the treatment.

From Table 3.3.4 we can see that in Scenarios 1-4, where there was a tolerated dose available for each subgroup, most trials ran to the maximum number of patients with less than 10% of trials stopping early for safety in one subgroup. In these scenarios, the average proportion of toxicities observed overall was between 12 and 16%. Although the average proportion of toxicities observed was fairly consistent across scenarios in the biomarker negative subgroup (under Methods 1-3), that in the biomarker positive subgroup increased as the true subgroup effect increased. This is in part due to the higher toxicity levels of all available doses.

Scenario	Escalation method	Average number patients			Average proportion toxicities		
		Overall	$\mathbb{I}_+ = 0$	$\mathbb{I}_+ = 1$	Overall	$\mathbb{I}_+ = 0$	$\mathbb{I}_+ = 1$
1	Baseline	59.94	29.97	29.97	0.12	0.12	0.12
	1 and 2	58.59	29.45	29.14	0.12	0.14	0.15
	3	58.97	29.49	29.48	0.12	0.14	0.13
2	Baseline	60.00	30.00	30.00	0.12	0.10	0.15
	1 and 2	58.79	29.42	29.37	0.13	0.14	0.15
	3	58.96	29.48	29.48	0.13	0.13	0.15
3	Baseline	60.00	30.00	30.00	0.13	0.08	0.19
	1 and 2	58.36	29.57	28.80	0.14	0.13	0.18
	3	58.04	29.34	28.71	0.14	0.14	0.19
4	Baseline	59.67	29.84	29.84	0.16	0.05	0.27
	1 and 2	56.40	29.36	27.04	0.14	0.14	0.23
	3	56.38	29.45	26.93	0.15	0.14	0.24
5	Baseline	52.55	26.28	26.28	0.26	0.03	0.49
	1 and 2	35.87	29.30	6.57	0.19	0.14	0.70
	3	36.39	29.57	6.82	0.19	0.14	0.69
6	Baseline	18.88	9.44	9.44	0.55	0.55	0.56
	1 and 2	17.31	8.92	8.39	0.55	0.67	0.68
	3	18.57	9.32	9.26	0.54	0.66	0.66

Table 3.3.4: Average number of patients treated per trial in total and in each subgroup, average proportion of toxicities observed per trial in total and in each subgroup.

The average proportion of toxicities observed in the biomarker negative subgroup under the baseline method decreases for Scenario 1 through 5 while that in the biomarker negative group increases. This is for no substantial difference in the number

of patients treated between subgroups. This contrasting proportion of DLTs observed in the two subgroups demonstrates that throughout trials, most biomarker negative patients were being underdosed, with an average of only 3% experiencing DLTs in Scenario 5, while on average 49% of biomarker positive patients treated experienced DLTs in this scenario and hence many were likely overdosed.

We also see that, despite there being no tolerated dose in the biomarker positive subgroup in Scenario 5, under the baseline method, an average of 26.28 patients were treated in this subgroup per trial. This is compared to around 7 under the methods which accounted for a subgroup effect. It is the ability of the methods which account for a potential subgroup effect to stop for safety in one subgroup but continue escalation in the other that leads to this advantage.

The reduced number of patients treated in the biomarker positive subgroup under Methods 1-3 in Scenario 5 and the sample sizes observed for both subgroups in Scenario 6 show that the stopping criterion for safety is effective. It had the effect of reducing the overall average sample size from 60 to below 19 when there was no tolerated dose in either subgroup. In that scenario (Scenario 6), all methods were comparable, with around 90% of trials correctly identifying that there was no tolerated dose in either subgroup (Table 3.3.5). The baseline method was comparable to the alternative in this case because its underlying assumption, that there was no subgroup effect, was correct.

					Recommended dose												
	Escalation	Significant subgroup effect			$I_+ = 0$							$I_+ = 1$						
Scenario	Method	0	1	2	None	100	150	180	215	245	260	None	100	150	180	215	245	260
	Baseline	1000	0	0	0.01	0.01	0.05	0.49	0.36^X	0.07	0.02	0.01	0.01	0.05	0.49	0.36^X	0.07	0.02
1	1	0	951	49	0.02	0.02	0.11	0.39	0.33^X	0.08	0.04	0.03	0.02	0.10	0.38	0.33^X	0.09	0.04
	2	755	196	49	0.02	0.01	0.07	0.39	0.40^X	0.09	0.03	0.03	0.01	0.06	0.39	0.39^X	0.09	0.03
	3	666	298	36	0.03	0.01	0.09	0.40	0.36^X	0.09	0.03	0.02	0.01	0.10	0.40	0.36^X	0.08	0.03
	Baseline	1000	0	0	0.01	0.01	0.11	0.58	0.28^X	0.02	0.00	0.01	0.01	0.11	0.58^X	0.28	0.02	0.00
2	1	0	962	38	0.03	0.01	0.11	0.42	0.32^X	0.07	0.04	0.02	0.03	0.25	0.49^X	0.19	0.02	0.00
	2	745	217	38	0.02	0.01	0.09	0.47	0.32^X	0.06	0.03	0.02	0.01	0.15	0.52^X	0.25	0.04	0.02
	3	662	304	34	0.02	0.02	0.11	0.45	0.32^X	0.06	0.03	0.02	0.03	0.20	0.50^X	0.22	0.02	0.01
	Baseline	1000	0	0	0.00	0.01	0.34	0.59	0.06^X	0.00	0.00	0.00	0.01	0.34^X	0.59	0.06	0.00	0.00
3	1	0	945	55	0.02	0.02	0.13	0.36	0.32^X	0.10	0.04	0.04	0.13	0.55^X	0.26	0.01	0.00	0.00
	2	662	283	55	0.02	0.01	0.19	0.43	0.24^X	0.08	0.02	0.05	0.06	0.36^X	0.41	0.10	0.01	0.02
	3	423	511	66	0.03	0.01	0.17	0.41	0.26^X	0.08	0.04	0.05	0.10	0.47^X	0.35	0.03	0.00	0.00
	Baseline	1000	0	0	0.01	0.30	0.68	0.01	0.00^X	0.00	0.00	0.01	0.30^X	0.68	0.01	0.00	0.00	0.00
4	1	0	871	129	0.03	0.02	0.12	0.40	0.32^X	0.08	0.03	0.11	0.76^X	0.13	0.00	0.00	0.00	0.00
	2	519	352	129	0.04	0.05	0.25	0.35	0.23^X	0.06	0.03	0.12	0.37^X	0.25	0.16	0.06	0.01	0.02
	3	73	804	123	0.02	0.04	0.13	0.36	0.34^X	0.09	0.03	0.11	0.74^X	0.15	0.00	0.00	0.00	0.00
	Baseline	1000	0	0	0.17	0.83	0.00	0.00	0.00^X	0.00	0.00	0.17^X	0.83	0.00	0.00	0.00	0.00	0.00
5	1	0	69	931	0.03	0.02	0.11	0.39	0.32^X	0.09	0.04	0.95^X	0.05	0.00	0.00	0.00	0.00	0.00
	2	64	5	931	0.04	0.02	0.11	0.38	0.31^X	0.09	0.05	0.95^X	0.00	0.01	0.01	0.01	0.00	0.01
	3	7	62	931	0.02	0.02	0.11	0.37	0.36^X	0.08	0.04	0.95^X	0.05	0.00	0.00	0.00	0.00	0.00
	Baseline	1000	0	0	0.89^X	0.10	0.00	0.00	0.00	0.00	0.00	0.89^X	0.10	0.00	0.00	0.00	0.00	0.00
6	1	0	183	817	0.89^X	0.10	0.00	0.00	0.00	0.00	0.00	0.91^X	0.09	0.00	0.00	0.00	0.00	0.00
	2	183	0	817	0.89^X	0.10	0.00	0.00	0.00	0.00	0.00	0.91^X	0.09	0.00	0.00	0.00	0.00	0.00
	3	323	0	677	0.90^X	0.10	0.00	0.00	0.00	0.00	0.00	0.90^X	0.10	0.00	0.00	0.00	0.00	0.00

Table 3.3.5: Number of trials which identify a subgroup effect (0 = no subgroup effect, 1 = significant subgroup effect, 2 = defaulted to subgroup effect after stopping for safety in one subgroup) and proportion of times each dose was recommended by subgroup out of trials giving a recommended dose (based on a frequentist calculation). Grey cells highlight dose-pairs with probability of causing a DLT in a patient greater than 0.35. The 'X' marks the dose with probability of toxicity closest to 0.16.

In Scenario 1, the bulk of recommended doses by all methods are split between 180mg/m^2 and 215mg/m^2. This is not completely unexpected as the true TD16 for this scenario is 206mg/m^2 which falls between the two but being slightly closer to 215mg/m^2. The true recommended doses, along with the probability of toxicity for all scenarios are given in Table 3.3.3. The locations of the recommended doses in Scenario 1 were also similar across all methods. This suggests that when a suitable number of patients are treated in each subgroup (with 30 appearing to be suitable), the recommended dose is identified with a reasonable level of accuracy. So, even when there is no subgroup effect, there is no clear loss in using a method which accounts for a potential subgroup effect compared to the baseline method when a sufficient number

of patients are treated. Note that, if the baseline method was run with a total of 30 patients (based on number treated per assumed homogeneous population), then the recommended dose locations would be the same (aside from simulation error) as those from one or other subgroup under Method 1.

Now consider the locations of recommended doses from Scenarios 2-5 (Table 3.3.5). As the subgroup effect increases, the baseline method gets progressively worse. This is because under the baseline method, the assumption is that all observations arise from the same population; the resulting recommended dose is effectively a compromise between the true recommended doses from the two subgroups. The most undesirable outcome from the baseline method arises from Scenario 5 where the true recommended dose in the biomarker negative subgroup is 215mg/m^2 and there is no tolerated dose in the biomarker positive subgorup. In 17% of trials the baseline method stops for safety in both subgroups, and in the remaining trials it identified the recommended dose for the entire population as 100mg/m^2. This means that 83% of the time a dose which has 'true' probability of DLT of 0.02 (expected to be inefficacious) and 0.42 (undesirably toxic) in the two subgroups is recommended for further testing.

Method 1, which considers a potential subgroup effect throughout escalation and in dose recommendation, performs much better than the baseline. This suggests that 30 patients, with the levels of variability observed here, are suitable to identify a recommended dose in a homogeneous population with reasonable accuracy. As previously discussed, ideally we would like some idea of whether a subgroup effect was in fact observed. This could be achieved using a hypothesis test (as in Method 2). In Scenario 1, the proportion of correctly identified recommended doses by Method 2

was greater than that from Method 1. This is because 79% of trials failed to reject the null hypothesis, hence correctly concluding the absence of a subgroup effect. Based on this conclusion, data were pooled and a more accurate dose recommendation obtained than would be based on half the data (as in Method 1).

Unfortunately, the low power of the hypothesis test means that although the recommended doses from Method 2 improve upon the baseline in Scenarios 2-4, they are notably worse than those from Method 1. Method 3 was designed to avoid this problem and does so fairly successfully. Only small differences in recommended dose locations are seen between the baseline method and Method 3 in Scenario 1, with a conclusion of no subgroup effect under Method 3 66.6% of the time. Under Scenarios 2-5, the recommended doses by Method 3 are improved upon those from Method 2, suggesting that it has more power to detect a subgroup effect than the hypothesis test. In the presence of a medium subgroup effect (as in Scenarios 3 and 4), the spike and slab priors are effective in identifying a subgroup effect. The proportion of times a subgroup effect is correctly identified is 57.7% and 92.7% in Scenarios 3 and 4, respectively. This is compared to only 33.8% and 48.1% under Method 2.

Additional scenarios

In addition to the simulations described above, Method 3 was run with a maximum of 120 patients per subgroup. From these results we were able to conclude that given a suitable number of patients, this method provides good estimation of the recommended dose in each subgroup. The results tables are given in Appendix 3.5.6.

Some additional scenarios were also run; the purpose was to investigate the sensitivity of the methods to different parameter values in the data generating dose-toxicity model. The same parameter values used to generate data for both subgroups in Scenario 1 were used for the biomarker negative subgroup, resulting in a true recommended dose of 215mg/m^2 in this subgroup in all cases. For the biomarker positive subgroup, the values of β_2 and β_3 were altered to create different scenarios but in a way that resulted in a true recommended dose of 150mg/m^2 in each case. The resulting dose-toxicity curves are shown in Figure 3.3.1. The corresponding true probability of DLT at each available dose is given in Appendix 3.5.5 along with a table giving some additional operating characteristics of the design.

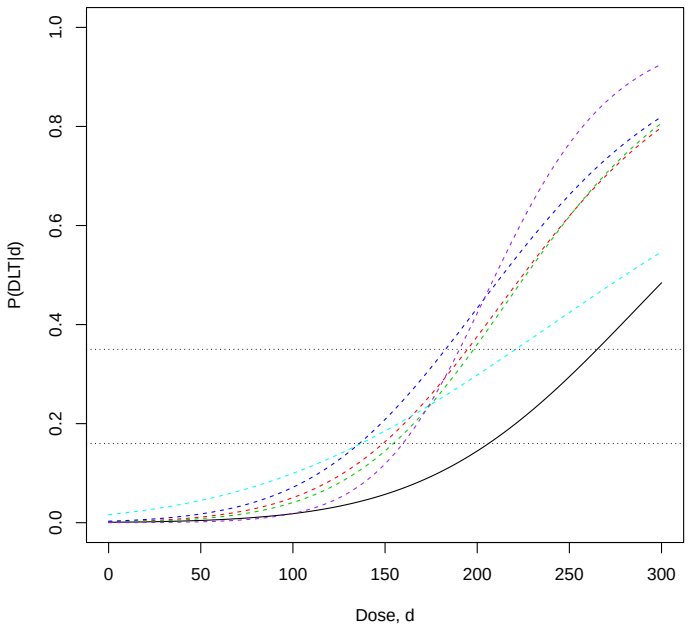

Figure 3.3.1: The dose-toxicity curves used to generate data in additional Scenarios 7-11. Horizontal lines are references at $\mathbb{P}(\text{DLT}|d) = 0.16$ and 0.35. The solid black curve on each plot represents that of the biomarker negative subgroup in all scenarios. The dose-toxicity curves for the biomarker positive group in these scenarios are shown for Scenarios 7-11 by the dashed red, green, dark blue, light blue and purple curves, respectively.

From the locations of the recommended doses for these additional scenarios, which are presented in Table 3.3.6, we can confirm that we have run a suitable number of simulations to be relatively certain in our conclusions drawn, for the given setting. This is seen from the consistency in the outcomes of the biomarker negative subgroup. The rest of this discussion is focussed on operating characteristics in the biomarker positive subgroup.

					Recommended dose													
	Escalation	Significant subgroup effect			$I_+ = 0$							$I_+ = 1$						
Scenario	Method	0	1	2	None	100	150	180	215	245	260	None	100	150	180	215	245	260
	Baseline	1000	0	0	0.01	0.02	0.48	0.45	0.04^X	0.00	0.00	0.01	0.02	0.48^X	0.45	0.04	0.00	0.00
7	1	0	936	64	0.02	0.02	0.13	0.37	0.33^X	0.09	0.04	0.05	0.21	0.58^X	0.15	0.01	0.00	0.00
	2	602	334	64	0.02	0.01	0.20	0.42	0.24^X	0.08	0.03	0.05	0.11	0.41^X	0.33	0.07	0.01	0.02
	3	364	567	69	0.02	0.02	0.21	0.37	0.29^X	0.06	0.03	0.06	0.20	0.53^X	0.20	0.02	0.00	0.00
	Baseline	1000	0	0	0.01	0.02	0.42	0.52	0.04^X	0.00	0.00	0.01	0.02	0.42^X	0.52	0.04	0.00	0.00
8	1	0	928	72	0.03	0.01	0.11	0.41	0.33^X	0.07	0.04	0.05	0.16	0.57^X	0.21	0.01	0.00	0.00
	2	649	279	72	0.03	0.01	0.20	0.44	0.24^X	0.06	0.02	0.06	0.07	0.39^X	0.39	0.08	0.01	0.01
	3	399	540	61	0.02	0.02	0.19	0.39	0.28^X	0.07	0.03	0.04	0.15	0.51^X	0.27	0.03	0.00	0.00
	Baseline	1000	0	0	0.00	0.07	0.65	0.27	0.01^X	0.00	0.00	0.00	0.07	0.65^X	0.27	0.01	0.00	0.00
9	1	0	896	104	0.02	0.01	0.13	0.41	0.32^X	0.07	0.04	0.10	0.40	0.45^X	0.06	0.00	0.00	0.00
	2	562	334	104	0.03	0.03	0.24	0.38	0.22^X	0.07	0.02	0.11	0.20	0.37^X	0.24	0.05	0.01	0.01
	3	268	636	96	0.02	0.03	0.21	0.36	0.28^X	0.07	0.04	0.09	0.33	0.46^X	0.11	0.01	0.00	0.00
	Baseline	1000	0	0	0.01	0.08	0.42	0.41	0.07^X	0.01	0.00	0.01	0.08	0.42^X	0.41	0.07	0.01	0.00
10	1	0	860	140	0.02	0.01	0.14	0.40	0.31^X	0.08	0.04	0.14	0.33	0.36^X	0.14	0.03	0.00	0.00
	2	553	307	140	0.03	0.03	0.22	0.38	0.26^X	0.07	0.03	0.14	0.17	0.31^X	0.26	0.08	0.01	0.02
	3	336	535	129	0.03	0.03	0.18	0.37	0.29^X	0.07	0.04	0.12	0.31	0.33^X	0.19	0.04	0.00	0.00
	Baseline	1000	0	0	0.00	0.00	0.38	0.60	0.01^X	0.00	0.00	0.00	0.00	0.38^X	0.60	0.01	0.00	0.00
11	1	0	972	28	0.02	0.02	0.13	0.40	0.32^X	0.08	0.04	0.01	0.08	0.65^X	0.26	0.00	0.00	0.00
	2	684	288	28	0.02	0.01	0.19	0.46	0.24^X	0.06	0.02	0.01	0.03	0.41^X	0.42	0.09	0.01	0.02
	3	406	559	35	0.03	0.01	0.19	0.37	0.27^X	0.09	0.04	0.02	0.08	0.57^X	0.31	0.02	0.00	0.00

Table 3.3.6: Number of trials which identify a subgroup effect (0 = no subgroup effect, 1 = significant subgroup effect, 2 = defaulted to subgroup effect after stopping for safety in one subgroup) and proportion of times each dose was recommended by subgroup out of trials giving a recommended dose (based on a frequentist calculation), for Scenarios 7-11. Grey cells highlight dose-pairs with probability of causing a DLT in a patient greater than 0.35. The 'X' marks the dose with probability of toxicity closest to 0.16.

It is difficult to make any firm conclusions concerning the effect of each of the parameters on the methods but it is clear that the overall comparisons between the methods which we have already made stand in all cases. Despite the different parameter values used to generate data in Scenarios 7 and 8, the resulting dose-toxicity

curves are fairly similar over the dose range of interest. This is likely to be the reason that the operating characteristics of these scenarios are similar. Although the dose-toxicity curve for Scenario 9 is not greatly dissimilar to those of Scenarios 7 and 8, there appears to an increased chance of stopping early. This could be down to the value of β_2 being greater than β_3 because this observation is more evident in Scenario 10 which has an even larger difference in parameter values. Scenario 11 results in a dose-toxicity curve with low toxicity at low doses but then increases steeply. The average proportion of toxicities observed in the trial are therefore decreased and fewer trials stop for safety.

Allowing early stopping for accuracy

Although a total of 30 patients (or more) in each subgroup is desirable, it is not always feasible. Along with the stopping rules which were used in the previous simulations (for safety in a subgroup or having treated the maximum number of patients in each subgroup), we now include one for accuracy. That is, the trial can stop for accuracy in a subgroup if a minimum of 5 patients have been treated at the dose advised for administration to the next cohort of patients and the ratio of the upper and lower bounds of the 95% credible interval around the estimate of that dose is less than 5 (as used by Whitehead et al., 2006a). We compare the impact of this stopping rule on Methods 1 and 3. The baseline design is not considered here because we have already confirmed that it is not suitable when a subgroup effect is present. In a homogeneous population, the effect of stopping rules is similar to that seen in one subgroup for Method 1. Method 2 is not considered because it would require use of an interim

analysis. The reduced number of patients available at the interim analysis, as well as control for multiplicity, would result in the test being even lower powered.

Introducing the stopping rule for accuracy was effective in reducing the sample size of the trial; this can be seen from the operating characteristics of the methods presented in Table 3.3.7. In Scenarios 1-4, where there was a tolerated dose in each subgroup, the average number of patients in the trial is between 45 and 51 in both methods. Even based on these reduced sample sizes, the locations of the recommended doses are still compacted around the true recommended dose; this can be seen in Table 3.3.8. Table 3.3.9 shows the reason that trials stopped. We see that in Scenario 1, under both methods, 45-49% of trials stopped early for accuracy in both subgroups. In Method 1 for Scenarios 2-5, the proportion of trials which stopped early for accuracy was consistently around these values when there was a tolerated dose in the subgroup.

Scenario	Escalation method	Average number patients			Average proportion toxicities		
		Overall	$\mathbb{I}_+ = 0$	$\mathbb{I}_+ = 1$	Overall	$\mathbb{I}_+ = 0$	$\mathbb{I}_+ = 1$
1	1	48.80	24.12	24.68	0.12	0.14	0.14
	3	47.36	23.70	23.66	0.11	0.13	0.12
2	1	48.22	24.39	23.83	0.13	0.14	0.16
	3	47.90	23.31	24.59	0.12	0.13	0.15
3	1	49.29	24.99	24.29	0.14	0.13	0.18
	3	47.77	22.01	25.77	0.13	0.11	0.18
4	1	50.84	24.51	26.33	0.15	0.14	0.23
	3	45.40	18.94	26.46	0.15	0.12	0.26
5	1	32.55	25.58	6.97	0.19	0.14	0.68
	3	26.87	20.03	6.84	0.20	0.12	0.71
6	1	19.19	9.45	9.74	0.53	0.65	0.66
	3	18.80	9.13	9.66	0.53	0.67	0.65

Table 3.3.7: Average number of patients treated per trial in total and in each subgroup, average proportion of toxicities observed per trial in total and in each subgroup, in simulations which allow early stopping for accuracy.

In Method 3, the proportion of trials which stop for accuracy in the biomarker negative subgroup increases as the true subgroup effect increases, while decreasing in the biomarker positive subgroup. The reason for this large discrepancy is the model selection identifying the presence of a subgroup effect. It is therefore better able to estimate the dose-toxicity curve in the biomarker negative subgroup. This is because of the spread of data. On the other hand the high uncertainty surrounding the estimation of the dose-toxicity curve in the biomarker positive subgroup, caused by a lack of data at higher doses, leads to the reduced number of trials which stop for accuracy as the subgroup effect increases.

		Recommended dose													
	Escalation	$I_+ = 0$							$I_+ = 1$						
Scenario	Method	None	100	150	180	215	245	260	None	100	150	180	215	245	260
1	1	0.04	0.02	0.13	0.33	0.32^X	0.08	0.09	0.03	0.01	0.14	0.35	0.34^X	0.06	0.08
	3	0.03	0.01	0.11	0.43	0.26^X	0.10	0.06	0.02	0.02	0.10	0.42	0.24^X	0.12	0.07
2	1	0.03	0.02	0.14	0.34	0.32^X	0.06	0.09	0.03	0.03	0.24	0.43^X	0.23	0.02	0.02
	3	0.03	0.02	0.12	0.44	0.22^X	0.10	0.07	0.03	0.03	0.23	0.48^X	0.17	0.05	0.02
3	1	0.02	0.03	0.12	0.35	0.32^X	0.08	0.09	0.04	0.12	0.46^X	0.32	0.05	0.00	0.00
	3	0.02	0.02	0.16	0.41	0.20^X	0.11	0.07	0.05	0.10	0.45^X	0.36	0.04	0.01	0.00
4	1	0.03	0.01	0.13	0.35	0.31^X	0.07	0.09	0.11	0.74^X	0.14	0.00	0.00	0.00	0.00
	3	0.03	0.02	0.11	0.41	0.20^X	0.13	0.10	0.12	0.74^X	0.14	0.00	0.00	0.00	0.00
5	1	0.02	0.02	0.12	0.38	0.29^X	0.09	0.07	0.94^X	0.06	0.00	0.00	0.00	0.00	0.00
	3	0.02	0.02	0.11	0.43	0.25^X	0.11	0.07	0.93^X	0.07	0.00	0.00	0.00	0.00	0.00
6	1	0.89^X	0.11	0.00	0.00	0.00	0.00	0.00	0.88^X	0.12	0.00	0.00	0.00	0.00	0.00
	3	0.90^X	0.10	0.00	0.00	0.00	0.00	0.00	0.89^X	0.11	0.00	0.00	0.00	0.00	0.00

Table 3.3.8: Number of trials which identify a subgroup effect (0 = no subgroup effect, 1 = significant subgroup effect, 2 = defaulted to subgroup effect after stopping for safety in one subgroup) and proportion of times each dose was recommended by subgroup out of trials giving a recommended dose (based on a frequentist calculation), for Scenarios 7-11, in simulations which allow early stopping for accuracy. Grey cells highlight dose-pairs with probability of causing a DLT in a patient greater than 0.35. The 'X' marks the dose with probability of toxicity closest to 0.16.

Scenario	Escalation method	Reason trial stopped					
		$\mathbb{I}_+ = 0$			$\mathbb{I}_+ = 1$		
		Safety	Max	Accuracy	Safety	Max	Accuracy
1	1	0.03	0.50	0.49	0.02	0.55	0.45
	3	0.02	0.54	0.46	0.01	0.54	0.45
2	1	0.02	0.53	0.47	0.03	0.50	0.49
	3	0.02	0.52	0.47	0.03	0.62	0.36
3	1	0.01	0.56	0.45	0.04	0.57	0.40
	3	0.01	0.41	0.59	0.04	0.73	0.23
4	1	0.02	0.52	0.47	0.11	0.84	0.05
	3	0.02	0.23	0.77	0.12	0.87	0.01
5	1	0.01	0.56	0.46	0.92	0.08	0.00
	3	0.01	0.24	0.77	0.92	0.09	0.00
6	1	0.85	0.15	0.00	0.83	0.16	0.00
	3	0.87	0.13	0.00	0.85	0.15	0.00

Table 3.3.9: Proportion of trials which stopped for safety, having treated the maximum number of patients and for accuracy in each subgroup.

As expected, the stopping rule for accuracy does not come in to play in a subgroup in which there is no tolerated dose (as in the biomarker positive subgroup in Scenario 5 and both subgroups in Scenario 6). This is down to the stopping rule for safety being met.

3.4 Discussion

We extended a traditional dose-toxicity model, by including terms for subgroup membership, used in dose-escalation to account for a potential subgroup effect. In doing so, the assumption of a homogeneous trial population is removed, reducing the risk of a missed or masked treatment effect due to variability between subgroups of the population. The proposed dose-escalation methods, which account for a potential subgroup effect, follow a similar procedure to the standard Bayesian model-based de-

sign to which they were compared. In this way, after the initial set-up of the trial, they should be no more difficult to employ.

Simulation results showed that accounting for subgroup membership in dose-escalation can increase the safety of escalation. Importantly, Methods 1-3 had the ability to stop early for safety in a subgroup if there was no tolerated dose, reducing the number of overdoses recommended for use in future trials. Although a hypothesis test was low powered to detect a subgroup effect (shown by Method 2), simulation results showed that a proposed method, which used spike and slab priors on the terms for subgroup membership (presented as Method 3), was reasonably good at identifying the presence of an underlying subgroup. The recommended dose locations from Method 3 were similar to those from Method 1 but with the advantage of providing exploratory information concerning the presence of a subgroup effect. Also, when there was no identifiable subgroup effect, escalation and identification of the recommended dose makes better use of available data than Method 1.

The methods were initially compared with a total of 30 patients available for treatment in each subgroup. Although such a sample size would be desirable, it is not always feasible. The use of a stopping rule for accuracy demonstrated that an overall sample size of 45-50 is suitable for Methods 1 and 3 to identify a recommended dose with a relatively small loss in accuracy.

As with standard Bayesian model-based designs, the proposed methods are flexible and practical since available doses and cohort sizes, among other design factors, can be altered throughout the trial. We considered the optimal setting with cohorts of size two, consisting of one biomarker positive and one biomarker negative patient (unless

one subgroup had stopped for saftey). This could be altered but the more unevenly distributed the patients are between subgroups, the lower powered the hypothesis test (in Method 2) and worse the variable selection algorithm (in Method 3) will perform. The proposed methods can allow for different values of θ to be used in each subgroup, if required. In practice it is also still possible for the clinical team to over-ride the model decision based on any available data.

The methods proposed here only have the potential to highlight subgroup effects between the two pre-defined subgroups of the population. It could be beneficial to extend this to the ordinal setting (similar to that of Tighioutart et al., 2012). However, the sample size in dose-escalation trials is simply too small to consider identification of a subgroup effect, with suitable power, within the trial. Rogatko et al. (2005) propose extending the search for the optimal dose, and consideration of a subgroup effect, beyond dose-escalation. This can also help account for population changes and longer-term endpoints in the identification of an optimal dose.

3.5 Appendix

3.5.1 Power Calculations

When considering Method 2, we looked into the choice of hypothesis test on the presence of a subgroup effect in dose-escalation trials. Initially, we considered using a difference hypothesis test on the parameters for subgroup membership such that:

$$H_0 : \beta_2 = 0 \text{ and } \beta_3 = 0 \quad \text{versus} \quad H_1 : \beta_2 \neq 0 \text{ and/or } \beta_3 \neq 0.$$

An alternative was equivalence testing which, for a dose-escalation trial conducted using a method which accounts for a potential subgroup effect, may feel more natural. Specifying equivalence hypotheses on the parameters for subgroup membership is difficult. This is because, without knowing how β_2 and β_3 are related the equivalence bounds are hard to specify. An alternative is to base the test on the difference in recommended dose between the subgroups. In this case, the equivalence bounds could be more intuitively based on a dose difference expected to have a relevant difference in efficacy. For equivalence bound c and dose recommendations $\widehat{\text{TD}}100\theta_-$ and $\widehat{\text{TD}}100\theta_+$ from the biomarker negative and positive subgroups, respectively:

$$H_0 : \widehat{\text{TD}}100\theta_- - \widehat{\text{TD}}100\theta_+ \notin [-c, c] \quad \text{versus} \quad H_1 : \widehat{\text{TD}}100\theta_- - \widehat{\text{TD}}100\theta_+ \in [-c, c].$$

A difference hypothesis can also be defined which is based on the difference in recommended dose between the subgroups such that,

$$H_0 : \widehat{\text{TD}}100\theta_- - \widehat{\text{TD}}100\theta_+ = 0 \quad \text{versus} \quad H_1 : \widehat{\text{TD}}100\theta_- - \widehat{\text{TD}}100\theta_+ \neq 0.$$

This final hypothesis test is the one that was chosen for use in simulations. The reason is that it is the best powered hypothesis test. This is because it does not account for mulitplicity (as in the first difference hypothesis test) and the tail probabilities used an equivalence hypothesis. Despite this being the preferred test of the three, it is still low-powered. Also, when testing for a subgroup effect, we would ideally consider the entire dose-toxicity curve as opposed to the point estimate. Alternative tests, using closed testing procedures for example, could avoid the issue of multiplicity. However, decisions over which parameter to investigate first and the distribution of the type I error rates could be difficult.

We can confirm mathematically that the test will be low powered: We have assumed that the β parameters are normally distributed with means at the corresponding maximum likelihood estimates (equivalent to MAP estimates when prior data is incorporated in the regression in the same manner as trial data) and known variances. Since maximum likelihood estimates are invariant to transformation, it can be said that asymptotically, d is normally distributed with mean $\mathbb{E}[d]$ and variance $\text{Var}(d)$. Solving Equation 3.1.1 for d we find that

$$\mathbb{E}[d] = d^*(\mathrm{e}^y - 1) \quad \text{for} \quad y = \frac{\log\{\theta/(1-\theta)\} - \hat{\beta}_0}{\hat{\beta}_1}. \tag{3.5.1}$$

Using the delta method we find

$$\text{Var}(d) \approx \left(\frac{d^*}{\beta_1}\right)^2 \mathrm{e}^{2y}\{y^2\sigma_1^2 + 2y\text{cov}(\beta_0, \beta_1) + \sigma_0^2\}.$$

Fitting a logistic regression model to the pooled data from the paediatric trial of Temzolomide reported by Nicholson et al. (1998), we can obtain mean estimates of the β's and the covariance matrix of the terms. From these estimates we use the formulae for $\mathbb{E}[d]$ and $\mathrm{Var}(d)$ to find the dose recommendation to be approximately 206mg/m^2 and its standard deviation to be about 168.

In the simulation study, we specified the type I error rate $\alpha = 0.3$ and a maximum of 30 patients per subgroup. The power of such a trial to detect a difference of 35mg/m^2 in recommended dose between the two subgroups is

$$\Phi\left\{\frac{35\sqrt{n}}{168\sqrt{2} - \Phi^{-1}(1-\alpha/2)}\right\} = 0.41.$$

To achieve a type I error rate of 0.3 and power $1 - \beta = 0.2$ to detect a difference of 35mg/m^2 in recommended dose between the two subgroups then the sample size required is

$$\left\{\frac{\Phi^{-1}(1-\alpha/2) + \Phi^{-1}(1-\beta)168\sqrt{2}}{35}\right\}^2 = 163.$$

These calculations can be checked through simulation. They show that, as expected, our hypothesis test is low powered. In practice, due to the nature of dose-escalation, the spread of observed dose levels is unlikely to be equal and is likely to differ between subgroups and scenarios. This will lead to larger variances and differences in variance between the groups than those here, and hence even lower powered test for a fixed sample size.

3.5.2 Specifics of Variable Selection in Method 3

The four-parameter dose-toxicity model defined in Equation 3.2.1 to be able to account for subgroup membership, as well as dose of the treatment, is;

$$\begin{aligned} \log\left\{\frac{\pi(x,\mathbb{I}_+)}{1-\pi(x,\mathbb{I}_+)}\right\} &= \beta_0+\beta_1\log\left(\frac{x}{d^*}+1\right)+\mathbb{I}_+\left\{\beta_2+\beta_3\log\left(\frac{x}{d^*}+1\right)\right\} \\ \text{where } \pi(x,\mathbb{I}_+) &= \mathbb{P}(\text{DLT}|x,\mathbb{I}_+). \end{aligned}$$

In the absence of alternative information to base the prior on, the same pseudo data is used to define the prior for variable selection as was used for the purpose of model fitting in all methods. That is, in each subgroup, a prior probaility of toxicity at 100mg/m^2 of 16% and a prior probability of toxicity at 260mg/m^2 of 50% with 3 patients treated at each of the two prior doses. This data can be fitted to the four-parameter dose-toxicity model to obtain MAP estimates of the parameter vector;

$$\begin{aligned} \boldsymbol{\beta} &= [\beta_0,\beta_1,\beta_2,\beta_3]^T \\ &= [-3.136, 3.765, 5.993\times10^{-16}, -7.195\times10^{-16}]^T. \end{aligned}$$

The design matrix has columns: 1) the intercept term, 2) the transformed dose $(x/d+1)$, 3) the indicator of membership of the biomarker positive subgroup, 4) the interaction between transformed dose and the indicator of subgroup membership, for all possible combinations of the four-parameter dose-toxicity model terms. In our case it is given by,

$$\boldsymbol{X} = \begin{bmatrix} 1 & \log\left(\frac{100}{200}+1\right) & 0 & 0 \\ 1 & \log\left(\frac{100}{200}+1\right) & 1 & \log\left(\frac{100}{200}+1\right) \\ 1 & \log\left(\frac{150}{200}+1\right) & 0 & 0 \\ 1 & \log\left(\frac{150}{200}+1\right) & 1 & \log\left(\frac{150}{200}+1\right) \\ 1 & \log\left(\frac{180}{200}+1\right) & 0 & 0 \\ 1 & \log\left(\frac{180}{200}+1\right) & 1 & \log\left(\frac{180}{200}+1\right) \\ 1 & \log\left(\frac{215}{200}+1\right) & 0 & 0 \\ 1 & \log\left(\frac{215}{200}+1\right) & 1 & \log\left(\frac{215}{200}+1\right) \\ 1 & \log\left(\frac{245}{200}+1\right) & 0 & 0 \\ 1 & \log\left(\frac{245}{200}+1\right) & 1 & \log\left(\frac{245}{200}+1\right) \\ 1 & \log\left(\frac{260}{200}+1\right) & 0 & 0 \\ 1 & \log\left(\frac{260}{200}+1\right) & 1 & \log\left(\frac{260}{200}+1\right) \end{bmatrix} = \begin{bmatrix} 1 & 0.405 & 0 & 0 \\ 1 & 0.405 & 1 & 0.405 \\ 1 & 0.560 & 0 & 0 \\ 1 & 0.560 & 1 & 0.560 \\ 1 & 0.642 & 0 & 0 \\ 1 & 0.642 & 1 & 0.642 \\ 1 & 0.730 & 0 & 0 \\ 1 & 0.730 & 1 & 0.730 \\ 1 & 0.800 & 0 & 0 \\ 1 & 0.800 & 1 & 0.800 \\ 1 & 0.833 & 0 & 0 \\ 1 & 0.833 & 1 & 0.833 \end{bmatrix},$$

From this information, the prior response vector $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}$ can be calculated as

$$\boldsymbol{y} = [-1.609, -1.609, -1.029, -1.029, -0.719, -0.719,$$
$$-0.388, -0.388, -0.125, -0.125, -0.728 \times 10^{-15}, -0.728 \times 10^{-15}]^T.$$

The vector of prior probability of toxicity for each dose and subgroup membership combination is then

$$\frac{\mathrm{e}^{\boldsymbol{y}}}{1+\mathrm{e}^{\boldsymbol{y}}} = [0.167, 0.167, 0.263, 0.263, 0.328, 0.328,$$
$$0.404, 0.404, 0.469, 0.469, 0.500, 0.500]^T.$$

Based upon this we can see that the prior implies that doses of 100, 150 and 180mg/m^2 are tolerated (i.e. have prior probability of toxicity less than 0.35). From these doses, 100mg/m^2 is the optimal dose (based on the escalation criteria defined in Step 3 of the proposed methods) as it has prior probability of toxicity closest to the target rate of 0.16.

Using *BoomSpikeSlab* to specify a spike and slab prior for variable selection

In the *BoomSpikeSlab* package, specification of the spike and slab prior was achieved using the function *SpikeSlabPrior*. The arguments passed to this function are used to define the spike and slab components of the prior. The design matrix `x` $= \boldsymbol{X}$ and response vector `y` $= \boldsymbol{y}$ (defined above) are the arguments involved in specifying the slab component of the prior that we defined. For specifying the slab component of the prior, we defined the argument `prior.inclusion.probabilities = c(1, 1, 0.5, 0.5)` (for prior setting c and d of Table 3.5.9). The initial two components of this vector translate to β_0 and β_1 being forced into the dose-toxicity model. The last two components define the prior probability of inclusion of the terms for subgroup membership. In our simulation study, different settings (given in Table 3.5.9) of these parameters were investigated.

Defining these three arguments for *SpikeSlabPrior* meant that the alternative arguments (`mean.y`, `sd.y` and `expected.model.size`) do not need to be defined. Default values were used for the other arguments, details of which are given in Table 3.5.1 and reasons for this use are given in the remainder of this paragraph. The precision matrix of the βs should always be full rank in our setting and so `diagonal.shrinkage`

is not relevant. A lack of prior information to base estimates on makes defining values of `expected.r2`, `prior.df` and `optional.coefficient.estimate`, which are more suitable to our setting than the default values, difficult. We investigated the value of `prior.information.weight` argument and values significantly greater than the default caused the prior to be more influential on variable selection than we felt desirable under a range of scenarios. For this reason, the default value seemed to be as good a choice as any. Given that we are sampling only two inclusion indicators (γ_2 and γ_3) it does not appear to be detrimental to sample both at each iteration.

Using *BoomSpikeSlab* for variable selection with a spike and slab prior

In the *BoomSpikeSlab* package, variable selection on the logistic regression model using a spike and slab prior was achieved using the function *logit.spike*. The available prior and trial data was supplied to this function in the form of a data frame via the argument `data`. The data frame has one row of information for each (prior and available) dose of treatment with columns;

- `responses`: a two column matrix with the columns giving the number of successes (experienced a DLT in their first cycle of treatment) and failures,
- `log_TrD`: the log-transformed dose administered to the patient,
- `subgroup`: a 0/1 indicator of whether the patient is in the biomarker bositive subgroup.

The logistic regression model to be fitted to the available data is passed to the function via the argument; `formula = responses` $\sim$ `log_TrD + subgroup*log_TrD`. The

output of *SpikeSlabPrior* is used for the argument `prior` which defines the spike and slab prior on this logistic regression model. After checking for convergence from a range of scenarios, we specified that the chain should be run for `niter = 20,000` iterations with $5,000$ iterations for burn-in (although this is not used in *logit.spike*).

The additional arguments `ping` and `nthreads` do not affect the output of the function. Default values were used for the other arguments, details of which are given in Table 3.5.2 and reasons for this use are given in the remainder of this section. Since we wish to fit the model based on available data, which contains no missing values, the arguments `subset` and `na.action` are not relevant. In defining the seed for a larger function, the `seed` argument is also not required. The generated initial values are suitable to fit our relatively simple model efficiently and so the default was used for `initial.value`. The standard setting for `contrasts` was used, this is used in a varienty of R functions and appears to be suitable here too. Since we are only sampling for two parameters, it is suitable to sample them both at each step in the sampling algorithm so we can leave `mh.chunk.size` at its default value. The default value of `proposal.df` appears to work reasonably well and we have no information to support the use of an alternative proposal distribution any more than this one.

We consider a set of available dose pairs instead of a continuous range. For this reason, `drop.unused.levels` has no effect on inferences for our model; it is similar to having the reduced available dose range. The argument `clt.threshold` specifies when asymptotic results should be used and after testing a range of values of this argument we found that its effect on escalation decisions was minimal and so we chose to use the default value for this parameter.

Parameter	Function	Default value
`diagonal.shrinkage`	The weight placed on the diagonal of the precision matrix of the β's to make it full rank	`0.5`
`expected.r2`	The inverse gamma prior on the residual variance of the slab component of the prior is set equal to $\sigma^2 = \text{var}(y)(1 - \texttt{expected.r2})$	`0.5`
`prior.df`	The weight given to `expected.r2` in calculating σ^2 for the inverse gamma prior on the residual variance	`0.01`
`optional.coefficient.estimate`	An optional estimate of the regression coefficients	`NULL`
`prior.information.weight`	The number of observations worth of weight given to the prior estimate of the β's	`0.01`
`max.flips`	The maximum number of variable inclusion indicators sampled at each iteration (a value ≤ 0 causes all indicators to be sampled)	`-1`

Table 3.5.1: Name, function and default value of the arguments of *SpikeSlabPrior* for which the default values were used.

Parameter	Function	Default value
`subset`	An optional definition of a subset of observations to use in variable selection	`N/A`
`na.action`	Specifies how to handle missing values in the data	`options(‘‘na.action’’)`
`seed`	Seed for the C++ random number generator	`NULL`
`initial.value`	Initial value for Markov chain	`NULL`
`contrasts`	Used in handling of factors	`NULL`
`mh.chunk.size`	Maximum number of coefficients to draw in each Metropolis Hastings update	`10`
`proposal.df`	Degrees of freedom of t-distribution of the Metropolis Hastings proposal distribution	`3`
`drop.unused.levels`	Indicator of whether unobserved factor levels should be dropped	`TRUE`
`clt.threshold`	Number of successes/failures above which asymptotic result used to increase efficiency of algorithm	`2`

Table 3.5.2: Name, function and default value of the arguments of *logit.spike* for which the default values were used.

3.5.3 Prior specification

We chose to specify the prior to control the operating characteristics of the trial. This required investigation of the likely escalation patterns of a range of prior settings. We specify no prior subgroup effect (to aid comparison of the methods) and weight the prior data to $1/10^{\text{th}}$ of the planned trial size. So, in selecting a prior we investigated priors consisting of 3 patients worth of data under dose-escalation Method 1 in one subgroup.

In order to get a start dose of 100, this is selected as the lower of the prior doses with a prior probability of DLT at this level equal to 0.16, the target toxicity level. The higher prior dose, prior proportion of toxicities at that dose and the weighting of patients at each of the two doses was then altered in the investigated prior settings. These are given in Table 3.5.3.

Prior setting	Dose					
	100	150	180	215	245	260
1	1/6 (1.5)	-	-	1/3 (1.5)	-	-
2	1/6 (1.5)	-	-	-	-	1/3 (1.5)
3	1/6 (1.5)	-	-	-	-	1/2 (1.5)
4	1/6 (1.5)	-	-	-	-	2/3 (1.5)
5	1/6 (2)	-	-	-	-	1/3 (1)
6*	1/6 (2)	-	-	-	-	1/2 (1)
7	1/6 (2)	-	-	-	-	2/3 (1)
8	1/6 (1)	-	-	-	-	1/3 (2)
9	1/6 (1)	-	-	-	-	1/2 (2)

Table 3.5.3: Prior settings tested given in terms of the prior proportion of DLTs observed at each dose and in brackets, the number of prior patients observed at that dose out of the total of 3 patients. The '*' indicates the prior setting used in the simulation study.

The initial scenario that we looked at was that when no DLTs were observed. In this Scenario, prior setting 1 was found to escalate undesirably quickly, especially upon reaching 215mg/m^2 (the higher prior dose in this setting). This led to the high prior dose used being at the top of the available dose range (as in prior settings 2-9). In scenario 2 we can see that this has the effect of slowing escalation slightly at higher dose levels and reducing the chance of the curve flipping. It is however still fast escalation. This setting, as well as prior settings 5 and 8 were felt to escalate too quickly in this likely scenario to be used in the study. Settings 4, 6 and 7 are more cautious with escalation seemingly more controlled over the dose range. These patterns are shown in Table 3.5.4.

Dose	Prior Scenario								
	1	2	3	4	5	6	7	8	9
100	1	1	1	2	1	2	3	1	1
150	2	1	2	3	1	2	3	1	1
180	2	1	3	4	3	2	4	1	3
215	1	2	3	6	5	2	3	2	5
245	0	0	1	3	3	1	1	1	3
260	-	-	-	-	-	-	-	-	-

Table 3.5.4: Escalation pattern under a range of prior settings when no DLTs are observed. Entries are the number of patients treated at each dose before the model escalates to the next highest dose.

We went on to investigate the case where a DLT was observed. Table 3.5.5 shows the escalation pattern for the scenario in which a DLT is observed in an early cohort of patients at 100mg/m^2. If a DLT was observed in the first patient, then the remaining prior settings led to the escalation being stopped for safety. In settings 6 and 7 escalation could continue if a DLT was observed in the second patient treated at 100mg/m^2 while the other scenarios required treatment of two patients.

Dose	Prior Scenario				
	3	4	6	7	9
100	2, DLT, 7	2, DLT, 10	1, DLT, 9	1, DLT, 11	2, DLT, 7
150	5	7	4	6	6
180	3	4	2	3	4
215	1	3	1	1	2
245	1	1	0	1	1
260	-	-	-	-	-

Table 3.5.5: Escalation pattern under a range of prior settings with a DLT observed at 100mg/m^2. Entries are the number of patients treated at each dose before the model escalates, given a DLT observed at 100.

Dose	Prior Scenario				
	3	4	6	7	9
100	1 ; 6	2 ; 8	2 ; 7	3 ; 12	1 ; 5
150	DLT ; 7	DLT ; 9	DLT ; 7	DLT ; 9	DLT ; 7
180	7	10	5	8	7
215	3	5	2	3	3
245	1	2	1	2	1
260	-	-	-	-	-
100	1 ; 1	2	2	3	1 ; 2
150	2 ; 6	3 ; 6	2 ; 7	3 ; 7	1 ; 5
180	DLT ; 8	DLT ; 13	DLT ; 7	DLT ; 10	DLT ; 10
215	5	8	3	5	6
245	2	2	1	1	2
260	-	-	-	-	-

Table 3.5.6: Escalation pattern under a range of prior settings with a DLT observed at 150mg/m^2 and 180mg/m^2 for the respective table sections. Entries are the number of patients treated at each dose before the model escalates. A semi-colon represents a break in dosing at that level (i.e. escalation and de-escalation).

In further scenarios, observation of a DLT at 150mg/m^2 and 180mg/m^2 was considered. These results are given in Table 3.5.6. With observation of a DLT at 150mg/m^2, all prior settings led to administration of 100 for several patients before re-escalating. Similarly, for observation of a DLT at 180mg/m^2. In this case, scenarios 6 and 7 de-escalate by only one dose while the others de-escalate by two dose levels. Prior

setting 6 was selected as being the most suitable prior because of its consistent escalation in the case of no DLTs and the reduced number of patients (compared to setting 7) required to re-escalate if a DLT is observed early on in the trial.

In Section 3.3.2, the recommended dose locations given are based on a frequentist model fit to the data. Table 3.5.7 presents the recommended dose locations which would be identified using Bayesian and Table 3.5.8 the frequentist estimates from our prior set up for Method 2. The small difference in recommendations between the two suggest that weighting the prior to $1/10^{\text{th}}$ of the final expected data appears to be suitable to have limited effect on the final dose recommendation if a Bayesian estimate is to be used with the sample sizes considered here.

	Bayesian estimate of the recommended dose													
	$\mathbb{I}_+ = 0$							$\mathbb{I}_+ = 1$						
Scenario	None	100	150	180	215	245	260	None	100	150	180	215	245	260
1	0.02	0.02	0.11	0.42	0.35	0.05	0.03	0.03	0.02	0.10	0.43	0.32	0.05	0.04
2	0.02	0.01	0.12	0.44	0.32	0.04	0.04	0.02	0.03	0.24	0.54	0.15	0.01	0.00
3	0.02	0.01	0.13	0.41	0.33	0.06	0.04	0.04	0.13	0.58	0.24	0.01	0.00	0.00
4	0.03	0.02	0.12	0.44	0.32	0.05	0.04	0.11	0.78	0.12	0.00	0.00	0.00	0.00
5	0.03	0.01	0.10	0.42	0.34	0.05	0.04	0.98	0.02	0.00	0.00	0.00	0.00	0.00
6	0.95	0.05	0.00	0.00	0.00	0.00	0.00	0.97	0.03	0.00	0.00	0.00	0.00	0.00

Table 3.5.7: Bayesian calculations of the proportion of times each dose was recommended by subgroup out of trials giving a recommended dose, based on dose-escalation Method 2.

	Frequentist estimate of the recommended dose													
	$\mathbb{I}_+ = 0$							$\mathbb{I}_+ = 1$						
Scenario	None	100	150	180	215	245	260	None	100	150	180	215	245	260
1	0.02	0.02	0.11	0.39	0.33	0.08	0.04	0.03	0.02	0.10	0.38	0.33	0.09	0.04
2	0.03	0.01	0.11	0.42	0.32	0.07	0.04	0.02	0.03	0.25	0.49	0.19	0.02	0.00
3	0.02	0.02	0.13	0.36	0.32	0.10	0.04	0.04	0.13	0.55	0.26	0.01	0.00	0.00
4	0.03	0.02	0.12	0.40	0.32	0.08	0.03	0.11	0.76	0.13	0.00	0.00	0.00	0.00
5	0.03	0.02	0.11	0.39	0.32	0.09	0.04	0.95	0.05	0.00	0.00	0.00	0.00	0.00
6	0.89	0.10	0.00	0.00	0.00	0.00	0.00	0.91	0.09	0.00	0.00	0.00	0.00	0.00

Table 3.5.8: Frequentist calculations of the proportion of times each dose was recommended by subgroup out of trials giving a recommended dose, based on dose-escalation Method 2.

3.5.4 Investigating Inclusion Probabilities

We investigated the effect of the prior inclusion probability of β_2 and β_3 and also the boundary on the inclusion probability for inclusion of terms in the fitted model. The combinations investigated are given in Table 3.5.9.

Method	Prior setting	Prior inclusion probability on		Boundary for inclusion of term in model
		β_2	β_3	
3	a	0.3	0.3	0.25
	b	0.3	0.3	0.35
	c	0.5	0.5	0.25
	d	0.5	0.5	0.35
	e	0.7	0.7	0.25
	f	0.7	0.7	0.25

Table 3.5.9: Combinations of prior inclusion probability and boundary for inclusion of terms included in the model investigated in Method 3.

As expected, the average number of patients and proportion of DLTs were very similar in each of the inclusion probability settings. This confirms the safety criterion on escalation is effective and that, in general, escalation is targeting suitable doses. The effect of the inclusion probability parameters on the model choice also agreed with expectations. This can be seen from the number of trials which declared a significant subgroup effect in escalation, as shown in Table 3.5.10 for prior settings 1 and 3. Increasing the prior inclusion probability of the parameters lead to the terms for subgroup membership being included in the model more often. Increasing the bound for inclusion of a term in the model led to a decrease in how often the terms for subgorup membership were considered in the model, and hence how many trials concluded that a significant subgroup effect was present.

Scenario	PIP seting	Significant subgroup effect			Recommended dose													
					$\mathbb{I}_+ = 0$							$\mathbb{I}_+ = 1$						
		0	1	2	None	100	150	180	215	245	260	None	100	150	180	215	245	260
1	a	889.00	78.00	33.00	0.02	0.01	0.04	0.44	0.44	0.04	0.01	0.01	0.00	0.05	0.44	0.44	0.04	0.01
	b	906.00	46.00	48.00	0.02	0.01	0.05	0.48	0.39	0.05	0.01	0.03	0.00	0.04	0.48	0.38	0.06	0.01
	c	732.00	244.00	24.00	0.02	0.01	0.06	0.43	0.41	0.07	0.02	0.01	0.01	0.08	0.42	0.40	0.06	0.02
	d	827.00	140.00	33.00	0.02	0.01	0.06	0.45	0.39	0.07	0.02	0.02	0.01	0.05	0.44	0.40	0.07	0.02
	e	287.00	668.00	45.00	0.03	0.01	0.10	0.43	0.32	0.07	0.03	0.03	0.01	0.09	0.41	0.37	0.07	0.02
	f	535.00	431.00	34.00	0.02	0.01	0.07	0.41	0.39	0.07	0.03	0.02	0.01	0.08	0.42	0.38	0.07	0.02
3	a	636.00	307.00	57.00	0.02	0.01	0.20	0.51	0.20	0.06	0.01	0.04	0.06	0.41	0.46	0.03	0.00	0.00
	b	743.00	207.00	50.00	0.02	0.01	0.22	0.53	0.17	0.04	0.01	0.04	0.05	0.37	0.50	0.04	0.00	0.00
	c	429.00	514.00	57.00	0.02	0.00	0.14	0.44	0.30	0.07	0.02	0.04	0.08	0.50	0.35	0.04	0.00	0.00
	d	560.00	386.00	54.00	0.02	0.01	0.18	0.48	0.23	0.05	0.02	0.04	0.08	0.44	0.41	0.03	0.00	0.00
	e	133.00	819.00	48.00	0.02	0.01	0.10	0.42	0.35	0.08	0.03	0.03	0.09	0.57	0.30	0.01	0.00	0.00
	f	261.00	677.00	62.00	0.01	0.01	0.12	0.43	0.33	0.08	0.02	0.05	0.08	0.57	0.29	0.01	0.00	0.00

Table 3.5.10: Number of trials which identify a subgroup effect (0 = no subgroup effect, 1 = significant subgroup effect, 2 = defaulted to subgroup effect after stopping for safety in one subgroup) and proportion of times each dose was recommended by subgroup out of trials giving a recommended dose (based on a frequentist calculation) in Method 3 a range of inclusion probability settings.

3.5.5 Dose-toxicity Scenarios Investigated and Additional Results Table

The dose-toxicity curves corresponding to Scenarios 1-6 of the simulation study are presented in Figure 3.5.1. The parameter values and resulting probability of toxicity for data generated for the biomarker positive subgroup for additional Scenarios 7-11 are given in Table 3.5.11. A table showing some operating characteristics of this design are given in Table 3.5.12.

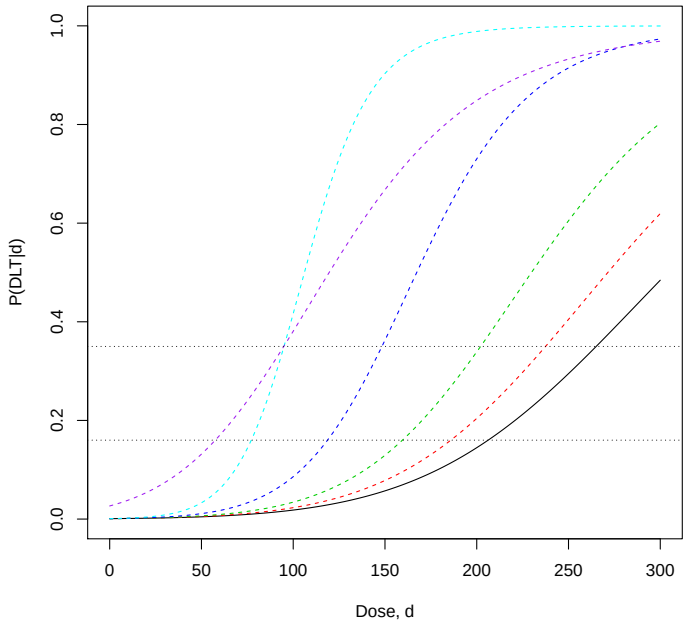

Figure 3.5.1: The dose-toxicity curves used to generate data in Scenarios 1-6 of the simulation study. Horizontal lines are references at $\mathbb{P}(\text{DLT}|d) = 0.16$ and 0.35. The solid black curve represents both subgroups in Scenario 1 and the biomarker negative subgroup in Scenarios 2-5. The dose-toxicity curve for the biomarker positive group in these scenarios are shown by the dashed red, green, dark blue and light blue curves, respectively. The dose-toxicity curves for both subgroups in Scenario 6 is shown by the dashed purple curve.

Scenario	Parameter value				P(DLT|d, I+ = 1)					
	β_0	β_1	β_2	β_3	100	150	180	215	245	260
1	-7.10	7.68	0.00	0.00	0.02	0.06	0.10	0.18^X	0.28	0.33
7	-7.10	7.68	0.75	0.75	0.05	0.16^X	0.28	0.45	0.60	0.66
8	-7.10	7.68	0.30	1.30	0.04	0.15^X	0.26	0.44	0.59	0.66
9	-7.10	7.68	1.30	0.30	0.07	0.21^X	0.34	0.51	0.64	0.70
10	-7.10	7.68	3.00	-3.00	0.10	0.19^X	0.25	0.34	0.41	0.45
11	-7.10	7.68	-2.00	5.00	0.02	0.12^X	0.28	0.54	0.74	0.81

Table 3.5.11: Parameter value and simulated probability of DLT at each dose (in mg/m^2) to be tested in the additional simulations, given for biomarker positive subgroup. Dark grey cells highlight dose-pairs with probability of causing a DLT in a patient greater than 0.35. The 'X' marks the dose with probability of toxicity closest to 0.16, in cases where there is a tolerated dose.

Scenario	Escalation method	Average number patients			Average proportion toxicities		
		Overall	$\mathbb{I}_+ = 0$	$\mathbb{I}_+ = 1$	Overall	$\mathbb{I}_+ = 0$	$\mathbb{I}_+ = 1$
7	Baseline	59.83	29.91	29.91	0.14	0.07	0.20
	1 and 2	58.11	29.51	28.61	0.14	0.13	0.19
	3	57.96	29.57	28.39	0.14	0.13	0.21
8	Baseline	59.94	29.97	29.97	0.13	0.08	0.19
	1 and 2	57.93	29.40	28.53	0.14	0.14	0.20
	3	58.19	29.48	28.71	0.14	0.13	0.20
9	Baseline	59.95	29.97	29.97	0.14	0.06	0.22
	1 and 2	56.99	29.71	27.28	0.14	0.13	0.23
	3	57.23	29.54	27.69	0.15	0.13	0.23
10	Baseline	59.71	29.86	29.86	0.14	0.08	0.20
	1 and 2	55.78	29.54	26.25	0.15	0.13	0.27
	3	56.20	29.51	26.70	0.15	0.13	0.25
11	Baseline	59.94	29.97	29.97	0.13	0.08	0.19
	1 and 2	59.20	29.54	29.66	0.13	0.13	0.16
	3	59.00	29.45	29.55	0.14	0.13	0.17

Table 3.5.12: The number of patients treated per trial in total and in each subgroup, average proportion of toxicities observed per trial in total and in each subgroup, for Scenarios 7-11.

3.5.6 Long-run Simulations

Method 3 was run with 120 patients in each subgroup (with prior scaled up respectively) to confirm that the method works in theory, given suitable amounts of data. The results in Tables 3.5.13 and 3.5.14 confirm this.

	Average number patients			Average proportion toxicities		
Scenario	Overall	$\mathbb{I}_+ = 0$	$\mathbb{I}_+ = 1$	Overall	$\mathbb{I}_+ = 0$	$\mathbb{I}_+ = 1$
1	240.00	120.00	120.00	0.12	0.12	0.12
2	240.00	120.00	120.00	0.12	0.11	0.13
3	240.00	120.00	120.00	0.13	0.11	0.15
4	239.32	120.00	119.32	0.12	0.11	0.13
5	143.45	120.00	23.45	0.16	0.12	0.55
6	84.48	42.92	41.56	0.44	0.49	0.49

Table 3.5.13: Average number of patients treated per trial in total and in each subgroup, average proportion of toxicities observed per trial in total and in each subgroup for long-run simulations under Method 3.

				Recommended dose													
	Significant subgroup effect			$\mathbb{I}_+ = 0$							$\mathbb{I}_+ = 1$						
Scenario	0	1	2	None	100	150	180	215	245	260	None	100	150	180	215	245	260
1	911.00	89.00	0.00	0.00	0.00	0.01	0.44	0.52	0.02	0.00	0.00	0.00	0.01	0.43	0.53	0.03	0.00
2	818.00	182.00	0.00	0.00	0.00	0.01	0.63	0.34	0.01	0.01	0.00	0.00	0.06	0.72	0.21	0.00	0.00
3	279.00	721.00	0.00	0.00	0.00	0.07	0.49	0.40	0.04	0.01	0.00	0.01	0.71	0.27	0.01	0.00	0.00
4	2.00	992.00	6.00	0.00	0.00	0.01	0.49	0.45	0.04	0.01	0.01	0.96	0.03	0.00	0.00	0.00	0.00
5	0.00	32.00	968.00	0.00	0.00	0.01	0.47	0.47	0.04	0.01	0.99	0.01	0.00	0.00	0.00	0.00	0.00
6	557.00	0.00	443.00	0.89	0.10	0.01	0.00	0.00	0.00	0.00	0.89	0.09	0.01	0.00	0.00	0.00	0.00

Table 3.5.14: Number of trials which identify a subgroup effect (0 = no subgroup effect, 1 = significant subgroup effect, 2 = defaulted to subgroup effect after stopping for safety in one subgroup) and proportion of times each dose was recommended by subgroup out of trials giving a recommended dose (based on a frequentist calculation) for long-run simulations under Method 3.

Chapter 4

A Practical Design for a Dual-agent Dose-escalation Trial that Incorporates Pharmacokinetic Data

Abstract

Traditionally, model-based dose-escalation trial designs recommend a dose for escalation based on an assumed dose-toxicity relationship. Pharmacokinetic data are often available but are currently only utilised by clinical teams in a subjective manner to aid decision making if the dose-toxicity model recommendation is felt to be too high. Formal incorporation of pharmacokinetic data in dose-escalation could therefore make the decision process more efficient and lead to an increase in the precision

of the resulting recommended dose, as well as decreasing the subjectivity of its use. Such an approach is investigated in the dual-agent setting using a Bayesian design, where historical single-agent data are available to advise the use of pharmacokinetic data in the dual-agent setting. The dose-toxicity and dose-exposure relationships are modelled independently and the outputs combined in the escalation rules. Implementation of stopping rules highlight the practicality of the design. This is demonstrated through an example which is evaluated using simulation.

Keywords: Dose-escalation, pharmacokinetic data, dual-agent, escalation rules, combination treatment.

4.1 Introduction

Dose-escalation trials are usually first-in-man trials of a treatment for a given application. They proceed by administering successive cohorts of patients with increasing doses of the treatment in order to identify a recommended dose for use in efficacy trials (Pocock, 2004). Despite the need for an accurate dose recommendation, to maximise the treatment's chance of success in efficacy trials, identification of the recommended dose is typically based only on short-term, binary toxicity data. That is, an indicator of whether a dose-limiting toxicity (DLT) is observed in a patient during the first cycle of treatment. Other factors to consider in the design of a dose-escalation trial concern patient ethics and practical issues. These considerations include minimising the number of patients treated at sub-optimal dose levels and limits on time and patient resources.

A desirable dose-escalation trial design identifies the recommended dose rapidly and reliably while maintaining patient safety as a priority. Achieving these properties requires a compromise between the rate, in terms of the speed and efficiency of escalation, and safety of escalation. These properties are controlled largely by specification of two decision rules:

i. The escalation rule controls which dose is administered to a cohort of patients;

ii. The stopping rule controls when the trial is stopped.

Traditional, algorithmic dose-escalation trial designs are simple to implement as they rely on fixed escalation and stopping rules. For example, the 3+3 design (Storer, 1989) requires pre-specified, available doses and treats patients in cohorts of size three. Escalation proceeds using pre-specified decision rules such as those given here which are described and illustrated in Jaki et al. (2013):

i. Escalation rule:

   - If no DLTs are observed in a cohort of three patients, then we say that $0/3$ patients in a cohort experience a DLT and administer the next cohort of patients with the next higher pre-specified dose level;
   - If $1/3$ patients in a cohort experience a DLT, treat another cohort of patients at the same dose level.

ii. Stopping rule:

   - If $\geq 2/6$ patients treated at a dose level experience a DLT, stop the trial.

- The recommended dose is declared as the dose below that observed to have an unacceptable level of toxicity. The recommended dose from such a trial is often referred to as the maximum tolerated dose (MTD).

An alternative family of designs are model-based. These designs assume some model for the dose-toxicity relationship, enabling quantitative definition of the recommended dose as the dose with probability θ of causing a DLT in a patient. An overview of single-agent trial designs and the advantages of model-based designs over other available options are given by Jaki et al. (2013).

A combination of drugs may be required to increase the effectiveness of a treatment. The aim of dose-escalation of a combination treatment is traditionally to identify a recommended dose combination with probability θ of causing a DLT in a patient. For a dual-agent treatment the recommended dose combination will be a dose-pair. That is, a dose of each of the two drugs which, when administered together, have probability θ of causing a DLT in a patient. An additional complication of dual-agent over single-agent escalation is that there are now two drugs, both of which may have to be escalated to find the recommended dose-pair. In modelling the dose-response relationship, possible drug-drug interactions (DDIs) must also be accounted for. DDI's can act at the pharmacodynamic or pharmacokinetic level. The result is an increase, or decrease, in toxicity and response (for a DDI acting at the pharmacodynamic level) or in exposure (for a DDI acting at the pharmacokinetic level), compared with the case of no interaction. The case of no interaction is defined in Section 4.2 in relation to the relevant models.

Harrington et al. (2013) provide a good review of dual-agent dose-escalation trial designs and emphasise the advantages of model-based over algorithmic designs in such a setting. The advantages include improvements in operating characteristics such as administering fewer sub-optimal dose-pairs more reliable identification of the recommended dose-pair. These improvements come largely from the ability of model-based designs to use all available trial information to advise escalation decisions and attempt to estimate the entire dose-toxicity surface. This is in contrast to algorithmic designs which search for the recommended dose-pair with escalation decisions based largely on information from the previous cohort.

A common method of escalation in the combination setting involves fixing the dose of one drug and escalating the other (Dejardin et al., 2014; Ellis et al., 2013). Where both drugs require escalation, Yuan and Guosheng (2008) propose a sequential dose-escalation procedure. Ordering of dose-pairs in the dual-agent setting is difficult; although pair-wise ordering is possible, it is not ideal. When more than a few doses of each drug are being considered, then approaches (such as that of Bailey et al., 2009) which use a single-agent model with covariates also encounter difficulties. It is therefore preferable to model the entire dose-toxicity surface. A range of models for dual-agents have been suggested (Braun and Wang, 2010; Huo et al., 2012; Neuenschwander et al., 2015; Thall et al., 2003; Wang and Ivanova, 2005; Yin and Yuan, 2009) and different escalation rules considered (Sweeting and Mander, 2012; Wheeler et al., 2014). In this chapter we base our proposed designs on the dual-agent dose-toxicity model specified in Neuenschwander et al. (2015), details of this model are given in Section 4.2.1.

Model-based designs can be employed in a Bayesian manner, allowing incorporation of prior knowledge along with all available trial data. Although Bayesian methods are arguably subjective, their use in early phase clinical trials has been endorsed (CHMP et al., 2006) and can be beneficial at this stage in trials where little data are available. Basing priors upon relevant, available data can reduce the subjectivity of the design but care must still be taken to ensure sensible weighting of this historical information. Given reasonable priors, an assumed model (which suitably describes the dose-toxicity relationship) and specified escalation rules, an advised dose for escalation can be obtained from the model. When escalation is complete, a stopping rule is satisfied and a recommended dose-pair can be identified. In practice, at each stage in dose-escalation, a clinical team use all available data observed but not accounted for in the model (such as safety, efficacy, pharmacodynamic and pharmacokinetic data) to select a dose that is typically less than or equal to that advised by the model. Use of the data in this way is subjective and inefficient because this data is rarely modelled at this stage in trials and its use is inconsistent.

Dual-agent dose-escalation trial designs have been proposed which account for both binary (or ordinal) efficacy data and toxicity data (Mandrekar et al., 2007; Whitehead et al., 2006b, 2011). Dragalin et al. (2008) allow for continuous efficacy data but base decisions on the four-option probability combination set of binary efficacy and toxicity outcomes. The inclusion of continuous pharmacokinetic exposure data has not been considered in dose-escalation in the combination setting. However, several methods have been proposed in the single-agent setting (Holford, 1995; Newell, 1994; Piantadosi and Liu, 1996; Whitehead et al., 2007).

In this chapter, we present a simple method of formally incorporating pharmacokinetic data into a Bayesian, model-based, dual-agent dose-escalation trial design in order to improve escalation decisions. In the dual-agent setting, historical single-agent data can be incorporated into the model through the use of informative priors. Basing the design on single-agent data in conjunction with clinical knowledge is more favourable than relying on pre-clinical data alone, which does not transfer reliably to the clinical setting. In Section 4.2, the dose-toxicity and dose-exposure models are presented. In Section 4.3, the proposed method of dose-escalation, which is practical and utilises both dose-toxicity and dose-exposure models, is built up in four stages from an initial basic method. In this way, the impact of the desicion rules introduced at each stage can be seen. The practicality and overall benefits of the final proposed method become especially clear in Section 4.4 through presentation of the results of a simulation study comparing the methods discussed in Section 4.3. The chapter concludes with a discussion in Section 4.5.

4.2 Modelling the Data

In this section, we describe the dose-response models underlying the proposed dose-escalation procedure. Inferences drawn from the fitted models are used in the trial escalation and stopping rules which are discussed in Section 4.3. Suggested prior distributions are given with each of the models. These are used to illustrate the proposed method in Section 4.4 but could of course be altered if available information suggested another prior distribution might be better suited to the particular situation.

A detailed example of prior derivation for the dual-agent trial using historical single-agent data is given in Appendix 4.6.1. To obtain the posterior distributions of the parameters of the dose-response models, we use Markov Chain Monte Carlo methods as closed form solutions do not exist.

The dose-response models are presented in terms of a dual-agent dose-escalation trial of drug A and drug B for which sufficient single-agent trial data are available. As general notation for the dose-response relationships, take $i = \{\text{A, B}\}$ as an indicator of the administered drug for which the dose set $\mathbf{d}_i$ is available for treatment. Define d_i^* as some fixed reference dose of drug i used to standardise the doses. For $x_\text{A} \in \mathbf{d}_\text{A}$ and $x_\text{B} \in \mathbf{d}_\text{B}$, let $\{x_\text{A}, x_\text{B}\}$ denote the dose-pair administered to a patient. Both of the dose-response models use the transformed, standardised dose $x_i/d_i^* + 1$. This is done so that the dual-agent models reduce to the single-agent models if a dose of zero is used for the other drug.

4.2.1 The Dose-toxicity Model

The dose of a treatment administered to a cohort of patients is usually selected only after the previous cohort has been treated, their responses observed and the model updated based upon these responses. This is done to reduce the risk of toxic side-effects for patients involved in the trial by basing decisions on all available information, including the most current and in the case of dose-escalation potentially most relevant, on the treatment. In order to control the trial duration, this data must be available relatively soon after treatment. The toxicity data used in dose-escalation are typically a binary indicator of whether a patient experienced a DLT during the first cycle of

treatment (21 days, say).

Define $\pi(x_i)$ as the probability that a patient experiences a DLT given dose x_i of drug i. Although a one-parameter power model (as used in O'Quigley et al., 1990) can have improved identification of a target dose for a single target toxicty rate, the two-parameter logistic regression model (used for example in Neuenschwander et al., 2008) better estimates the entire dose-toxicity relationship (O'Quigley et al., 1990). This improved modelling of the entire dose-toxicity relationship provides flexibility for secondary objectives that concern toxicity rates besides θ. We therefore use the following two-parameter model as the single-agent dose-toxicity model upon which the dual-agent model is based:

$$\log\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = \log(\alpha_i) + \beta_i \log\left(\frac{x_i}{d_i^*} + 1\right) \quad \text{where } \pi_i = \mathbb{P}(\text{DLT}|x_i). \tag{4.2.1}$$

To extend this single-agent model to the dual-agent setting, we need to allow for the dose of the second drug and for a potential toxicity DDI. This is achieved by considering the odds of toxicity at a dose of each treatment and introducing the interaction term, ζ. For interaction parameter ζ, the dual-agent dose-toxicity model with dose-dependent interaction term is defined as (Neuenschwander et al., 2015):

$$\text{odds}(\pi(x_\text{A}, x_\text{B})) = \text{odds}(\pi^0(x_\text{A}, x_\text{B})) \exp\left(\zeta \frac{x_\text{A}}{d_\text{A}^*} \frac{x_\text{B}}{d_\text{B}^*}\right), \tag{4.2.2}$$

$$\text{where } \pi(x_\text{A}, x_\text{B}) = \mathbb{P}(\text{DLT}|x_\text{A}, x_\text{B}, \text{possible DDI}),$$

$$\text{odds}(\pi^0(x_\text{A}, x_\text{B})) = \frac{\pi^0(x_\text{A}, x_\text{B})}{1-\pi^0(x_\text{A}, x_\text{B})},$$

$$\text{and } \pi^0(x_\text{A}, x_\text{B}) = \pi(x_\text{A}) + \pi(x_\text{B}) - \pi(x_\text{A})\pi(x_\text{B}). \tag{4.2.3}$$

From this formulation, it can be seen that a value of $\zeta = 0$ implies no toxicity interaction under the assumption of Bliss independence (as defined in Equation 4.2.3). The assumption of Bliss independence holds if, for example, drugs A and B were selected for the dual-agent trial because they target different cell pathways and, hence, are expected to yield non-overlapping toxicities. The combination would therefore be expected to have increased efficacy over the single-agents for a given toxicity rate. When there is a toxicity interaction, a value of $\zeta > 0$ implies an increase in the odds of toxicity, while $\zeta < 0$ implies a decrease in the odds of toxicity, in the dual-agent setting at the reference doses compared with the case of no toxicity interaction, assuming Bliss independence.

Multivariate normal prior distributions are specified on the single-agent parameters $\{\log(\alpha_A), \log(\beta_A)\}$ and $\{\log(\alpha_B), \log(\beta_B)\}$. A normal prior distribution is specified on the interaction parameter ζ. In the presence of a lack of reliable prior information on the potential DDI, this distribution could be centered at zero with a large variance. Alternatively, a cautious prior could be centered on the case of an interaction which leads to an increase in the odds of toxicity. More information on the choice of prior is given in Appendix 4.6.1.

4.2.2 The Dose-exposure Model

To ensure the feasibility of utilising pharmacokinetic data in dose-escalation, this data should also be available within the first cycle of treatment. The pharmacokinetic parameters of interest to us are measures of exposure to the drug and can be obtained from a concentration-time curve. Two useful exposure parameters are the area under

the curve (AUC), a measure of the average drug concentration over a fixed period of time, and the maximum concentration after administration of treatment (C_{max}) (Jambhekar et al., 2009). The single-agent model used as the basis for the dual-agent dose-exposure model is a linear regression model for the logarithm of a selected pharmacokinetic exposure parameter (PK), given dose x_i of drug i is

$$\log(PK(x_i)) \sim \mathrm{N}\left(\phi_{1i} + \phi_{2i}\log\left(\frac{x_i}{d_i^*} + 1\right), \sigma_i^2\right). \tag{4.2.4}$$

Note that we consider a single exposure parameter (e.g. AUC or C_{max}), which has been chosen for each drug. The choice of exposure parameter should be motivated by whether toxicity in the single-agent trials appeared to be driven by AUC or C_{max}. This model is equivalent to the standard regression model for dose-proportionality, which is frequently utilised in pharmacokinetic studies, with ϕ_2 as the power coefficient. The only difference is that we use the transformed, standardised dose. This is done here so that in the dual-agent extension of this model, the same transformation of dose is used for both drugs. The transformation we have chosen allows for zero doses of the drugs which is important for consistency in the dual-agent setting.

The case of no exposure interaction for drug A is defined such that the exposure to drug A is equal to that expected if administered as a single-agent. Similarly for no exposure interaction for drug B. So, for interaction parameters, ϕ_{3A} and ϕ_{3B}, the dual-agent dose-exposure models are as follows:

$$\log(PK(x_\mathrm{A})) \sim \mathrm{N}\left(\phi_{1A} + \phi_{2A}\log\left(\frac{x_A}{d_A^*} + 1\right) + \phi_{3A}\log\left(\frac{x_B}{d_B^*} + 1\right), \sigma_A^2\right) \tag{4.2.5}$$

$$\text{and } \log(PK(x_\mathrm{B})) \sim \mathrm{N}\left(\phi_{1B} + \phi_{2B}\log\left(\frac{x_B}{d_B^*} + 1\right) + \phi_{3B}\log\left(\frac{x_A}{d_A^*} + 1\right), \sigma_B^2\right) \tag{4.2.6}$$

The result is two dose-exposure models: one for the exposure to drug A and another for exposure to drug B. From this formulation, it can be seen that for $PK(x_{\mathrm{A}})$, a value of $\phi_{3\mathrm{A}} = 0$ (or $x_{\mathrm{B}} = 0$) implies no exposure interaction. When there is an exposure interaction, a value of $\phi_{3\mathrm{A}} > 0$ implies an increase in exposure to drug A, while $\phi_{3\mathrm{A}} < 0$ implies a decrease in exposure to drug A, at the reference doses compared to the case of no exposure interaction. Similarly for exposure to drug B.

Multivariate normal prior distributions are used on the single-agent parameters $\{\phi_{1A}, \phi_{2A}\}$ and $\{\phi_{1B}, \phi_{2B}\}$ and inverse gamma prior distributions used for each of the between-patient variability parameters, σ_A^2 and σ_B^2. As with the interaction parameter for the dose-toxicity model, a normal prior distribution is specified on each of the interaction parameters ϕ_{3A} and ϕ_{3A}, and the same logic stands in the face of little prior information on these parameters. More information on the choice of prior is given in Appendix 4.6.1.

4.2.3 Applying the Models

In Zhou et al. (2008), for the case of a single-agent dose-escalation trial, dose-toxicity and dose-exposure relationships are modelled independently by Equations 4.2.1 and 4.2.4, respectively. An optimum dose for administration to the next cohort of patients is identified for each model independently as the largest dose which satisfies specified safety criterion. The dose actually administered to patients is the minimum of the doses advised by the independent models.

We extend the single-agent models from Zhou et al. (2008) to the dual-agent setting (maintaining independent models for the dose-response relationships), resulting in the

dual-agent models in Equations 4.2.2, 4.2.5 and 4.2.6. We then combine the outputs of the models in the trial escalation rules. Our proposed escalation rules differ from those in Zhou et al. (2008) but incorporate similar safety constraints and targeting of exposure values, when this is accounted for. The method we propose is one possible extension of Zhou et al. (2008) in which models for the dose-toxicity and dose-exposure relationships are independent. The resulting model formulations lend themselves in a relatively straight-forward manner to prior specifications based on dose-escalation data from the corresponding single-agent trials.

4.3 Dual-agent Trial Designs

One of the implicit assumptions underlying dose-escalation trials is that efficacy monotonically increases with toxicity. For this reason, toxicity can be classified in relation to its expected effect on efficacy, as in Neuenschwander et al. (2008). It is often more realistic to target a desirable toxicity range, rather than a single value, and so we define the following toxicity intervals:

- $\pi_{\text{AB}} \in [0.00, 0.16)$ as an underdose;
- $\pi_{\text{AB}} \in [0.16, 0.35)$ as in the target toxicity interval; and
- $\pi_{\text{AB}} \in [0.35, 1.00]$ as an overdose.

Based on the assumption of monotonicity, and in terms of these toxicity classifications, the recommended dose-pair from the trial would be the dose-pair with greatest posterior probability of estimated toxicity being in the target toxicity interval.

We define a similar classification system for exposure with desirable exposures of drug i lying within $[L_i, U_i]$. Using single-agent data, U_i can typically be easily defined, but identifying L_i can be more difficult. This is because exposure levels corresponding to excessive toxicity are more easily identified from historical data than relationships with efficacy, which would be better suited to selecting the lower bound. So instead of directly defining the boundaries, a single, target exposure $E_i < U_i$ is defined for each drug. Doses with desirable exposures are then chosen to have exposure values within a certain percentage (20% for our evaluations) of this target level. As with toxicity, categorise exposure resulting from a dose-pair of drugs A and B such that: an undesirably low exposure has $PK(x_{\mathrm{A}}) < L_{\mathrm{A}}$ and $PK(x_{\mathrm{B}}) < L_{\mathrm{B}}$ and an undesirably high exposure has $PK(x_{\mathrm{A}}) > U_{\mathrm{A}}$ or $PK(x_{\mathrm{B}}) > U_{\mathrm{B}}$; this leaves a desirable exposure to result in at least one exposure in the target interval and neither exposure being greater than the corresponding upper limit. In relation to these exposure classifications, the recommeneded dose-pair from the trial would be the dose-pair that leads to posterior estimates $\widehat{PK}(x_{\mathrm{A}})$ and $\widehat{PK}(x_{\mathrm{B}})$ closest to their corresponding target exposures values.

From these classification systems for toxicity and exposure, it is clear that the instinctive definition of the recommended dose-pair is dependent upon the trial escalation and stopping rules. However, this does not mean that when exposure data is not considered during escalation that the definition of the recommended dose-pair based upon exposure classification is irrelevant. When there is no reliable prior information on target exposure values, or in the unlikely event that cycle 1 binary toxicity and exposure data are highly correlated, then there will be no benefit to considering exposure data. However, when there is prior knowledge linking exposure data to

long-term toxicity and/or effiacy, then this data should be considered for the benefit of patients and suitable drug development decisions to be made. Instead of defining a 'true' recommended dose-pair based on a combination of toxicity and efficacy, we have chosen to highlight the recommendations by each classification separately. This is done for clarity in comparisons of the methods and to highlight the difference in outcomes between scenarios.

In the remainder of this section some base decision rules, which are employed in each of four further methods, are described. The four methods each have specific decision rules (on top of the base ones) which are built on from Method 1 to Method 4. Method 1 is a simple method of dose-escalation concerned only with identifying the recommended dose-pair from toxicity data. Method 2 has more focus on patient safety and is a standard method of dual-agent model-based dose-escalation. The proposed method is presented as Method 3 and incorporates pharmacokinetic data into escalation decisions. Method 4 is equivalent to Method 3 except that it allows the study to stop based on sufficient precision about the recommended dose-pair. This final method is included to show that the proposed method (Method 3) is practical to employ in terms of the required number of patients. Results of a simulation study comparing these four methods are presented in Section 4.4.

The following base decision rules which constrain the step size in escalation and stop the trial for patient resources are employed in all methods:

i. Escalation rule:

- Escalate by a maximum of one dose-level of each drug from the dose-pair administered to the most recently treated cohort of patients.
  - This constraint is included to make escalation safer for patients by controlling the speed of escalation.

ii. Stopping rule:

- If 60 patients have been treated, stop the trial.
  - The recommended dose-pair is declared as the dose-pair which would be chosen for escalation out of those doses already administered in the trial, were the trial to continue.

Method 1: Optimise the probability of being in the target toxicity interval

This method is concerned only with optimising the probability of being in the target toxicity interval. This is achieved by using the following decision rule in addition to the base rules:

i. Escalation rule:

- Administer the dose-pair which **maximises the posterior probability**, $\mathbb{P}(\pi(x_{\mathrm{A}}, x_{\mathrm{B}}) \in [0.16, 0.35) | \{x_{\mathrm{A}}, x_{\mathrm{B}}\})$.

Within constraints of the base escalation rules, escalation under this method will occur rapidly to the dose-pair with maximum probability of toxicity in the target toxicity interval, out of available dose-pairs. No account is taken of sub-optimal dosing in

terms of toxicity or exposure. Escalation decisions are based soley on the dual-agent dose-toxicity model in Equation 4.2.2 with the corresponding normal priors described in Section 4.2.1.

Method 2: Optimise the probability of being in the target toxicity interval, within safety constraints

Patient safety is of priority in a dose-escalation trial and so it is intuitive for escalation to be restricted by some safety criteria. This method utilises a safety constraint. We define the safety criterion as only allowing escalation to dose-pairs with posterior probability of overdose less than 25%, mathematically $\mathbb{P}(\pi(x_{\mathrm{A}}, x_{\mathrm{B}}) \in [0.35, 1.00] | \{x_{\mathrm{A}}, x_{\mathrm{B}}\}) < 0.25$. Using the safety criterion in escalation implies an additional stopping rule when no dose-pairs satisfy the safety constraint. The decision rules for this method, in addition to the base rules, are therefore the following:

i. Escalation rule:

   - Administer the dose-pair which maximises the posterior probability, $\mathbb{P}(\pi(x_{\mathrm{A}}, x_{\mathrm{B}}) \in [0.16, 0.35) | \{x_{\mathrm{A}}, x_{\mathrm{B}}\})$;
   - **within dose-pairs which satisfy the safety criterion**.

ii. Stopping rule:

   - **If no dose-pairs satisfy the safety criterion**, stop the trial.
     - No recommended dose-pair declared.

As with Method 1, this method bases escalation decisions soley on the dual-agent dose-toxicity model in Equation 4.2.2 with the corresponding normal priors described

in Section 4.2.1. However, escalation under this method is more cautious, given the additional safety constraint on escalation. This design also enables the trial to stop for safety concerns if none of the available dose-pairs satisfy the safety criterion based on the available data. These additional benefits of the design over that of Method 1 are the reason that it is often used in practice.

Method 3: Use pharmacokinetic information to select doses, within safety constraints

On paper, Method 2 is used in current practice. However, in reality additional, non-formal decision rules are often used by the clinical team, enabling them to incorporate additional data without formalising its use. The subjectivity in decisions based on this additional data will lead to inefficiencies and inconsistencies in its use. In this chapter we are interested in formalising the use of pharmacokinetic data.

Pharmacokinetic information can often be an indicator of efficacy and/or long-term safety (Clark et al., 1994). It is reasonable to argue that if a suitable level of efficacy is reached, then it is unnecessary to escalate beyond this dose, risking greater toxicity. Also, unreasonably high exposure values should be avoided as they may indicate increased risk of toxicity being observed within, or after, the first cycle of treatment. Considering the benefit-risk ratio of the treatment in this way in dose-escalation trials can be beneficial to patient safety and increase the chance of the treatment being found to be efficacious and not overly toxic in later phase trials.

Safety of patients is a priority and for this reason, we maintain the toxicity safety constraint in our proposed design. The difference between this method and that of

Method 2 is that rather than escalation being based on optimising a toxicity criterion, we instead escalate based on an exposure criterion, within dose-pairs which satisfy the toxicity safety criterion. This is achieved using the following decision rules in addition to the base rules:

i. Escalation rule:

- Administer the dose-pair which **minimises the generalised squared inter point distance** (Deza and Deza, 2009) **of expected posterior exposure parameters from the target values**;
- within dose-pairs which satisfy the safety criterion.

ii. Stopping rule:

- If no dose-pairs satisfy the safety criterion, stop the trial.
  - No recommended dose-pair declared.

The distance measure used in this method is calculated as $g = \frac{\sum_{h=1}^{H} g_h}{H}$ for H iterations of a Markov chain, where

$$g_h = \sqrt{\left(\frac{\widehat{PK}(x_{\mathrm{A}h}) - E_{\mathrm{A}}}{\widehat{\sigma}_{\mathrm{A}h}}\right)^2 + \left(\frac{\widehat{PK}(x_{\mathrm{B}h}) - E_{\mathrm{B}}}{\widehat{\sigma}_{\mathrm{B}h}}\right)^2}$$

for $i = \{\mathrm{A}, \mathrm{B}\}$ with E_i, the target exposure values defined when classifying exposure data and $\widehat{PK}(x_{i\mathrm{A}})$ and $\widehat{\sigma}_{ih}$, the estimated exposure value and standard deviation in exposure at iteration h of the Markov chain.

In contrast to Methods 1 and 2, the escalation decisions in this method are based on toxicity and exposure criteria. The models for each of these relationships are fitted independently using Equations 4.2.2, 4.2.5 and 4.2.6 and corresponding priors described in Sections 4.2.1 and 4.2.2 for the dose-toxicity and dose-exposure models, respectively. The output of the independent models are then combined via the escalation rules.

Method 4: Allow for early stopping

This method uses exactly the same escalation rules as Method 3. The method of fitting the dose-toxicity and dose-exposure models independently using Equations 4.2.2, 4.2.5 and 4.2.6, with corresponding priors described in Sections 4.2.1 and 4.2.2, is therefore the same with the output of the independent models again combined via the escalation rules. The difference is that additional stopping rules are specified. This introduces the option for the trial to stop early, having identified the recommended dose-pair.

A sample size of 60 (as used in the previous methods) or greater would be desirable for a trial. However, this is not always feasible or necessary. Stopping rules can therefore be implemented which allow early stopping if the estimate of the recommended dose-pair is reasonably accurate. To achieve this, the following decision rules are used in addition to the base rules:

i. Escalation rule:

- Administer the dose-pair which minimises the generalised squared inter point distance of expected posterior exposure parameters from the target values;
- within dose-pairs which satisfy the safety criterion.

ii. Stopping rule:

- If no dose-pairs satisfy the safety criterion, stop the trial.
  - No recommended dose-pair declared.
- **If at the recommended dose for escalation**, criteria (a)-(c) are satisfied:
  (a) 9 patients have already been treated;
  (b) No higher adjoining dose-pair satisfies the safety criterion;
  (c) One or both of the following criteria are satisfied and is the highest among dose-pairs which satisfy the safety criterion:
    · **Toxicity stopping criterion**: The posterior probability of being in the target toxicity interval is greater than 0.70, that is

$$\mathbb{P}(\pi(x_{\mathrm{A}}, x_{\mathrm{B}}) \in [0.16, 0.35) | \{x_{\mathrm{A}}, x_{\mathrm{B}}\}) > 0.70$$

    · **Exposure stopping criterion**: The posterior probability of at least one drug having exposure within the desirable exposure interval and neither drug having undesirably high exposure, that is

$$\mathbb{P}(\{PK(x_{\mathrm{A}}) \in [L_{\mathrm{A}}, U_{\mathrm{A}}] \cup PK(x_{\mathrm{B}}) \in [L_{\mathrm{B}}, U_{\mathrm{B}}]\}$$
$$\cap \{PK(x_{\mathrm{A}}) < U_{\mathrm{A}}\} \cap \{PK(x_{\mathrm{B}}) < U_{\mathrm{B}}\} | \{x_{\mathrm{A}}, x_{\mathrm{B}}\}) > 0.25.$$

  - The recommended dose-pair is declared as the dose-pair which would be chosen for escalation, out of those doses already administered in the trial, were the trial to continue.

Design features, including the maximum sample size and tolerances for the safety criterion and stopping rules, used in the decision rules specified in this section, as well

as in the simulation example in the next section, are flexible. The maximum sample size of 60 patients was chosen as a desirable but not often feasible sample size to obtain an idea of long-term operating characteristics of the methods. The tolerances for stopping rules were then selected as values which produce desirable early stopping characteristics under the example scenario given in Section 4.4. Choices of doses, target values and distance measure used in the example were chosen as values felt to be reasonable based on available data. These values are flexible and can be adjusted based on available information and desired operating characteristics of a trial.

4.4 Simulation Study Results

Data from two single-agent trials was used as the basis for the simulation study presented in this section. Single-agent data for drug A was taken from Bristol-Myers Squibb (2007-2011) and that for drug B from Merck Sharp & Dohme Corp. (2009-2012). Single-agent dose-toxicity and dose-exposure models, as given in Equations 4.2.1 and 4.2.4 respectively, were fitted independently to the single-agent data in a frequentist manner. The resulting single-agent parameter estimates were used as the basis for priors on the parameters, after accounting for between-trial heterogeneity (which was discussed in Section 2.2.2 and is described in relation to the example used in simulations in Appendix 4.6.1). In this case, no information was available on the interaction between the two drugs and so weakly informative priors, centered on the case of no interaction, were taken for these parameters. The data and method of prior elicitation used to obtain the following prior distributions, as used in the simulation

study, are explained in detail in Appendix 4.6.1.

The priors for the dose-toxicity model are then given by:

$$\begin{aligned}
\begin{bmatrix} \log(\alpha_{\mathrm{A}}) \\ \log(\beta_{\mathrm{A}}) \end{bmatrix} &\sim \mathrm{MVN}_2\left(\begin{bmatrix} -5.65 \\ 1.81 \end{bmatrix}, \begin{bmatrix} 14.25 & 0 \\ 0 & 0.79 \end{bmatrix}\right), \\
\begin{bmatrix} \log(\alpha_{\mathrm{B}}) \\ \log(\beta_{\mathrm{B}}) \end{bmatrix} &\sim \mathrm{MVN}_2\left(\begin{bmatrix} -15.36 \\ 2.93 \end{bmatrix}, \begin{bmatrix} 91.90 & 0 \\ 0 & 0.41 \end{bmatrix}\right), \\
\zeta &\sim \mathrm{N}(0, 0.23)
\end{aligned}$$

and for the dose-exposure model by:

$$\begin{aligned}
\begin{bmatrix} \phi_{1\mathrm{A}} \\ \phi_{2\mathrm{A}} \\ \phi_{3\mathrm{A}} \end{bmatrix} &\sim \mathrm{MVN}_3\left(\begin{bmatrix} 2.35 \\ 4.87 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.23 & 0 & 0 \\ 0 & 0.58 & 0 \\ 0 & 0 & 0.47 \end{bmatrix}\right), \\
\begin{bmatrix} \phi_{1\mathrm{B}} \\ \phi_{2\mathrm{B}} \\ \phi_{3\mathrm{B}} \end{bmatrix} &\sim \mathrm{MVN}_3\left(\begin{bmatrix} 3.66 \\ 4.69 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.04 & 0 & 0 \\ 0 & 0.09 & 0 \\ 0 & 0 & 0.47 \end{bmatrix}\right), \\
1/\sigma_{\mathrm{A}}^2 &\sim \mathrm{Gamma}(15, 1/7.63), \\
1/\sigma_{\mathrm{B}}^2 &\sim \mathrm{Gamma}(15, 1/2.47).
\end{aligned}$$

Reference doses and target exposure values of the two drugs were selected at the single-agent maximum tolerated doses according to the single-agent models in Equations 4.2.1 and 4.2.4. Although the model does not require pre-specification of available doses, this was done for the purpose of simulation. Available doses were

selected for drug A as $d_{\mathrm{A}} = \{10, 15, 20, 25, 30\}$ with reference dose $d^*_{\mathrm{A}} = 25$ and target exposure value 300 and for drug B as $d_{\mathrm{B}} = \{20, 40, 60, 80, 100\}$ with reference dose $d^*_{\mathrm{B}} = 80$ and target exposure value $1,000$. The 'available dose-pairs' refer to any combination (one of drug A and one of drug B), of the available doses. The starting dose-pair was taken to be the lowest available dose-pair (10mg of drug A and 20mg of drug B in this case) and patients were treated in cohorts of size 3 for all simulations.

In the simulation study, toxicity and exposure data were generated from the models given in Equations 4.2.2, 4.2.5 and 4.2.6, with parameter values equal to their corresponding prior means, with the exception of the interaction parameters values which were varied depending upon the simulated scenario. A 'true' classification refers to the toxicity or exposure classification (defined at the start of Section 4.3) that the dose-pair of interest falls into. This is based upon the model specified and parameter values which data were simulated from. The 'true' recommended dose-pair is the dose-pair which optimises the escalation criteria under the specified models and parameter values. The models, with corresponding priors, were fitted to the data using the `Rstan` package (Stan Development Team, 2013) in *R* (R Core Team, 2014).

Results are presented for the following five scenarios based on estimates from $1,000$ simulated trials under the given scenario and method. The corresponding true probabilities of toxicity for each of the scenarios are given in Table 4.6.3 in Appendix 4.6.2:

1. No toxicity and no exposure interaction: This, perhaps unlikely, scenario is included for comparison of the methods in a scenario where the 'true' recom-

mended dose-pairs based on toxicity classification and that based on exposure classification are similar;

2. No toxicity interaction but a 4-fold increase in exposure to drug B at the reference doses: This scenario differs from Scenario 1 only in the exposure interaction. Although it appears to be an extreme scenario it is highly important as it represents the case of an unexpected dose-exposure interaction. If the pharmacokinetic data are not accounted for in dose-escalation, a dose-pair with suitable toxicity but high exposure could be identified as the recommended dose-pair. When exposure data are considered as an indicator for long-term safety concerns then this treatment at the recommended dose level could be found to be unsafe in later trials. Escalation following Methods 1 and 2, which base decisions on toxicity data alone, is not affected by this change of scenario;

3. A 3-fold increase in the odds of toxicity and a 2-fold increase in exposure of drug B at the reference doses: This is a more realistic scenario where there is some level of dose-toxicity and dose-exposure DDI. It is also included as a difficult scenario in terms of 'true' classifications to demonstrate the safety criterion;

4. A 2-fold increase in odds of toxicity at the reference doses but no exposure interaction: This scenario covers the case when the toxicity interaction is not directly driven by an exposure interaction. It is similar to Scenario 2, but this time the exposure escalation criteria of Method 3 will be pushing for escalation, based on links to efficacy maybe, but based on short-term toxicity those dose-pairs are not desirable;

5. A 10-fold increase in the odds of toxicity and a 5-fold increase in exposure of drug B at the reference doses with available dose range restricted to $d_{\mathrm{A}} = \{20, 25, 30\}$ and $d_{\mathrm{B}} = \{60, 80, 100\}$: This scenario is included to demonstrate the methods in a setting where no available dose-pairs have desirable safety characteristics. We want to ensure that the safety criteria are effective in such a case to reduce the number of patients treated with a highly toxic treatment.

Tables of the operating characteristics of the methods under these scenarios are given in Appendix 4.6.2. Initially we consider the results of Methods 1-3. Under these methods, dose-escalation continued until a total of 60 patients had been treated in the trial, unless (in Methods 2 and 3) the trial was stopped for safety before this point.

Consistency and accuracy of the recommended dose-pair

The proportion of times each available dose-pair was declared as the recommended dose-pair is given in Table 4.6.4 (in Appendix 4.6.2) and 'true' recommended dose-pair based on toxicity and exposure classification are marked. From this table we see that the recommendations by Method 1 and 2 are fairly similar. More recommendations by Method 1 are classified, based on the toxicity criteria, as being overdoses than recommendations from Method 2. The lack of safety criteria in Method 1 also means that escalation by this method is unethical.

From Table 4.6.4, we can see that the recommended dose-pairs from Method 3 were more consistent than those of Method 2. That is, the number of dose-pairs at which one or more simulated trials declared a recommended dose-pair was less, and more condensed, in Method 3 than those by Method 2. This is most noticeable in Scenario 2

where 97% of recommended dose-pairs were spread over only two dose-pairs, compared to nine dose-pairs by Method 2. This can be seen clearly for all scenarios in Table 4.6.4 and is comforting given that in reality we only have one attempt to identify the 'best' dose for patients. This improved consistency in recommended dose-pairs when exposure data are used stems from exposure data being continuous. The result is that escalation paths are more varied and escalation of both drugs, as opposed to escalation of one drug at a time, is more likely than when these decisions are based soley on toxicity data. This reduces the chance of escalation 'sticking' at a certain dose of one or both drugs.

As well as improvements in consistency of the recommended dose-pair, other benefits of Method 3 were observed. Under the setting of no DDIs (Scenario 1), the 'true' recommended dose-pair based on the toxicity classification and that based on exposure classification are similar (as can be seen in Table 4.6.4). Despite the similarity in location of the 'true' recommended dose-pairs, Method 3 led to a 6.4% decrease in the proportion of recommended dose-pairs in the target toxicity interval compared to Method 2. However, this compromise was for a 17.4% increase in the percentage of recommended dose-pairs with desirable exposures.

A similar pattern is seen in Scenario 2 when there was a stronger dose-exposure than dose-toxicity interaction. In this case, the 'true' recommended dose-pairs by the toxicity and exposure classification differ, but are both still within the target toxicity interval. Based on the toxicity classification, 6.7% more recommended dose-pairs were classed as under-doses by Method 2 than by Method 3. However, there was an 85.4% increase in the proportion of these dose-pairs proportion with desirable

exposure values between these two methods.

In Scenario 3, there was both a toxicity and an exposure interaction. The result was that the 'true' recommended dose-pair by the exposure classification edged into the overdose category (with true probability of DLT of 0.37). Under this scenario, Method 3 led to an increase in the proportion of recommended dose-pairs classified as overdoses because escalation was effectively targeting the defined ideal exposure values which occur at an overly toxic dose. In this case, the safety criterion was not effective in stopping escalation to the dose with true probability of DLT of 0.37 in either Method 2 or Method 3. However, in Scenario 4 for example, the 'true' recommended dose-pair by the exposure classification has probability of DLT of 0.42. In this case, the probability of DLT is great enough that in general the safety criterion is effective. In this scenario, recommended dose-pairs with target toxicity classification by Method 3 were actually slightly increased over those of Method 2 due to improved exploration of available dose-pairs.

Overdosing and undesirable exposures

From Figure 4.4.1 we see that on average 16-24% of patients in a trial experienced a DLT under Scenarios 1-4. Upon investigation, this value is reasonable because most of the observed DLTs occured at the recommended dose-pair. It can also be seen that the average proportion of toxicities per trial decreased from that in Method 1 as safety constraints (Method 2) and pharmacokinetic data (Method 3) were incorporated into the dose-escalation trial design. There was also a general decrease in the proportion of undesirably high exposure values observed. This was most noticeable in Scenario

2 when there was an exposure interaction but no toxicity interaction. Accounting for pharmacokinetic data in escalation therefore has a notable effect on escalation, which is reflected in the exposure values observed.

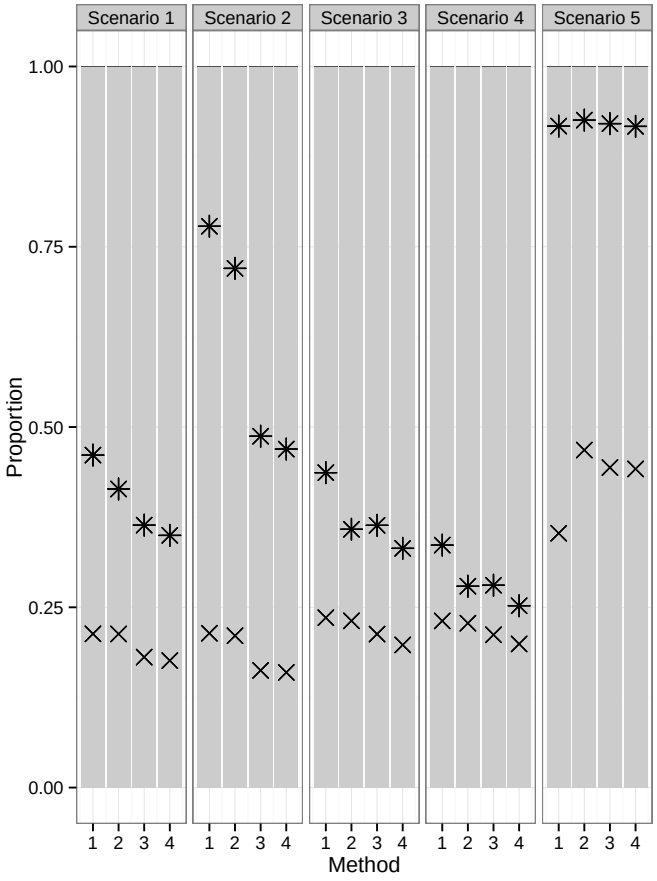

Figure 4.4.1: Average proportion of patients experiencing DLTs (marked by a cross) and undesirably high exposures (marked by a star) per trial under each dose-escalation method and scenario.

Table 4.6.5 (in Appendix 4.6.2) and Figure 4.4.1 show the operating characteristics of the trial simulations. Under Scenarios 1-4, Method 2, which employs the safety criterion, was observed to have some benefit over Method 1 in terms of administered doses and toxicity classification of the recommended dose-pair. This was most noticable in Scenario 3 where the percentage of recommended dose-pairs classified as an

overdose was reduced from 27.8% to 16.4% in Method 2 compared to Method 1. In Scenario 5, when no dose-pairs were tolerated by safety criteria, the safety criterion was effective in reducing the number of patients treated per trial. The number of trials which identified a recommended dose-pair classified as an overdose was reduced from 100% under Method 1 to 1.9% under Method 2. Methods 2 and 3 were comparable under this scenario because they employ the same safety criteria and corresponding stopping rule.

Using the safety criterion in escalation is therefore beneficial for patient safety, in the case of unexpectedly high exposure or a badly chosen dose range, without considerable compromise in identification of the recommended dose-pair. However, since Method 2 does not account for exposure data, there are high numbers of recommended dose-pairs with undesirably high exposure values, especially in Scenarios 1 and 2. These undesirably high exposures, as well as being undesirable, may indicate that an efficacious dose has already been reached. The additional risk to patients of administering higher doses is therefore unnecessary and unethical. Alternatively, it could indicate possible long-term safety concerns and in practice the exposure data could well be used subjectively to over-ride model recommendations. The reduction in undersirably high exposure levels experienced by patients was especially clear in Scenario 2 where 87.7% of administered dose-pairs by Method 2 had a true exposure classification of undesirably high.

The observed decrease in undesirably high exposures and average proportion of DLTs patients experienced in a trial under Method 3 compared to Method 2 was due to escalation being more cautious when exposure data were considered. This is

due to the non-binary nature of exposure data leading to increased exploration of the available dose-pairs.

One of the big assumptions made in the above evaluation was that the prior perfectly reflected the truth under which data were generated. A detailed sensitivity analysis of Method 3 to prior specification (details of which are given in Appendix 4.6.3), however, showed that the method was found to be robust to priors deviating from the true models. For example, the proportion of recommended dose-pairs classified as being in the target toxicity interval under Method 3 in Scenario 1 was reduced from 77.1% to 67.1% based on a prior with only one tolerated start dose and half the variance of that given at the start of this section.

Method 4: Allow for early stopping

The benefits observed from using pharmacokinetic data in escalation, such as the improved consistency of the recommended dose-pair and general reduction in proportion of patients experiencing DLTs and undesirably high exposures in a trial, are only beneficial if the trial is practical to carry out. Method 4 considers the practicality of the trial design in terms of the number of patients treated in the trial. Early stopping of the trial (before the maximum of 60 patients have been treated) for accuracy of the estimate, as well as for safety, was allowed in this method. The additional stopping rules were based on a high probability of either toxicity or exposure being in the corresponding target interval, with no option to escalate to a higher dose-pair. Under the stopping rules specified in Section 4.3, only small losses in operating characteristics were observed from those observed in Method 3. This suggests that the stopping rules

used were reasonable.

Figure 4.4.2 shows the proportion of times each stopping rule was met under each scenario using Method 4. In Scenario 5, most trials stopped for safety due to the lack of available tolerated dose-pair. In Scenarios 1, 3 and 4, practical use of the early stopping rules is more clear. The toxicity and exposure stopping criteria are met a reasonable number of times, in between 41% and 75% of trials in Scenarios 1-4. These values are reflected in the average number of patients treated per trial, presented in Table 4.6.5. The average number of patients required per trial is down to about 34 patients in Scenario 3 and around 40 patients in Scenarios 1, 2 and 4.

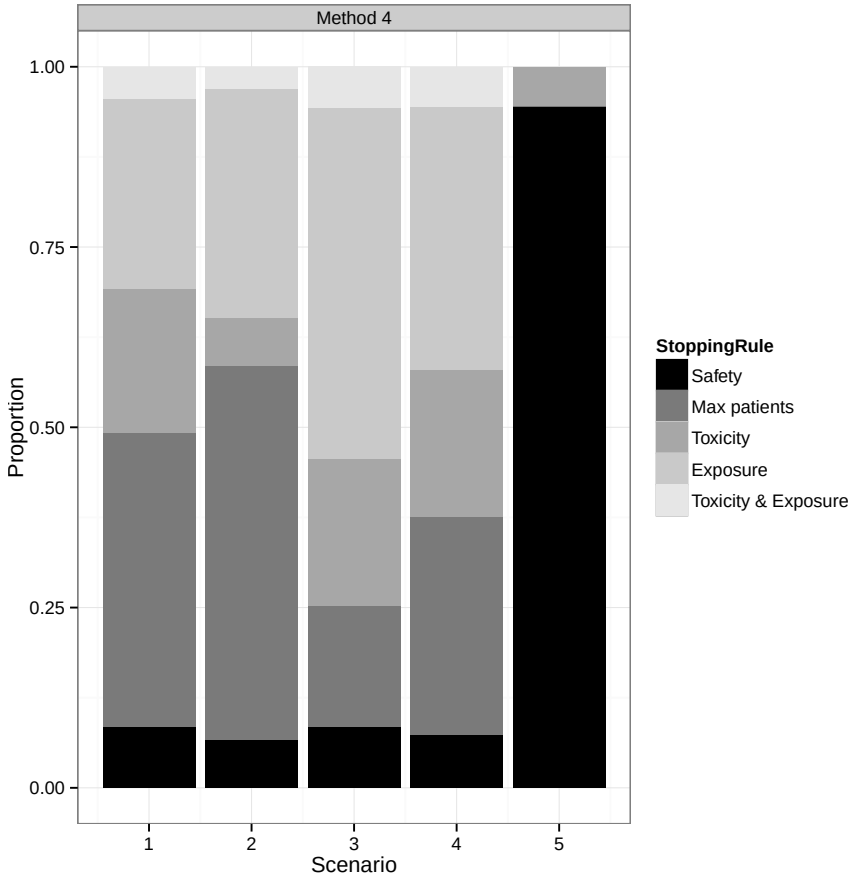

Figure 4.4.2: Reasons trial stopped under dose-escalation Method 4 for the given scenarios.

In Scenario 2, the 'true' recommended dose-pair based on the toxicity classification is higher than that targeted by exposure classification (this can be seen from Table 4.6.3 in Appendix 4.6.2). The toxicity stopping criterion is therefore highly unlikely to be met. Therefore, stopping rules such as those employed in Method 4 which are based around the precision of estimates, effectively make the trial size practical when the interaction scenario and available dose levels allow this.

4.5 Discussion

We proposed a method of dose-escalation (presented as Method 3) for a dual-agent treatment which, through the escalation rules specified, enables formal integration of exposure information into dose-escalation decisions. The specific escalation rules used in this chapter illustrate the design but these could be adjusted to cater for a specific trial. Exposure data is typically available during dose-escalation trials but is only used in a subjective manner. When prior knowledge links exposure data to long-term toxicity and/or effiacy then pharamcokinetic data should be considered for the safety and benefit of trial and future patients. The novel method is relatively simple to implement and simulation results show good operating characteristics. Early stopping of the trial (presented as Method 4), for safety concerns or accuracy of the estimate, was also investigated and results show that the method is practical to employ.

In this work, we have used a two-parameter model for the dose-toxicity relationship. This was chosen over a one-parameter model because of its flexibility to change the target toxicity level corresponding to the recommended dose-pair. Although the

target toxicity level is fixed throughout the trial, other considerations or new data can lead to the target toxicity level being changed from that originally specified after the trial has been conducted. For example, the target toxicity interval could be changed from $[0.16, 0.35)$ to a lower interval such as $[0.10, 0.30)$. This is a request that we have experienced on multiple occasions in practice. Additionally, more than one dose can be taken to further trials from a dose-escalation trial (be this further phase I trials such as dose expansion, or initial efficacy trials). In such a case, it is beneficial to be able to obtain reasonable estimates of the probability of toxicity for a range of doses below the recommended dose. This is to reduce the chance that any of the doses taken for further testing has too low toxicity (and hence low efficacy).

Simulation results showed that formal incorporation of exposure data into dose-escalation decisions can lead to a decrease in the proportion of patients who experience toxicities, and generally also undesirably high exposures, within the trial. In addition, the continuous nature of exposure data makes escalation along the diagonal of available dose-pairs more likely and means that escalation is unlikely to stick on a dose level of one or both drugs, as can occur when toxicity data alone is considered. This dose-sticking was apparent when dose-escalation patterns in individual trials were investigated. This property can be especially beneficial when no/few DLTs are observed early on in the dose-escalation trial, as is often the case. The result is that dose-recommendations are more compacted around suitable dose-pairs when pharmacokinetic data is utilised, along with toxicity data, in escalation decisions.

The proposed method is flexible and practical since it can be used throughout escalation, even in cases where pharamcokinetic data is delayed, for example. In practice

it is also still possible for the clinical team to over-ride the model decision based on any available data. The proposed method was presented for the case where 'ideal' values of the exposure parameters had been identified and were in effect targeted, within dose-pairs classified as safe by the safety criterion. The resulting recommended dose-pair is therefore hoped to have an improved benefit-risk ratio over the dose-pair selected as having highest toxicity within doses which satisfy the safety criterion. Equally, a pharmacodynamic response could be used in place of the exposure parameter. Another simple alternative, which in some cases may better model the dose-exposure relationship, is to model the exposure interaction in terms of exposure, rather than dose, of the second drug. That is, in place of Equations 4.2.5 and 4.2.6, for reference exposure values PK_A^* and PK_B^* using the models.

$$\begin{aligned} \log(PK_A) &\sim \mathrm{N}\left(\phi_{1A} + \phi_{2A}\log\left(\frac{x_A}{d_A^*}+1\right) + \phi_{3A}\log\left(\frac{PK_B}{PK_B^*}+1\right), \sigma_A^2\right) \\ \text{and} \quad \log(PK_B) &\sim \mathrm{N}\left(\phi_{1B} + \phi_{2B}\log\left(\frac{x_B}{d_B^*}+1\right) + \phi_{3B}\log\left(\frac{PK_A}{PK_A^*}+1\right), \sigma_B^2\right). \end{aligned}$$

A hierarchical model with probability of toxicity defined as a function of exposure, in turn defined as a function of dose, may better model the effect of dose on toxicity. Use of this model in dose-escalation enables decision rules to be based on toxicity criteria alone, while formally including pharmacokinetic data in the model itself. This could be advantageous if people were adverse to making escalation decisions on exposure data, as required in our proposed method. However, the hierarchical model requires direct modelling of the correlation between toxicity and exposure. Our proposed method of incorporating pharmacokinetic data using independent dose-toxicity

and dose-exposure models does not require such a strong assumption, and is therefore less prone to model mis-specification. On top of this it is computationally much simpler than a hierarchical model.

4.6 Appendix

4.6.1 Using Single-agent Data for Prior Derivation

The historical single-agent trial data used in Section 4.3 to illustrate the proposed dual-agent dose-escalation trial design came from Bristol-Myers Squibb (2007-2011) and Merck Sharp & Dohme Corp. (2009-2012) for drugs A and B respectively. The relevant historical data from these publications which was used to derive priors on the single-agent parameters of the dual-agent trial are presented in Tables 4.6.1 and 4.6.2. Where we were not able to obtain the required data for these tables directly, the derivation is given here. The exposure values of both treatments were provided as the summary statistics shown in the tables with no additional information on their method of calculation.

Historical single-agent trial data

In the single-agent trial of drug A, drug A was administered once daily on days 1-5 of a 21 day cycle. Escalation followed a $3 + 3$ design with identification of the MTD at 25mg. An additional dose-expansion cohort was then treated at this dose, resulting in a total of 44 patients being treated in the trial. The AUC over the 24 hour period on day 5 of cycle 1 was available for all 44 patients and was the pharmacokinetic parameter of drug A chosen for use in the dual-agent trial.

DLT data provided for drug A is given in Merck Sharp & Dohme Corp. (2009-2012) in terms of the total number of DLTs experienced at each dose level. For example, a cohort of three patients from which two patients did not experience a DLT, while

the other experienced two DLTs, is recorded as two DLTs in three patients. For our evaluations we were instead interested in the number of patients who experienced a DLT at each dose-level. So, for the example, we record one out of three patients having experienced a DLT. In Table 4.6.1, the derived values for the number of patients who experienced a DLT at each dose level is given. These were calculated based upon a $3 + 3$ design. At dose levels where it was unclear how many patients experienced a DLT, the highest option was used.

Dose drug A	5	10	15	20	25	30
Number of patients	3	6	3	3	23	6
Number of DLT's	0	1	0	0	2	4
$AUC_{(0-24)}$ on day 5 mean (sd)	21.83 (28.59)	143.85 (134.15)	179.82 (126.95)	222.73 (136.49)	348.07 (370.95)	545.42 (441.93)

Table 4.6.1: Data used to obtain priors for drug A (obtained/derived from Bristol-Myers Squibb, 2007-2011).

In the single-agent trial of drug B, drug B was administered once daily on days $1, 3, 8, 10, 15$ and 17 of a 28 day cycle. We assumed that escalation followed a $3 + 3$ design resulting in declaration of the MTD at 80mg. This was based upon the following logic: before the multiple dosing trial, a single-dosing investigation, in which patients were treated at $20, 40, 80$ and 120mg, was carried out. We expect that the multiple dosing trial had the same planned doses as the single-dosing trial. We expected that escalation proceeded as planned to 120mg, at which dose an undesirable number of toxicities was observed. Instead of de-escalating to 80mg, we suspected that an additional dose level was introduced at 100mg. This is the only dose level for which cohorts do not appear to be of size three. We expected that observation of two

DLTs in two patients treated at 100mg led to a decision not to recruit any further patients at this dose level. The MTD was therefore declared at 80mg.

For drug B, the AUC over the 24 hour period on day 1 of each cycle was available for all 17 patients enrolled in the trial. The summary statistics were for all day 1 AUC values and so these are effectively based upon a total of 51 observations. This measure may not be overly helpful for understanding the pharmacokinetics of the drug but can be used for simulations by assuming that the dual-agent trial records the same data. We also assumed that both drugs have a 21 day cycle and observations of DLTs for drug B over the reduced time period would be the same as those recorded in the historical trial.

Dose drug B	20	40	80	100	120
Number of patients	3	3	6	2	3
Number of DLT's	0	0	0	2	2
$AUC_{(0-24)}$ on day 1 of each week mean (sd)	122 (85.3)	302 (47.7)	936 (364.0)	2530 (81.2)	2320 (673.0)

Table 4.6.2: Data used to obtain priors for drug B (obtained/derived from Merck Sharp & Dohme Corp., 2009-2012).

Obtaining prior distributions

The historical, single-agent data was fitted, in a frequentist manner, to the dose-response models given in Equations 4.2.1 and 4.2.4. We have no additional beliefs to incorporate and so we felt the additional complication and subjectivity of a Bayesian model fit to be unnecessary in this case. For the dose-toxicity model, a logistic regression model was fitted to the historical toxicity data to obtain regression coefficient estimates and corresponding standard deviations. For the exposure data of both

drugs, only summary values of the exposure parameters were available. To overcome this, data were simulated to obtain estimates for the mean and standard deviations of the dose-exposure model parameters. To obtain mean estimates of the dose-exposure model parameters, $1,000$ data points were simulated at each dose and a linear regression of $\log(PK)$ against log(dose) fitted. To obtain a reasonable estimate of the prior standard deviations of the dose-exposure parameters, data were again simulated at each dose, but this time only for the number of patients observed at each dose.

The regression coefficient estimates for the parameters of the single-agent dose-toxicity and dose-exposure models were used directly as the mean prior parameter values. The estimated variances, however, were increased to account for heterogeneity between trials. This was achieved using a power prior (Ibrahim and Chen, 2000). In a power prior, the likelihood of the historical data is raised to some power, $a \in [0, 1]$, effectively down-weighting the historical data in relation to the trial data. In the case of a normal likelihood with known variance, this is equivalent to increasing the variance by a factor of $1/a$. In Section 2.2.2 under Step 4, the reasons for accounting form heterogeneity and the choice of weight of the historical data were discussed. Since we expect the single-agent data to be highly relevant to the action of the treatments in combination, and given that the dose-toxicity model is specified as a combination of two single-agent models, we assumed relatively low between-trial heterogeneity. We so chose to make the prior data worth about $2/3$ of the dual-agent data. So, assuming we have normally distributed data with known variance, we increased the variances of the historical posterior parameter estimates by a heterogeneity factor of $3/2$. The prior standard deviations were used as those obtained from the regression fits, multiplied

by the heterogeneity factor. In addition, the prior correlation between parameters of the models was set to 0. We chose to do this in order to further reduce the prior information (since we are assuming relatively low between-trial heterogeneity). Of course, covariance estimates obtained from the model fitted to historical data could be used here instead.

The dose-exposure model also requires a prior distribution to be specified on the parameter for inter-patient variability. The residual standard error (from the regression fit of the dose-exposure model which was used to estimate variance of the relevant parameters) was set as the mean of this distribution. The variance of the prior distribution was adjusted until the 95^{th} percentile of the distribution was equal to residual standard error multiplied by 3/2 (the heterogeneity factor). The result in our case was a coefficient of variation of 25% around the exposure data mean. Target exposure values were identified as the expected exposures based on this linear regression fit at the single-agent maximum tolerated doses (300 for drug A and $1,000$ for drug B).

The plots in Figure 4.6.1 show the resulting prior distributions. The corresponding prior distributions are given in Section 4.4, along with those on the interaction parameters. No information was available on the interaction parameters and so we specified normal priors on them centred at 0. The variance was chosen so that the 99^{th} percentile of the distribution corresponded to a 3-fold DDI at the reference doses (25mg for drug A and 80mg for drug B).

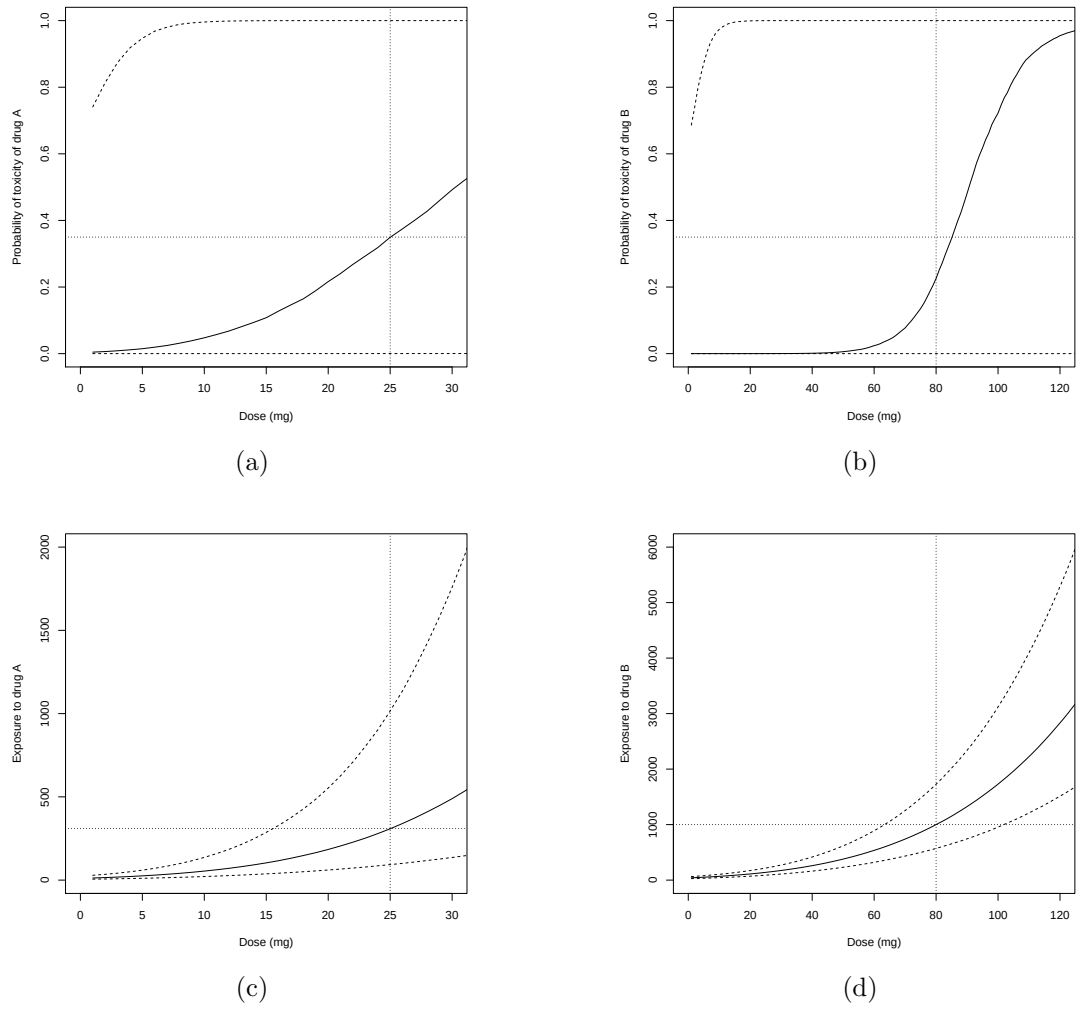

Figure 4.6.1: Median and 90% credible interval for priors on the single-agent models for a) dose-toxicity relationship of drug A, b) dose-toxicity relationship of drug B, c) dose-exposure relationship of drug A, d) dose-exposure relationship of drug B. Dotted lines indicate the recommended dose based on single-agent data.

4.6.2 Results Tables

			Probability of toxicity					PK of drug A					PK of drug B				
			Dose of drug B					Dose of drug B					Dose of drug B				
			20	40	60	80	100	20	40	60	80	100	20	40	60	80	100
Scenario 1	Dose of drug A	10	0.03	0.03	0.03	0.11	0.47	54.0	54.0	54.0	54.0	54.0	111.3	260.7	535.6	999.2	1732.0
		15	0.06	0.06	0.07	0.14	0.49	103.4	103.4	103.4	103.4	103.4	111.3	260.7	535.6	999.2	1732.0
		20	0.11	0.11	0.12	0.19	0.52	183.6	183.6	183.6	183.6	183.6	111.3	260.7	535.6	999.2	1732.0
		25	0.20	0.20	0.20	0.26*	0.56	306.6	306.6	306.6	306.6	306.6	111.3	260.7	535.6	999.2	1732.0
		30	0.30	0.30	0.31[X]	0.36	0.62	487.7	487.7	487.7	487.7	487.7	111.3	260.7	535.6	999.2	1732.0
Scenario 2	Dose of drug A	10	0.03	0.03	0.03	0.11	0.47	54.0	54.0	54.0	54.0	54.0	218.1	511.1	1049.8	1958.5	3394.8
		15	0.06	0.06	0.07	0.14	0.49	103.4	103.4	103.4	103.4	103.4	284.9	667.5	1371.2	2558.2	4434.0
		20	0.11	0.11	0.12	0.19	0.52	183.6	183.6	183.6	183.6	183.6	360.6	844.8	1735.4	3237.6	5611.8
		25	0.20	0.20*	0.20	0.26	0.56	306.6	306.6	306.6	306.6	306.6	445.1	1043.0	2142.4	3997.0	6928.1
		30	0.30	0.30	0.31[X]	0.36	0.62	487.7	487.7	487.7	487.7	487.7	538.6	1262.0	2592.4	4836.3	8383.0
Scenario 3	Dose of drug A	10	0.03	0.03	0.05	0.16	0.61	54.0	54.0	54.0	54.0	54.0	155.8	365.0	749.9	1398.9	2424.8
		15	0.07	0.08	0.10	0.24	0.69	103.4	103.4	103.4	103.4	103.4	178.0	417.2	857.0	1598.8	2771.2
		20	0.14	0.17	0.21	0.36	0.76	183.6	183.6	183.6	183.6	183.6	200.3	469.3	964.1	1798.6	3117.6
		25	0.24	0.30[X]	0.37*	0.52	0.84	306.6	306.6	306.6	306.6	306.6	222.6	521.5	1071.2	1998.5	3464.0
		30	0.38	0.46	0.55	0.68	0.90	487.7	487.7	487.7	487.7	487.7	244.8	573.6	1178.3	2198.3	3810.5
Scenario 4	Dose of drug A	10	0.03	0.03	0.04	0.14	0.56	54.0	54.0	54.0	54.0	54.0	111.3	260.7	535.6	999.2	1732.0
		15	0.06	0.07	0.09	0.20	0.62	103.4	103.4	103.4	103.4	103.4	111.3	260.7	535.6	999.2	1732.0
		20	0.13	0.14	0.17	0.29	0.68	183.6	183.6	183.6	183.6	183.6	111.3	260.7	535.6	999.2	1732.0
		25	0.22	0.26	0.30[X]	0.42*	0.75	306.6	306.6	306.6	306.6	306.6	111.3	260.7	535.6	999.2	1732.0
		30	0.35	0.40	0.45	0.57	0.82	487.7	487.7	487.7	487.7	487.7	111.3	260.7	535.6	999.2	1732.0
Scenario 5	Dose of drug A	10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		15	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		20	-	-*	0.35	0.59	0.91	-	-	183.6	183.6	183.6	-	-	2096.9	3912.0	6780.8
		25	-X	-	0.59	0.78	0.96	-	-	306.6	306.6	306.6	-	-	2678.1	4996.2	8660.1
		30	-	-	0.78	0.90	0.98	-	-	487.7	487.7	487.7	-	-	3341.4	6233.8	10805.2

Table 4.6.3: Tables of the 'true' probability of toxicity and exposure used to generate data for each scenario. Dark grey cells highlight dose-pairs with toxicity/exposure category overdose, light grey cells the target interval and white cells underdoses. The 'X' and '*' mark the 'true' recommended dose-pair for each scenario based solely on toxicity and exposure data respectively. A '-' indicates that a dose-pair was not available under the given scenario.

			Escalation method 1					Escalation method 2					Escalation method 3					Escalation method 4				
			Dose of drug B					Dose of drug B					Dose of drug B					Dose of drug B				
			20	40	60	80	100	20	40	60	80	100	20	40	60	80	100	20	40	60	80	100
Scenario 1	Dose of drug A	10	0	0	0	0.002	0	0	0	0	0.001	0	0	0	0	0	0	0	0	0	0.003	0
		15	0	0.001	0	0.001	0	0	0.001	0.001	0.011	0.001	0	0	0	0.024	0	0.001	0.001	0.002	0.021	0
		20	0.005	0.013	0.018	0.051	0	0	0.013	0.045	0.068	0.003	0.004	0.038	0.161	0.209	0	0	0.037	0.130	0.160	0
		25	0.028	0.088	0.131	0.327*	0	0.003	0.082	0.166	0.290*	0.001	0.001	0.021	0.091	0.450*	0	0.004	0.031	0.113	0.490*	0
		30	0.023	0.084	0.113^{X}	0.106	0	0.005	0.082	0.138^{X}	0.088	0	0	0	0^{X}	0.002	0	0	0	0.001^{X}	0.007	0
Scenario 2	Dose of drug A	10	0	0	0	0.002	0	0	0	0	0.001	0	0	0	0.001	0	0	0	0	0.001	0	0
		15	0	0	0	0.002	0	0	0.002	0.003	0.009	0.002	0	0.016	0.003	0	0	0	0.014	0	0	0
		20	0.007	0.066	0.016	0.061	0	0.001	0.023	0.040	0.056	0.002	0.002	0.123	0	0	0	0	0.161	0	0	0
		25	0.024	0.101*	0.135	0.334	0	0.006	0.090*	0.160	0.263	0	0.003	0.846*	0	0	0	0.008	0.812*	0	0	0
		30	0.021	0.072	0.108^{X}	0.105	0	0.003	0.098	0.144^{X}	0.096	0	0.002	0.003	0^{X}	0	0	0.002	0.002	0^{X}	0	0
Scenario 3	Dose of drug A	10	0	0	0	0.012	0	0	0	0.001	0.004	0	0	0	0.006	0	0	0	0	0.003	0	0
		15	0	0.002	0.009	0.027	0	0.002	0.018	0.048	0.050	0	0	0.003	0.075	0	0	0.001	0.009	0.083	0	0
		20	0.018	0.092	0.368	0.115	0	0.003	0.124	0.428	0.058	0	0.002	0.042	0.594	0	0	0.003	0.055	0.489	0	0
		25	0.052	0.142^{X}	0.149*	0.002	0	0.011	0.146^{X}	0.091*	0	0	0.010	0.062^{X}	0.200*	0	0	0.013	0.107^{X}	0.234*	0	0
		30	0.007	0.004	0.001	0	0	0.003	0.009	0.002	0	0	0.004	0.001	0	0	0	0.003	0	0	0	0
Scenario 4	Dose of drug A	10	0	0	0	0.003	0	0	0	0	0.004	0	0	0	0.001	0.003	0	0	0	0	0.003	0
		15	0	0.003	0	0.008	0	0	0	0.013	0.017	0	0	0.001	0.018	0.087	0	0	0.001	0.011	0.079	0
		20	0.011	0.040	0.196	0.202	0	0.001	0.083	0.297	0.124	0	0.011	0.055	0.363	0.302	0	0.005	0.046	0.290	0.272	0
		25	0.044	0.141	0.240^{X}	0.063*	0	0.014	0.143	0.234^{X}	0.021*	0	0.007	0.024	0.082^{X}	0.044*	0	0.005	0.042	0.160^{X}	0.083*	0
		30	0.013	0.021	0.015	0	0	0.003	0.028	0.016	0	0	0.001	0	0	0	0	0	0.001	0	0	0
Scenario 5	Dose of drug A	10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		15	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		20	-	-*	1	0	0	-	-*	1	0	0	-	-*	1	0	0	-	-*	1	0	0
		25	-X	-	0	0	0	-X	-	0	0	0	-X	-	0	0	0	-X	-	0	0	0
		30	-	-	0	0	0	-	-	0	0	0	-	-	0	0	0	-	-	0	0	0

Table 4.6.4: Proportion of times each available dose-pair declared as the recommended dose-pair, out of those trials which identified a recommended dose-pair, under each dose-escalation method and scenario. Dark grey cells highlight dose-pairs with toxicity category overdose, light grey cells the target interval and white cells underdoses. The 'X' and '*' mark the 'true' recommended dose-pairs for each scenario based solely on toxicity and exposure data respectively. A '-' indicates that a dose-pair was not available under the given scenario.

Scenario	Escalation method	Average number of patients per trial	Proportion of trials identified a dose recommendation	Average proportion recommended dose-pair's with true classification					
				by toxicity			by exposure of drugs A and B		
				under-dose	target interval	over-dose	undesirably low	desirable	undesirably high
1	1	60	1	0.040	0.845	0.115	0.037	0.628	0.335
	2	56.295	0.935	0.072	0.835	0.093	0.060	0.621	0.319
	3	55.725	0.925	0.227	0.771	0.002	0.203	0.795	0.002
	4	42.669	0.915	0.196	0.798	0.007	0.172	0.821	0.008
2	1	60	1	0.033	0.856	0.111	0.007	0.131	0.862
	2	56.010	0.930	0.078	0.822	0.100	0.003	0.119	0.877
	3	55.953	0.929	0.145	0.855	0	0.018	0.973	0.009
	4	43.656	0.933	0.176	0.824	0	0.014	0.982	0.004
3	1	60	1	0.041	0.681	0.278	0.112	0.720	0.168
	2	54.129	0.897	0.077	0.759	0.164	0.148	0.725	0.127
	3	54.129	0.897	0.086	0.709	0.205	0.054	0.941	0.006
	4	33.735	0.916	0.099	0.664	0.237	0.071	0.926	0.003
4	1	60	1	0.057	0.844	0.099	0.250	0.701	0.049
	2	55.212	0.916	0.102	0.833	0.066	0.394	0.558	0.048
	3	55.611	0.923	0.090	0.866	0.044	0.450	0.549	0.001
	4	40.320	0.926	0.067	0.849	0.084	0.354	0.645	0.001
5	1	60	1	0	0	1	0	0	1
	2	6.675	0.019	0	0	1	0	0	1
	3	7.215	0.026	0	0	1	0	0	1
	4	5.937	0.056	0	0	1	0	0	1

Table 4.6.5: Average number of patients observed per trial and proportion of dose-pairs identified as the recommended dose-pair in each of the defined toxicity and exposure categories under each dose-escalation method and scenario.

4.6.3 Sensitivity Analysis

In the main simulation study presented in Section 4.4, a series of scenarios with differing DDIs are presented. These scenarios were all run using the same prior. That is, priors on the single-agent parameters were derived from historical single-agent data and priors on the interaction parameters were centered on the case of no DDI. The dose-response models update well for the different scenarios (i.e. observed interactions), which is reflected in the recommended dose-pairs.

In the main simulation study, simulated data were based on prior means for the single-agent parameters. Here we consider the sensitivity of the dose-response models to the prior on the single-agent parameters. That is, we investigate the cases when simulated data is not generated from distributions with means equal to the prior means, as in practice.

The largest effect of the prior on escalation occurs when an informative, but incorrect, prior is specifed. That is, when the prior parameter values are different to those observed in practice and the prior variance of the parameter estimates is small. The following four prior settings which cover the extremes of the prior specification on single-agent model parameters (i.e. $\log(\alpha_A)$, $\log(\beta_A)$, $\log(\alpha_B)$, $\log(\beta_B)$, ϕ_{1A}, ϕ_{2A}, ϕ_{1B} and ϕ_{2B}) are considered;

1. Prior specified such that **only the lowest dose-pair** is tolerated based on toxicity criteria and the lowest dose-pair also has target exposure values based on the prior.

   (a) With parameter **variances set to be the same** as those used in main

simulation study.

(b) With parameter **variances set to be half** those used in main simulation study.

2. Prior specified such that **all dose-pairs** are tolerated based on toxicity criteria and the highest dose-pair also has target exposure values based on the prior.

   (a) With parameter **variances set to be the same** as those used in main simulation study.

   (b) With parameter **variances set to be half** those used in main simulation study.

The prior probability of toxicity for each dose combination based on the two settings for the prior mean are given in Table 4.6.6. Simulations were carried out for the proposed method (Method 3) with the four prior settings for Scenario 1 (the case of no interaction) and Scenario 3 (a 3-fold increase in the odds of toxicity and a 2-fold increase in the exposure of drug B). The results are presented in Figure 4.6.2 and Tables 4.6.7 and 4.6.8.

			Probability of toxicity					PK of drug A					PK of drug B				
			Dose of drug B					Dose of drug B					Dose of drug B				
			20	40	60	80	100	20	40	60	80	100	20	40	60	80	100
Prior setting 1	Dose of drug A	10	0.16	0.35	0.73	0.93	0.98	294	294	294	294	294	933	2787	2028	15660	31748
		15	0.45	0.58	0.83	0.95	0.99	573	573	573	573	573	933	2787	2028	15660	31748
		20	0.77	0.82	0.93	0.98	1.00	1032	1032	1032	1032	1032	933	2787	2028	15660	31748
		25	0.92	0.94	0.98	0.99	1.00	1747	1747	1747	1747	1747	933	2787	2028	15660	31748
		30	0.97	0.98	0.99	1.00	1.00	2814	2814	2814	2814	2814	933	2787	2028	15660	31748
Prior setting 2	Dose of drug A	10	0.01	0.01	0.01	0.02	0.14	40	40	40	40	40	61	153	330	643	1158
		15	0.03	0.03	0.03	0.04	0.15	77	77	77	77	77	61	153	330	643	1158
		20	0.07	0.07	0.07	0.08	0.19	140	140	140	140	140	61	153	330	643	1158
		25	0.13	0.13	0.13	0.14	0.24	236	236	236	236	236	61	153	330	643	1158
		30	0.24	0.24	0.24	0.25	0.33	381	381	381	381	381	61	153	330	643	1158

Table 4.6.6: Tables of the prior probability of toxicity. Dark grey cells highlight dose-pairs with toxicity/exposure category overdose, light grey cells the target interval and white cells underdoses. The 'X' and '*' mark the 'true' recommended dose-pair for each scenario based solely on toxicity and exposure data respectively.

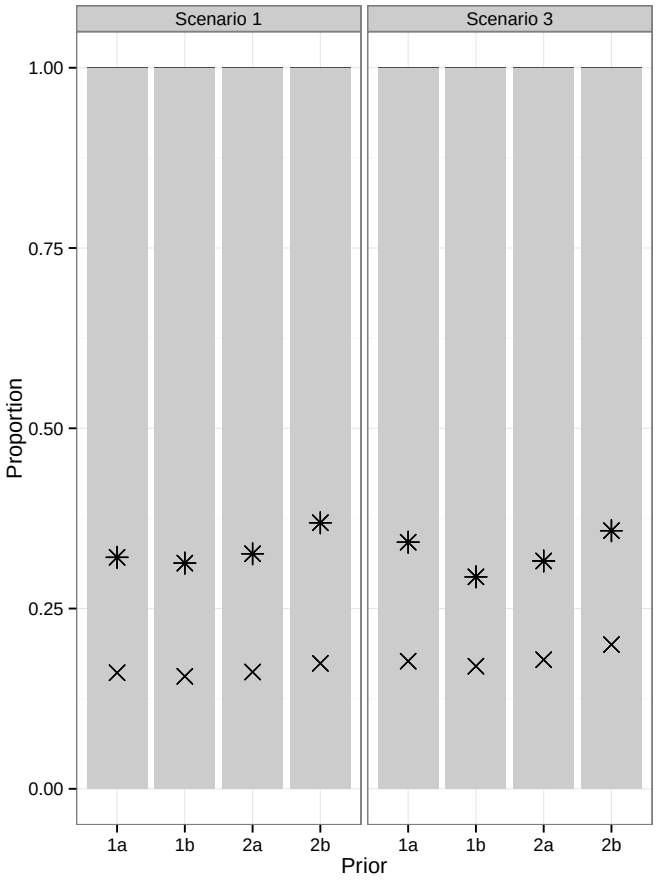

Figure 4.6.2: Average proportion of patients experiencing DLTs (marked by a cross) and undesirably high exposures (marked by a star) per trial under dose-escalation Method 3 with a range of prior settings for Scenarios 1 and 3.

			Prior setting 1a					Prior setting 1b					Prior setting 2a					Prior setting 2b				
			Dose of drug B					Dose of drug B					Dose of drug B					Dose of drug B				
			20	40	60	80	100	20	40	60	80	100	20	40	60	80	100	20	40	60	80	100
Scenario 1	Dose of drug A	10	0.005	0.009	0.004	0.009	0	0.001	0.012	0.008	0.007	0	0	0.006	0.005	0.002	0	0	0	0	0	0
		15	0	0.007	0.077	0.131	0	0	0.007	0.093	0.139	0	0	0.005	0.062	0.094	0	0	0.002	0.006	0.043	0
		20	0	0	0.080	0.378	0	0	0	0.062	0.423	0	0	0.005	0.074	0.320	0	0	0.011	0.082	0.188	0
		25	0	0	0	0.301*	0	0	0	0	0.248*	0.001	0.002	0.024	0.081	0.312*	0.002	0.003	0.053	0.186	00422*	0
		30	0	0	0^{X}	0	0	0	0	0^{X}	0	0	0	0.001	0^{X}	0.002	0	0	0.001	0^{X}	0.003	0
Scenario 3	Dose of drug A	10	0.006	0.004	0.008	0	0	0.007	0.007	0.015	0	0	0.005	0	0.013	0	0	0	0	0.008	0	0
		15	0.003	0.024	0.110	0.018	0	0.007	0.063	0.222	0.001	0	0.002	0.050	0.172	0.002	0	0.001	0.007	0.088	0	0
		20	0	0.028	0.301	0.090	0	0	0.047	0.538	0	0	0.003	0.057	0.517	0	0	0.004	0.059	0.578	0	0
		25	0	0.033^{X}	0.152*	0.221	0	0	0^{X}	0.093*	0	0	0.007	0.036^{X}	0.130*	0	0	0.014	0.080^{X}	0.154*	0	0
		30	0	0.002	0	0	0	0	0	0	0	0	0.003	0.001	0	0	0	0.004	0.003	0	0	0

Table 4.6.7: Proportion of times each available dose-pair declared as the recommended dose-pair, out of those trials which identified a recommended dose-pair, under dose-escalation Method 3 with a range of prior settings for Scenarios 1 and 3. Dark grey cells highlight dose-pairs with toxicity category overdose, light grey cells the target interval and white cells underdoses. The 'X' and '*' mark the 'true' recommended dose-pair for each scenario based solely on toxicity and exposure data respectively.

Scenario	Priorn setting	Average number of patients per trial	Proportion of trials identified a recommended dose-pair	Average proportion recommended dose-pairs with true classification					
				by toxicity			by exposure of drugs A and B		
				under-dose	target interval	over-dose	undesirably low	desirable	undesirably high
1	1a	55.155	0.915	0.321	0.679	0.000	0.181	0.819	0.000
1	1b	55.440	0.920	0.328	0.671	0.001	0.183	0.816	0.001
1	2a	55.896	0.928	0.255	0.740	0.004	0.159	0.835	0.005
1	2b	56.409	0.937	0.144	0.853	0.003	0.101	0.894	0.004
3	1a	54.699	0.907	0.155	0.379	0.465	0.073	0.596	0.331
3	1b	55.098	0.914	0.321	0.586	0.093	0.146	0.853	0.001
3	2a	54.933	0.911	0.247	0.619	0.134	0.132	0.862	0.007
3	2b	55.287	0.917	0.108	0.731	0.161	0.079	0.914	0.008

Table 4.6.8: Average number of patients observed per trial and proportion of dose-pairs identified as the recommended dose-pair in each of the defined toxicity and exposure categories under dose-escalation Method 3 with a range of prior settings for Scenarios 1 and 3.

From Figure 4.6.2, we can see that the average proportion of DLTs and undesirable exposures occuring in the trial is consistent, if not lower, than those observed in the main simulation study with the original prior. From Table 4.6.7, we see that the spread of recommended dose-pairs is different to that observed in the main simulation study but recommendations are still condensed around values with target toxicity and desirable exposure classifications. In Table 4.6.8, we see that in general there is a slight increase in the proportion of recommended dose-pairs classified as underdoses (by toxicity and exposure classifications) and a decrease in overdoses. Prior setting 1a under Scenario 3 is the only scenario where this is noticably not the case.

In Scenario 3, the true recommended dose-pair based on the exposure classification is only just an overdose by the toxicity classification (with true probability of causing a DLT in a patient equal to 0.37). In the main simulation results, the safety criterion was not suitable to completely avoid escalation to this border-line classification dose-pair, leading to 20% of recommended dose-pairs being classidied as overdoses by the toxicity criterion under Method 3. The priors investigated in the sensitivity analysis cause different patterns of escalation to become more likely. Observation of DLTs or high exposures at low dose-pairs will be more difficult to overcome under prior 1a (which reflects belief that there is only one safe dose) than under the original prior. This is reflected in an increase in recommended dose-pairs in the top left corner of the available dose grid. On the other hand, if few DLTs or high exposures are observed early on in the trial, prior setting 1a is easily overcome and escalation occurs rapidly, hence the increase in recommended dose-pairs classified as overdoses by the toxicity classification for scenarios in this setting.

Prior setting 1b was also investigated under Method 4 (because prior setting 2 is unlikely to arise in practice and prior setting 1a is easier to overcome with data than prior setting 1b). From Figure 4.6.3 and Tables 4.6.9 and 4.6.10, we see that the effect of the stopping rules on the operating characteristics of the dose-escalation trial is minimal. From Figure 4.6.4 we see that even under the more cautious prior setting 1b compared to the original prior setting used in the main chapter, the toxicity and exposure stopping rules are still effective, bringing the average trial size down to 44 and 40 patients in Scenarios 1 and 3, respectively.

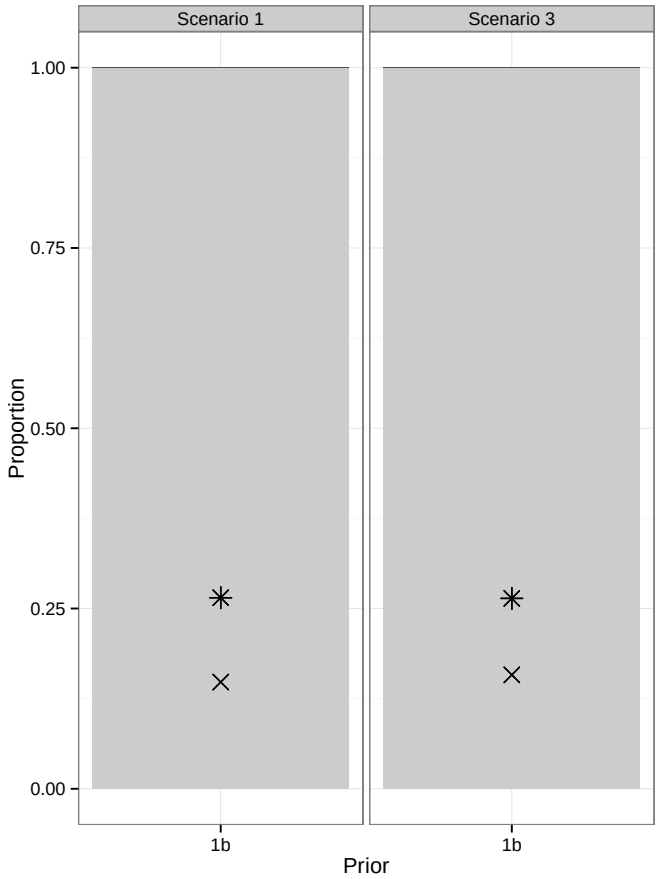

Figure 4.6.3: Average proportion of patients experiencing DLTs (marked by a cross) and undesirably high exposures (marked by a star) per trial under dose-escalation Method 4 with prior setting 1b for Scenarios 1 and 3.

			Prior setting 1b				
			Dose of drug B				
			20	40	60	80	100
Scenario 1	Dose of drug A	10	0.001	0.010	0.026	0.004	0
		15	0	0.002	0.083	0.185	0
		20	0	0	0.052	0.387	0
		25	0	0	0	0.246*	0.002
		30	0	0	0^{X}	0	0
Scenario 3	Dose of drug A	10	0.011	0.004	0.049	0	0
		15	0.001	0.078	0.216	0	0
		20	0	0.014	0.485	0.002	0
		25	0	0.001^{X}	0.138*	0	0
		30	0	0	0	0	0

Table 4.6.9: Proportion of times each available dose-pair declared as the recommended dose-pair, out of those trials which identified a recommended dose-pair, under dose-escalation Method 4 with prior setting 1b for Scenarios 1 and 3. Dark grey cells highlight dose-pairs with toxicity category overdose, light grey cells the target interval and white cells underdoses. The 'X' and '*' mark the 'true' recommended dose-pair for each scenario based solely on toxicity and exposure data respectively.

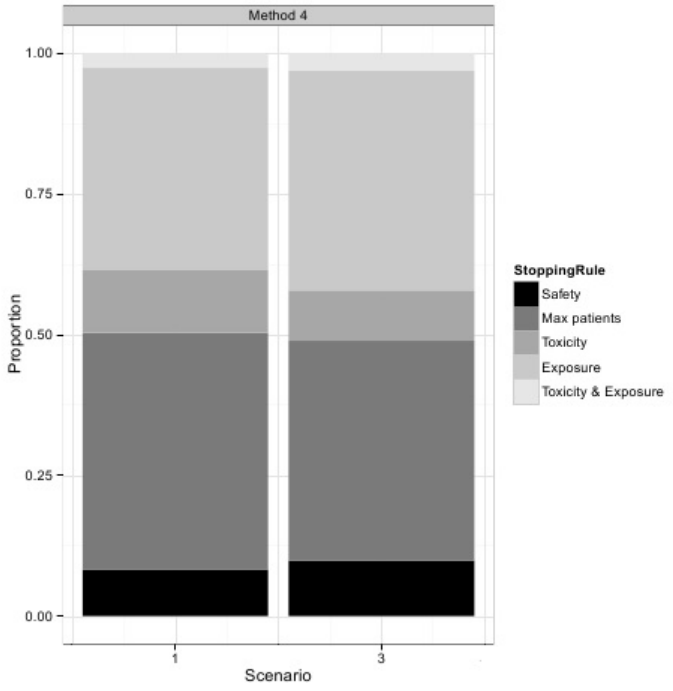

Figure 4.6.4: Reasons trial stopped under dose-escalation Method 4 with prior setting 1b for Scenarios 1 and 3.

Scenario	Priorn setting	Average number of patients per trial	Proportion of trials identified a recommended dose-pair	Average proportion recommended dose-pairs with true classification					
				by toxicity			by exposure of drugs A and B		
				under-dose	target interval	over-dose	undesirably low	desirable	undesirably high
1	1b	43.233	0.917	0.364	0.634	0.002	0.174	0.823	0.002
3	1b	39.819	0.901	0.360	0.501	0.140	0.158	0.840	0.002

Table 4.6.10: Average number of patients observed per trial and proportion of dose-pairs identified as the recommended dose-pair in each of the defined toxicity and exposure categories under dose-escalation Method 4 with prior setting 1b for Scenarios 1 and 3.

Chapter 5

Sample Size Calculation in Phase II Clinical Trials

5.1 Phase II Clinical Trials

Attention now switches from phase I dose-escalation trials to phase II clinical trials. Phase II is a broad term encompassing a range of trial types. In general, phase II clinical trials are hypothesis driven and non-confirmatory. They aim to confirm the safety of the recommended dose of the treatment that was identified in phase I, and to look for initial signs of the treatment's efficacy. If the safety and efficacy profile of the treatment appears to be promising after phase II trials, then the experimental treatment is taken forward to large-scale phase III trials which focus on the treatment's efficacy, while collecting information on its long-term toxicity (Pocock, 2004).

If the results of the phase III trial are positive, then the treatment can be made accessible to patients outside of clinical trials. For this reason, it is important that the

chance of an incorrect conclusion being drawn from the phase III trial is minimised. In light of this, phase III trials are usually large-scale randomised trials comparing the experimental treatment with a control treatment. The primary endpoint considered in the phase III trial is usually the outcome which is felt to provide the most relevant measurement of the treatment's efficacy for the patient population of interest. This is often a time-to-event endpoint, often death in oncology, although this can be slow to observe. The definitive nature of phase III trials mean that strict error controls are placed on the hypotheses concerning the primary endpoint of interest. It also means that frequentist methodology is generally used in order to avoid the subjective nature of Bayesian designs.

As in phase III, the primary objective of phase II clinical trials is generally efficacy related. However, phase II trials are considered to be non-confirmatory trials which are carried out in order to justify progression of the experimental treatment to phase III. For this reason, phase II trials tend to be much smaller and shorter in duration than phase III trials. Despite the restrictions on the design of phase II trials, the inferences drawn from the trial must be relatively accurate in order to minimise the chance of a treatment with an undesirable benefit-risk ratio being taken to phase III. The most common reason for failure in phase III trials is lack a efficacy of the experimental treatment (Arrowsmith and Millar, 2013). So, as well as being costly in terms of resources, failure at phase III can mean that large amounts of (potentially severely ill) patients have been administered an inefficacious treatment. An alternative outcome, which is also not favourable, is that in which patients in phase III trials are administered an efficacious but unacceptably/unnecessarily toxic treatment.

Collecting as much relevant information as possible in early phase trials could improve the reliability concerning the decision over whether an experimental treatment should progress to phase III clinical trials. Design restrictions on phase II trials are common; largely these (implicitly or explicitly) concern the cost implications of the trial. The more patients involved in a trial, the longer the duration (for a fixed recruitment rate), more patients put at risk and higher the cost of the trial. In an attempt to reduce the size of phase II trials there are certain design factors which are commonly used. Some of these are (Seymour et al., 2010):

Short-term endpoint: The actual endpoint of interest for a disease can sometimes take months, or even years, to observe. The trial duration is heavily dependent on this observation time (as well as recruitment rate into the trial). As an alternative, a short-term (often binary) endpoint which can be collected much sooner, and is thought to be highly correlated with the actual endpoint of interest, can be used. For example, in oncology trials, a binary indicator of tumour growth or a continuous time-to-progression endpoint can often be used as an alternative to the actual endpoint of interest, death (FDA et al., 2007).

- In some diseases there is no short-term endpoint which reliably predicts the occurance of the actual endpoint of interest. The use of a poorly predictive short-term endpoint in phase II would lead to an unreliable decision over whether to progress the treatment to phase III. In other cases, the actual endpoint of interest can be observed in a relatively short period, removing the need for an endpoint to be used.

Single-arm trial: In a single-arm trial, data are only collected on the experimental treatment. The trial data on the experimental treatment is then compared to historical control data. The number of patients required in the trial is therefore less than that for a randomised trial which tests hypotheses with the same error constraints. When the historical control data are suitably similar to that which could be obtained in the trial if it were randomised, then reliable conclusions can be drawn from the single-arm trial.

- Since the data used in analysing the results of a single-arm trial are not concurrent, issues concerning comparability can arise. The gold-standard is a randomised trial in which data are obtained concurrently on the experimental and a control treatment (Ratain and Sargent, 2009). The comparisons drawn from randomised trials are more reliable than those from a single-arm trial (with same error constraints on the hypotheses) but the number of patients required in the trial is increased.

Relaxed error constraints: The sample size of a trial is chosen such that the probability of making an incorrect decision concerning the trial hypotheses is controlled to be less than the specified error constraints. The sample size for the trial can be decreased if these error constraints are relaxed.

- Relaxing the error constraints on the hypotheses decreases the certainty with which correct trial conclusions are made; increasing the chance of wrongly abandoning a promising treatment or progressing with an undesirable treatment.

Early stopping allowed: If, at an interim analysis, the experimental treatment is looking very promising (or not) compared to the control treatment, then the trial can be stopped early, reducing the size and resource burden of the trial. Early stopping for efficacy and/or futility such as this is often used in phase III trials and is becoming more common in phase II.

- The use of interim analyses, to decide whether there is suitable evidence (or lack) of efficacy to stop the trial early, increases the complexity of trial designs (Whitehead, 1997). When a frequentist design is used, multiplicity issues arise from having multiple analysis points. Trial planning (in terms of funding) can also be more difficult when using an interim analysis because a single sample size cannot be obtained for the trials - instead they are calculated in terms of maximum or expected sample sizes.

Bayesian methodology: When designing phase II trials there is little physical data available concerning the efficacy of the experimental treatment. There is however, a range of less formal data (for example, knowledge of the treatment in a different application or efficacy observations in patients involved in phase I trials) which can be incorporated into the trial design and/or analysis using Bayesian methodology. When the available data are incorporated in a sensible manner (with thorough consideration of the effect of historical data on the outcome of the trial), and a confirmatory trial will follow, then Bayesian designs are often justified and can reduce the trial size.

- Bayesian methods account for uncertainty in both the outcome of the trial

and the model parameters and so the Bayesian sample size calculated for a trial will not always be lower than a frequentist alternative. The subjective nature of Bayesian designs, and strict regulatory control, mean that Bayesian methods are rarely used in confirmatory, phase III trials. They are however endorsed for use in small clinical trials within the pharmaceutical industry (CHMP et al., 2006).

Ideally, none of the above design factors would be used in phase II trials. However, within cost restrictions on phase II trials, one or more of these design factors are likely to be employed. In Section 5.2, a frequentist method of sample size calculation which is commonly used in single-arm, phase II trials with a binary endpoint is given. This example clarifies the standard set-up of phase II clinical trials and the use of error constraints in sample size calculation. In Section 5.3, the discussion is extended to time-to-event data. Methods of modelling the time-to-event data are described, as well as a frequentist method of sample size calculation. A literature review of Bayesian and frequentist alternatives for sample size calculation based on a time-to-event endpoint is given in Section 6.1.

In Section 6.2, Bayesian methods of sample size calculation for phase II clinical trials with a time-to-event endpoint are considered. Sample size calculations are presented for both single-arm and randomised trials and in each case the error constraints on the hypotheses can be specified as desired. Calculations are given for a single analysis at the end of the trial but could be extended to account for interim analyses. The methods are illustrated in Section 6.4 using uveal melanoma data but the method is

also applicable outside of oncology when a time-to-event endpoints is of interest.

The calculations involved in the work in Chapter 6 rely on historical data on the control treatment, together with a proportional hazards assumption (which is discussed in detail in Section 5.3), to find the number of events which need to be observed in order to test the trial hypothesis with given error constraints. Recruiting only the number of patients equal to the number of events required and waiting for them all to experience an event can be a lengthy process. An alternative, which can reduce trial duration, is to recruit more patients than the number of events required. For a given trial duration, the expected sample size can then be calculated. This additional calculation requires some information on the time to the event of interest for patients. In addition, for the case of a randomised design, selection of a suitable allocation ratio of patients between experimental and control treatments is considered.

The methodology used for the proposed sample size calculations in Chapter 6 is Bayesian; there are several advantages of using Bayesian methods in early phase clinical trials. At the design stage of the phase II trial, there will be some information available on the experimental treatment. For example, data could be available from the phase I trial of the treatment, use of the treatment in another application and informal use of the treatment. More complete and relevant data is likely to be available on the control treatment. The use of Bayesian methods for the phase II trial enables this information to be incorporated along with the observed trial data to obtain updated inferences on the experimental treatment. As well as being intuitive that available data should be utilised, this can reduce the sample size required for the current trial. Hence, reducing the cost and duration of the current trial.

5.2 Sample Size Calculation Based on a Binary Endpoint

At the design stage of a phase II clinical trial it is important to know what resources are likely to be needed for the planned trial and, based on this, whether the trial is indeed feasible. A sample size calculation can aid this decision. Analysis of the trials we are concerned with involve testing pre-defined hypotheses concerning efficacy response rates on the experimental (and control) treatment(s). Now, the more events observed in the trial, the more likely that the correct decision will be made concerning the trial hypotheses (in terms of type II error since type I error will be fixed in the design). In the sample size calculation, an estimate of the number of events required in order to control the probability of making an incorrect decision at a fixed error level, is calculated.

Initially, consider a single-arm trial based on a binary endpoint. A frequentist method of sample size calculation for such a trial is described in the following two paragraphs as an introduction to sample size calculation. A more detailed account of this approach can be found in Stallard (2008).

Take the endpoint of interest to be a positive binary response (observation that a patient's tumour has shrunk by some fixed amount, for example) at A years. Now, historical data can be used to define p_0, the expected response rate within A years on the control treatment. We consider the experimental treatment to be sufficiently promising to progress to phase III trials if the observed response rate at A years is at least p_1, for $p_1 > p_0$. Here, p_1 can be considered as the clinically worthwhile response

rate and it should be selected by assessing the needs of the treatment area and the potential benefits of the experimental treatment.

Now consider the trial design: In the trial, n patients will be administered with the experimental treatment. Say that m of these patients respond positively to treatment within A years. The experimental treatment is considered promising if $p_1 > p_0$, equivalently if $m \geq \kappa$ for some κ. The values of n and κ are chosen to control the risk of wrongly progressing the treatment to phase III (the one-sided type I error α, typically 0.05 or 0.10) and the risk of wrongly abandoning the treatment (the type II error β, typically 0.2 or 0.1). Letting p be the true probability that a patient administered the experimental treatment will respond within A years, a search procedure can be used to identify pairs (n, κ), which satisfy:

$$\mathbb{P}(m \geq \kappa | p = p_0) \leq \alpha \quad \text{and} \quad \mathbb{P}(m \geq \kappa | p = p_1) \geq 1 - \beta.$$

The pair with the smallest n value is that which minimises the sample size and is therefore the pair of interest. Note that, when analysing the trial data the value of κ corresponding to the actual sample size used in the trial should be used, and not that of the planned sample size.

5.3 Utilising Time-to-event Data in Sample Size Calculation for Phase II Clinical Trials

The sample size calculation presented in Section 5.2 is often suitable for traditional phase II cancer trials. In these trials, a binary response (such as tumour shrinkage) is often used as a short-term alternative for the more relevant endpoints of time to disease progression or mortality. When considering a cytotoxic treatment for solid tumours, where the aim is to reduce tumour size, an intermediate marker such as tumour shrinkage may be suitable. However, many new cancer treatments are intended to be cytostatic rather than cytotoxic; that is they will control the growth of the tumour rather than destroying it (Millar and Lynch, 2003). In such cases, destruction or shrinkage of the tumour is not anticipated and "tumour response" is no longer a sensible endpoint. In the case of Ipilimumab, an immunotherapy approved by the FDA in March 2011 for the treatment of uveal melanoma, no reliable alternative endpoint for time to mortality could be identified. This led to the endpoint in uveal melanoma trials of this treatment being changed from response to overall survival (Hodi et al., 2010; Robert et al., 2011).

Another example is that of diseases such as pancreatic cancer, for which the use of a short-term, binary endpoint appears to be unnecessary. In this disease, median survival is of the order of six months (Amikura et al., 1995; Kayahara et al., 1993). As a result, there is no substantial advantage in terms of trial duration in seeking earlier endpoints such as tumour response or progression-free survival; sadly the most objective endpoint, time to death, is likely to be quickly available. This can also

be the case in other therapeutic areas for rapidly lethal conditions such as alcoholic hepatitis (Ramond et al., 1992). Similarly in infectious diseases where time to fever clearance or viral clearance is often taken as the endpoint of interest (Fox et al., 2011).

Arrowsmith and Millar (2013) agree that there is a need in oncology to design phase II trials which "can deliver data that are sufficient to support good decision-making, and to have suitably discriminatory proof-of-concept criteria agreed prospectively". That is, using an endpoint in phase II which is directly relevant to the efficacy endpoint of interest even though this may require longer, larger trials than has become usual. The use of randomised, as opposed to single-arm, trials could also lead to more informed decision making from phase II trials.

It is widely agreed that randomised trials are preferable to single-arm designs (Ratain and Sargent, 2009). However, within the resources available for a phase II clinical trial, a randomised trial might not be feasible. Such a design could be made more feasible by relaxing the error constraints but this compromises certainty in the trial conclusion. Alternatively, a Bayesian design can be used. Bayesian designs enable incorporation of prior data on the experimental and/or control treatment. In this way, the required sample sizes can be reduced (Whitehead et al., 2008).

In frequentist sample size calculations, a 1:1 allocation ratio between the experimental and control treatments minimises the sample size. This is also be the case in a Bayesian design where equal amounts of prior data are available on both treatments. In reality, there will be more prior knowledge surrounding the control than experimental treatment; this imbalance in information can be incorporated into the Bayesian design. The result is that a non-equal allocation ratio may minimise the sample size

of the trial. In the case where a large amount of relevant data are available on the control treatment, this could lead to a decision not to allocate any patients to the control treatment (Whitehead et al., 2008).

An overview of methods of modelling time-to-event data is given in Section 5.3.1 in the context of a single-arm trial. The discussion is extended to the case which arises from a randomised trial in Section 5.3.2. The information in these two sections comes from Collett (2014), unless otherwise stated. This book contains additional details on the topics discussed here, as well as their extensions in survival analysis. More information on survival analysis can be found in Cox and Oakes (1996).

5.3.1 Modelling Time-to-event Data from a Single-arm Clinical Trial

The time between a patient's recruitment into the trial (also taken as the time they were administered with treatment) and their (treatment-related) death is considered as the survival time of the patient. As discussed, endpoints other than death, such as time to disease progression, are commonly used as efficacy endpoints which are available much sooner. The data arising from such endpoints is time-to-event data, as opposed to being true survival data. Since mortality can be considered as the event of interest, survival can also be considered as a time-to-event endpoint. The methods of modelling and analysing time-to-event data are the same as those for survival data. Consider time to the event of interest to be a continuous variable which is greater than 0, i.e. the event of interest has not occured at the time that a patient is recruited

into the trial.

In a clinical trial, patients are usually recruited over a period of months, or even years. The recruitment time of each patient is therefore likely to be different, as is their time of event. These key time-points are often both recorded in study time, the time from commencement of the trial. It is often more useful to consider patient time, the time from recruitment to event for each individual patient. Within the practicalities of a clinical trial, it is unlikely that the trial will continue until all patients have experienced an event. Instead the trial will end at a given time-point, by which some patients will have experienced an event (and their actual time-to-event can be calculated), while others have not. Those who have not experienced an event by the end of the study are considered to have censored time-to-event observations.

Censored time-to-event observations, such as those considered here, are right-censored. That is, the patient's time-to-event is greater than the time they were observed for in the trial. So, right-censoring can occur during the trial if a patient chooses to leave the trial before experiencing an event, is lost to follow-up, or outlives the final analysis point in the trial. Left and interval censored data can also arise but are not considered here, for more information on these see Chapter 1 of Collett (2014). The occurrance of censoring in time-to-event data is one of the main reasons, along with the bad fitting normal assumption to the data, for the special handling of this time-to-event data. In handling censored data, it is assumed that censoring occurs at random, between patients and with time.

Table 5.3.1 presents an example data set, in terms of study and patient time, for a clinical trial in which patients are recruited over 3 years and followed up for a further

2 years. The total duration of the trial is therefore 5 years. In Figure 5.3.1 these data are presented visally to clarify the concepts of patient time and censoring in the context of a clinical trial. We see that patients 1, 2, 5, 7 and 8 experienced an event within the trial and so their time-to-event is recorded (in Table 5.3.1 and by a cross in Figure 5.3.1) as not censored. The time-to-event of the other patients were censored, with patients 3 and 6 either leaving the trial or being lost to follow-up before experiencing the event of interest. Patient 4 on the other hand did not experience an event within the trial, leading to a censored time-to-event being recorded for them too.

Patient number	Recruitment time	Event time	Time-to-event (patient time)	Censoring indicator
1	0	3.5	3.5	1
2	0.6	2.8	2.2	1
3	1.0	2.2	1.2	0
4	1.3	5	3.7	0
5	2.1	2.6	0.5	1
6	2.2	4.5	2.3	0
7	2.5	4.8	2.3	1
8	3	3.4	0.4	1

Table 5.3.1: Example time-to-event data for 8 patients. Recruitment and event time are given in terms of study time while time-to-event uses patient time. The censoring indicator is equal to 0 if censored and 1 otherwise. These data are represented visually in Figure 5.3.1.

From the time-to-event data observed in a trial, it may be desirable to make inferences such as the probability that a patient experiences an event before a given time-point of interest. Let the time from recruitment of a patient to the time they experience the event of interest be $t > 0$ and let T be the random variable of which t is a realisation. The distribution function $F(t) = \mathbb{P}(T < t)$ gives the probability that

a patient experiences an event before time t. In the analysis of time-to-event data, the quantity $1 - F(t) = \mathbb{P}(T \geq t)$ is often of interest. This is known as the survivor function, $S(t)$.

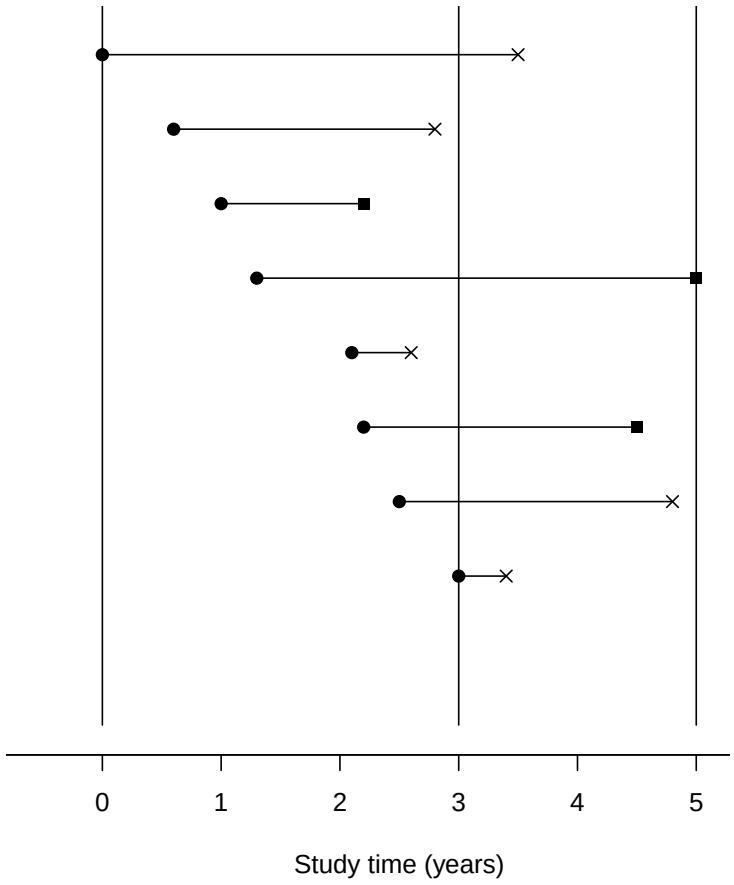

Figure 5.3.1: Visual representation of the time-to-event data given in Table 5.3.1. The patient time is shown with recruitment represented by a circle, an event by a cross and censoring by a square.

Now, say that the probability of experiencing an event at time t is of interest. Technically this probability is equal to 0. So, what we actually calculate is the probability of an event occurring between time t and $t + \delta t$; and consider the limit as δt tends to 0. Logically, this probability is conditioned on the patient not having experienced the event of interest by time t. The resulting probability is the hazard

function $h(t)$:

$$\begin{aligned} h(t) &= \lim_{\delta t \to 0} \left\{ \frac{\mathbb{P}(t \leq T < t + \delta t | T \geq t)}{\delta t} \right\}, \\ &= \frac{f(t)}{S(t)}, \\ \text{for} \quad f(t) &= \frac{d}{dt} F(t). \end{aligned}$$

These probabilities can be estimated from a given data set. When the data closely follow a parametric model, a probability distribution function can be used to accurately estimate $S(t)$, and hence $h(t)$, for a given t. The Weibull distribution has been used in medical statistics (such as in the analysis of pancreatic cancer survival data Ko et al., 2008; Lima et al., 2004) and in engineering applications (where the distribution was initially derived).

In the remainder of this section, time-to-event data are discussed in two settings: when the data can be modelled by a Weibull distribution and when a parametric assumption is not required/suitable. The Weibull distribution is also used in Chapter 6 when a parameteric distribution is considered for the time-to-event data. At the end of this section, methods of checking whether a Weibull assumption is sufficient are described. Although the parametric methods discussed are described and demonstrated under a Weibull assumption, alternative parametric distributions could be used with calculations and derivations following a similar logic.

Weibull distributed data

The Weibull distribution is an extension of the exponential distribution, which it holds as a special case. In the exponential special case, the hazard rate is constant with time such that $h(t) = \lambda$ for $\lambda > 0$. For a clinical trial with a time-to-event endpoint which is assumed to be exponentially distributed, the implication is that a patient has the same probability of experiencing an event at their time of recruitment into the trial as they have at 3 years into the trial, say. This assumption may be suitable to model the occurance of an adverse event, for which the patient's risk is unaffected by the duration of their treatment. However, it is clearly unrealistic for modelling the endpoint of mortality, for example.

In addition to the scale parameter λ, which comprises the hazard function of the exponential distribution, the Weibull distribution has a rate parameter, γ (with $\gamma > 0$). The hazard and survivor functions under a Weibull model can be derived from its probability distribution function as:

$$h(t) = \lambda\gamma t^{\gamma-1} \quad \text{and} \quad S(t) = e^{-\lambda t^{\gamma}}.$$

So, when $\gamma = 1$, the Weibull distribution reduces to the exponential special case in which the hazard rate is assumed to be constant with time. For $\gamma > 1$, the hazard rate increases monotonically with time and for $0 < \gamma < 1$, the hazard rate decreases monotonically with time. This additional flexibility provided by the rate parameter makes the Weibull distribution much better suited than the exponential distribution for modelling time-to-event data. The distributional assumption is discussed here but

this may still be too restrictive in practice. Non-parametric alternatives which may be better suited if the Weibull distribution cannot be assumed are discussed later in this section.

If the observed time-to-event data are assumed to follow a Weibull distribution, then maximum likelihood estimates $\hat{\lambda}$ and $\hat{\gamma}$ can be obtained. The maximum likelihood estimates are the parameter values which maximise the likelihood function. They correspond to the parameters of the best-fitting Weibull curve to the data. Here, the likelihood function for complete and censored Weibull distributed data is derived. A basic understanding of maximum likelihood methods is assumed and more information on these methods can be found in Pawitan (2001). As well as being used to obtain maximum likelihood estimates of the model parameters, which can be used to estimate the survivor and hazard functions at a given time, the likelihood function is used in deriving the posterior distribution of the model parameters. This topic was discussed in Section 2.2.1 on Bayesian methods.

Take a trial in which n patients were treated with the experimental treatment and their survival times observed. If $m = n$ of these patients died during the trial, then no censoring occurred and the likelihood of the survival times, t, based on the distribution parameters is:

$$\begin{aligned} L(t) &= \prod_{j=1}^{n} f(t_j), \\ &= \prod_{j=1}^{n} \lambda\gamma t_j^{\gamma-1} \mathrm{e}^{-\lambda t_j^{\gamma}} \text{ assuming Weibull distributed survival times.} \end{aligned}$$

When censoring is present in the data, we have $n > m$ with the actual survival

times of $n-m$ patients not available. We still have some information about the survival times of these patients; their survival times are greater than their last observation time. For censoring indicator d_j, equal to 0 if observation j is censored and 1 otherwise, this information can be incorporated into the likelihood function as follows:

$$\begin{aligned} L(t) &= \prod_{j=1}^{m} f(t_j) \prod_{j=m+1}^{n} S(t_j), \\ &= \prod_{j=1}^{n} \{f(t_j)\}^{d_j} \{S(t)\}^{1-d_j}, \\ &= \prod_{j=1}^{n} \left(\lambda\gamma t_j^{\gamma-1} \mathrm{e}^{-\lambda t_j^{\gamma}}\right)^{d_j} \left(\mathrm{e}^{-\lambda t_j^{\gamma}}\right)^{1-d_j} \text{ assuming Weibull distributed survival times.} \end{aligned}$$

No parametric assumption

The choice of parametric models may be too restrictive to sufficiently model the observed data in some cases. An alternative approach, which does not require a distributional assumption, is to use non-parametric methods. Some parametric methods for time-to-event data, which are used in the sample size calculation in Section 6.2 are described in this section.

If the time-to-event data set does not contain any censored observations, then the survivor function can be estimated using an intuitive estimate; the proportion of patients alive at time t. This is known as the empirical survivor function;

$$\hat{S}(t) = \frac{\sum_{j=1}^{n} \mathbb{I}_{t_j > t}}{n},$$

where $\mathbb{I}_{t_j>t}$ is an indicator equal to 1 if the survival time of patient i is greater than

t and 0 otherwise. Assuming that all patients are alive at the start of the trial, $\hat{S}(t)$ will be equal to 1 at commencement of the trial. The estimate will decrease at event times until all patients in the trial have experienced an event, at which time $\hat{S}(t) = 0$. Between the observed event times, the empirical survivor function is constant. The resulting curve is therefore a step function, of a similar form to that in Figure 5.3.2.

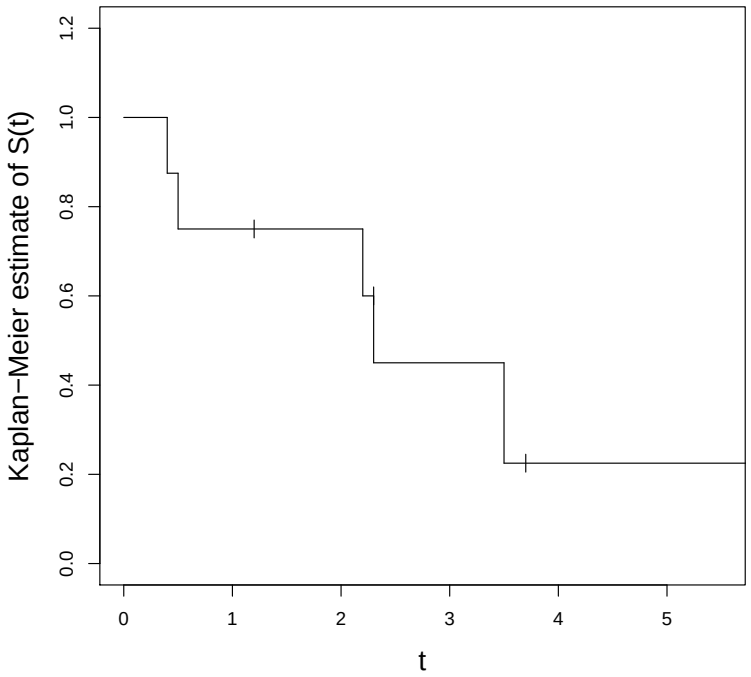

Figure 5.3.2: Plot of the Kaplan-Meier estimate of the survivor function for the data given in Table 5.3.1 with calculation of the estimate given in Table 5.3.2. Censored observations are marked by a vertical dash.

Event time	$n_j - r_j$	D_j	$(n_j - r_j - D_j)/(n_j - r_j)$	$\hat{S}(t)$
0	8	0	1	1.000
0.4	8	1	7/8	0.875
0.5	7	1	6/7	0.750
2.2	5	1	4/5	0.600
2.3	4	1	3/4	0.450
3.5	2	1	1/2	0.225

Table 5.3.2: Calculation of the Kaplan-Meier estimate of the survivor function, which is shown in Figure 5.3.2, for the time-to-event data given in Table 5.3.1.

It is unlikely that all of the trial data will be uncensored and so this empirical estimate must be extended to account for censoring: Define the start time of the trial as t_0 and let times $t_{(1)}, ...t_{(m)}$ be the ascending, non-censored event times. The intervals between consecutive event times may contain censored event times and at each time point there may be more than one event (this is most likely due to recording inaccuracies rather than truly spontaneous events). If $n - r_j$ patients have not experienced an event just before time $t_{(j)}$ and D_j events are observed at time $t_{(j)}$, then the probability of experiencing an event at time $t_{(j)}$ is $(n - r_j - D_j)/(n_j - r_j)$. It can be shown that this probability is in fact an estimate of the probability of experiencing an event between times $t_{(j)}$ and $t_{(j+1)}$ is $(n - r_j - D_j)/(n - r_j)$. A censored observation which occurs at an event time is included in n but not in D_j at that event time. The estimated survivor function is then the Kaplan-Meier estimate:

$$\hat{S}(t) = \prod_{j=1}^{m} \frac{n - r_j - D_j}{n - r_j}.$$

The calculation of the Kaplan-Meier estimate of the survivor function is demonstrated in Table 5.3.2 for the data given in Table 5.3.1. The estimate is plotted as in Figure 5.3.2. The Kaplan-Meier estimate is a step function from $\hat{S}(t) = 1$ at time $t = 0$. In the example data set in Table 5.3.1, the final event time was censored. So, instead of the Kaplan-Meier estimate of the survivor curve decreasing to $\hat{S}(t) = 0$ at the final event time, the function is undefined after the final uncensored event time (at 4.8 years). Observed event times are marked on the plot by decreses in the function, the size of the step determines the number of events which occurred at that time

point. Censoring is also be marked on the curve by a vertical dash at censored event times.

An alternative method of estimating the survivor function uses the Nelson-Aalen estimate. The Nelson-Aalen estimate is used to estimate the cumulative hazard function, $H(t) = \int_0^t h(u)du$, which is in turn used to estimate the survivor function, $\hat{S}(t) = e^{-\hat{H}(t)}$. In small samples, the properties of the Kaplan-Meier and Nelson-Aalen estimates differ, with the Nelson-Aalen estimate often performing better (Colosimo et al., 2002). However, asymptotically the estimates are the same and the Kaplan-Meier estimate derives naturally from the empirical survivor function, making its derivation and use more natural.

Using a Weibull assumption or non-parametric methods

The step-function approximation to the true survior function which results from non-parametric methods does not always produce reliable estimates; this can be especially problematic with sparse data sets. However, the use of parametric methods with an invalid distributional assumption can also lead to unreliable estimates. The amount and distribution of the data should be considered in order to select the most appropriate method of estimating the survivor function. If a parametric model is to be assumed, it is important to check the validity of the distributional assumption.

An idea of the validity of a parametric assumption can be obtained visually by plotting the non-parametric estimate of the survivor function alongside the best fitting curve based on the selected parametric model. Similarity between the two curves indicates a reasonably well fitting distributional model. When a Weibull model is

assumed, the properties of the distribution mean that this judgement call can be made clearer: Since the survivor function of the Weibull distribution is $S(t) = e^{-\lambda t^{\gamma}}$, it can be seen that $\log[-\log\{S(t)\}] = \log(\lambda) + \gamma \log(t)$. Now plot $\log[-\log\{\hat{S}(t)\}]$ against $\log(t)$ for Kaplan-Meier estimates $\hat{S}(t)$ of the survivor function. If the resulting curve roughly follows a straight line with intercept and gradient similar to the maximum likelihood estimates obtained for the logarithm of the scale parameter and the shape parameter of the Weibull distribution respectively, then the Weibull distribution is considered to fit the data reasonably well.

Numerical estimates for the goodness-of-fit of a parametric model to a data set can also be assessed (Collett, 2014). For example, obtaining narrow confidence intervals around the maximum likelihood parameter estimates can be indicative of a well-fitting distributional assumption. Similarly, small Akaike information criterion or Bayesian information criterion values indicate the model is a good fit to the data.

Unfortunatley, these numerical methods are only useful in selecting between a series of candidate models. They do not indicate whether any model is a particularly good fit. So, both numerical and visual methods suffer from vagueness over assessing the validity of a model assumption.

5.3.2 Modelling Time-to-event Data from a Randomised Clinical Trial

It is unlikely that there will be large amounts of relevant data available regarding the time to the event of interest for the experimental treatment. However, such data is

likely to be available for patients on the current standard or control treatment. The historical control data can be used in the trial design and sample size calculation together with any existing information on the experimental treatment.

A basic assumption in the analysis of time-to-event data from two treatments is that of the proportional hazards model (Cox, 1972). Under the proportional hazards model, the ratio of the hazard rate on the experimental treatment to that on the control treatment is considered to be constant with time. That is, for hazard rates $h_{\mathrm{E}}(t)$ and $h_{\mathrm{C}}(t)$ on the experimental and control treatments, respectively, $h_{\mathrm{E}}(t)/h_{\mathrm{C}}(t) = c$ for constant c and $t > 0$. By the definition of the hazard function, c must be greater than 0 and so set $c = e^g$.

If $h_{\mathrm{E}}(t)$ and $h_{\mathrm{C}}(t)$ are proportional, then a baseline hazard rate $h_0(t)$ can be defined to which they are also both proportional. Now, for constant g_i relating the baseline hazard function to the hazard rate on treatment i,

$$\begin{aligned} h_i(t) &= e^{g_i} h_0(t) \quad \text{for} \quad i = \mathrm{E},\, \mathrm{C}. \\ \text{This can be re-written as} \quad h_{\mathrm{E}}(t) &= e^g h_{\mathrm{C}}(t) \text{ for } g = g_{\mathrm{E}} - g_{\mathrm{C}}, \\ \text{equivalently,} \quad S_{\mathrm{E}}(t) &= S_{\mathrm{C}}(t)^{e^g}. \end{aligned}$$

The proportional hazards assumption underlies the log-rank test which is a commonly used method of analysis for time-to-event data from two treatments. The log-rank test is a hypothesis test of whether the time-to-event data from patients treated with the experimental and control treatments originated from the same population. Equivalently, it tests the null hypothesis of no difference in time to the event

of interest between the two treatment groups. The number of events observed in the trial in each group (O_i for $i = \mathrm{E}, \mathrm{C}$) is straight-forward to calculate from the time-to-event data. In addition, find the number of events which we would expect to observe in each group (E_i) if there was in fact no difference between the treatment groups. We have that $E_i = n_{\mathrm{i}} D/n$ where $n = n_{\mathrm{E}} + n_{\mathrm{C}}$ and D is the total number of events observed in all n patients. The log-rank test statistic, which can be compared to a χ^2_1 distribution, is then:

$$X^2 = \frac{(O_{\mathrm{E}} - E_{\mathrm{E}})^2}{O_{\mathrm{E}}} + \frac{(O_{\mathrm{C}} - E_{\mathrm{C}})^2}{O_{\mathrm{C}}}$$

In Section 6.2, historical time-to-event data on the control and experimental treatments are used to derive Bayesian sample sizes for single-arm and randomised trials with time-to-event endpoints. Sample sizes based on the log-rank test are used later in Chapter 6 as a frequentist sample size for comparison with those obtained from the proposed Bayesian calculations.

Chapter 6

Bayesian Methods for Setting Sample Sizes and Choosing Allocation Ratios in Phase II Clinical Trials with Time-to-event Endpoints

Abstract

Conventional phase II trials using binary endpoints as early indicators of a time-to-event outcome are not always feasible. Uveal melanoma has no reliable intermediate marker of efficacy. In pancreatic cancer and viral clearance, the time to the event of interest is short, making an early indicator unnecessary. In the latter application,

Weibull models have been used to analyse corresponding time-to-event data.

Bayesian sample size calculations are presented for single-arm and randomised phase II trials assuming proportional hazards models for time-to-event endpoints. Special consideration is given to the case where survival times follow the Weibull distribution. The proposed methods are demonstrated through an illustrative trial based on uveal melanoma patient data. A procedure for prior specification based on knowledge or predictions of survival patterns is described. This enables investigation into the choice of allocation ratio in the randomised setting to assess whether a control arm is indeed required.

The Bayesian framework enables sample sizes consistent with those used in practice to be obtained. When a confirmatory phase III trial will follow if suitable evidence of efficacy is identified, Bayesian approaches are less controversial than for definitive trials. In the randomised setting, a compromise for obtaining feasible sample sizes is a loss in certainty in the specified hypotheses: the Bayesian counterpart of power. However, this approach may still be preferable to running a single-arm trial where no data is collected on the control treatment. This dilemma is present in most phase II trials, where resources are not sufficient to conduct a definitive trial.

Keywords: Phase II trial; proportional hazards model; sample size calculation; time-to-event endpoint; Bayesian framework.

6.1 Introduction

This chapter concerns phase II trials in which a single novel treatment is to be assessed in terms of time-to-event data. The objective of such a trial is to establish whether the experimental treatment shows sufficient promise to justify large-scale, definitive investigation in phase III. Much has been written about the conduct of phase II trials in oncology, but the development of treatments for infectious diseases and potentially lethal conditions such as alcoholic hepatitis can involve similar investigations.

In conventional phase II clinical trials in oncology, all trial patients are allocated to the experimental treatment (Stallard, 2008). Data are then collected on 'tumour response', defined as a binary indicator of whether a patient's tumour disappears or shrinks by a pre-defined amount within a given follow-up period. Investigators will proceed to further development of the treatment if the number of responses exceeds some critical value, chosen together with the sample size to achieve specified risks of type I and type II error. The subsequent phase III trial will then be a randomised comparison of the experimental treatment with a placebo or standard control, in which time until progression or death is the primary endpoint.

Here we focus on cases where it is either not feasible or not necessary to use a short-term binary endpoint in place of the desired survival endpoint. In advanced pancreatic cancer, for example, typical survival times are 6 months or less, removing the necessity of using tumour response as an earlier indicator of survival. In primary uveal melanoma, there is no observable counterpart to the shrinkage of tumours and investigators have to rely on survival patterns. Similarly, new cytostatic cancer drugs

are designed to limit the growth of tumours rather than to kill them, so that tumour shrinkage is not a necessary condition for efficacy (Millar and Lynch, 2003). Outside the field of oncology, infectious diseases can be assessed in terms of the time to fever clearance or viral clearance: a desirable event occurring within a matter of days. Fox et al. (2011) describe an analysis of the latter. Another example is alcoholic hepatitis that can be rapidly lethal (Ramond et al., 1992) without an intermediate marker through which efficacy is signalled.

Frequentist approaches to the design of phase II trials yielding survival endpoints have been described by various authors. For single-arm studies, proposed designs include a two-stage procedure based on the Nelson-Aalen estimate (Case and Morgan, 2003) and one-stage procedures based on a one-sample log-rank test (Sun et al., 2011). Owzar and Jung (2008) consider various parametric and non-parametric approaches, while Whitehead (2014) constructs a method from survival rates past a limited number of landmark time points. Randomised studies can be based on more familiar frequentist survival approaches, devised to operate with small samples (Evans and Ildstad, 2001, for example).

The Bayesian method of sample size calculation described in this chapter is based on an idea briefly mentioned by Simon (2000) and developed for binary and normally distributed endpoints by Whitehead et al. (2008). We extend this work to the case of time-to-event endpoints when a proportional hazards assumption can be made. For trials where sample sizes are to be set in the absence of detailed knowledge of the likely survival pattern, the Weibull model is suggested for use at the design stage. The method has similarities to that of Thall et al. (2005), which adopts an expo-

nential model to construct a sequential version without explicit calculation of the required sample size. Gittins and Pezeshk (2000) consider sample size determination for survival data through consideration of the cost-benefit of a randomised clinical trial. Various other alternative methods for Bayesian sample size calculations have been described (Dong et al., 2012; O'Hagan and Stevens, 2002; Spiegelhalter et al., 2004; Zaslavsky, 2012; Zaslavsky and Whitehead, 2012; Zhao et al., 2012) but these use different Bayesian principles than those underlying the sample size calculations we present in this chapter.

The Bayesian approach has several advantages over frequentist methods in early phase trials. Incorporation of informative prior opinion about treatment properties allows a reduction in sample size. This strategy has to be used cautiously as it will reduce the amount of real phase II data available for planning later trials. However, particularly in rare diseases, it can enable otherwise infeasible studies to be conducted, and conclusions drawn will be confirmed in subsequent large-scale and probably frequentist phase III trials. Furthermore, in randomised trials the Bayesian approach offers a basis for choosing the ratio of patients allocated to experimental and control treatments. Time-to-event data is such that non-parametric approaches such as proportional hazards regression are difficult to implement without recourse to Markov chain Monte Carlo methods (such as those of Gelfand and Mallick, 1995), and these do not lend themselves for use at the design stage.

In this chapter we adopt a proportional hazards model for the time-to-event data to be collected. In Section 6.2, formulae are developed for the number of events required in a Bayesian trial with time-to-event endpoints in which all patients receive

the experimental therapy and for randomised trials with an R:1 allocation ratio. To transform the number of events required to a target sample size, further assumptions are required concerning the nature of the survival distribution. A detailed example for the special cases, where an exponential or Weibull model is assumed for the purpose of trial design, is presented for illustration in Section 6.3. The choice of R (including the single-arm option, $R = \infty$) is explored in the context of differential prior knowledge of the two treatments. Clinical data show good agreement between exponential models and survival experience following diagnosis of pancreatic cancer (Ko et al., 2008; Lima et al., 2004). In infectious diseases, Weibull models have been used to analyse viral clearance times in dengue fever (Fox et al., 2011). Alternatively, if an estimated survival function based on existing data is available, then a sample size calculation can be based on this in a similar manner to Whitehead (2001). This option is compared to the Weibull case through an illustration, based on data from uveal cancer patients, and is presented in Section 6.4. A discussion of the method and its applications is presented in Section 6.5.

6.2 Bayesian Approach to Sample Size Setting

6.2.1 A Model for the Data and Criteria for Sample Size

Consider testing the null hypothesis of no treatment difference between experimental (E) and control (C) treatments against the alternative of a clinically relevant advantage of E. Survival from time of entry to the trial to occurrence of a certain event is recorded. In this chapter, an event is taken as being undesirable, such as death or dis-

ease progression. Obvious modifications are required if the event is positive, such as viral clearance. Survival times t_{ij} and censoring indicators d_{ij} (= 0 if censored and 1 otherwise) are collected for the j^{th} trial patient receiving treatment i, for $j = 1, ..., n_i$; i = E, C, and they are assumed to be independent. The total number of patients treated in the trial is then $n = n_{\text{E}} + n_{\text{C}}$.

Let the survival function for a patient on treatment i be given by $S_i(t)$ for $t > 0$, i = E, C. In designing the study, a third survival function, $S_0(t)$, will also be considered. This represents the survival experience of patients with whom the investigators are already familiar. There might be historical data from such patients, or there might be a consensus concerning values of $S_0(t)$ for one or more values of t. It will be assumed that $S_i(t) = S_0(t)^{\lambda_i}$ for i = E,C leading to $S_{\text{E}}(t) = S_{\text{C}}(t)^{\text{e}^{-\theta}}$ (which is derived from the more familiar proportional hazards model $h_{\text{E}}(t) = h_{\text{C}}(t)\text{e}^{-\theta}$, where $h_i(t)$ denotes the hazard function for patients on treatment i for $i = \text{E}, \text{C}$) where the negative log-hazard ratio $\theta = -\log(\lambda_{\text{E}}/\lambda_{\text{C}})$ represents the advantage of treatment E over treatment C. Let $\theta_1 > 0$ be the negative log-hazard ratio corresponding to a clinically worthwhile treatment effect. It may be convenient, especially when designing a single-arm trial, to assume that $S_0(t) = S_{\text{C}}(t)$ for all t: that is $\lambda_{\text{C}} = 1$.

Denote all data collected as $\boldsymbol{x} = (\boldsymbol{t}, \boldsymbol{d})$ for $\boldsymbol{t} = \{t_{\text{E}1}, ..., t_{\text{E}n_{\text{E}}}, t_{\text{C}1}, ..., t_{\text{C}n_{\text{C}}}\}$ and $\boldsymbol{d} = \{d_{\text{E}1}, ..., d_{\text{E}n_{\text{E}}}, d_{\text{C}1}, ..., d_{\text{C}n_{\text{C}}}\}$. Then denote data from the n_i patients on treatment i as $\boldsymbol{x}_i = (\boldsymbol{t}_i, \boldsymbol{d}_i)$ for i = E, C. Letting $w_{ij} = -\log\{S_0(t_{ij})\}$ implies that the corresponding random variable W_{ij} follows the exponential distribution with parameter λ_i for i = E, C and all j. The likelihood of λ_i based on w_{ij}, which at this stage cannot be

calculated from the $\boldsymbol{x}$ as $S_0(t)$ is unknown, is then;

$$L(\lambda_i) = \lambda_i^{D_i} \mathrm{e}^{-S_i \lambda_i},$$

where D_i denotes the total number of events observed on treatment i and $S_i = w_{i1} + ... + w_{in_i}$, the sum of the transformed survival times of patients on treatment i.

In Sections 6.2.2 and 6.2.3 in the succeeding paragraphs, the numbers of events required for a single-arm trial ($n_{\mathrm{C}} = 0$) and for a randomised trial will be derived. To deduce the required sample size expected to yield this number of events in a given time-frame, some knowledge of the function $S_0(t)$ is required. If sufficient historic data exist, then a Kaplan-Meier estimate, $\bar{S}_0(t)$ of $S_0(t)$, might be used. Failing that, it might be possible to fix values for $S_0(t_1)$ and $S_0(t_2)$ for two time points t_1 and t_2 from clinical experience. This would allow $S_0(t)$ to be modelled as the Weibull distribution function with rate parameter ϕ_0 and shape parameter γ that takes the specified values at times t_1 and t_2. As we are also assuming that $S_i(t) = S_0(t)^{\lambda_i}$ for $i = \mathrm{E}, \mathrm{C}$, the Weibull assumption implies that survival times on treatment i follow the Weibull distribution function with rate parameter $\phi_0 \lambda_i$ and shape parameter γ, $i = \mathrm{E}, \mathrm{C}$. If the value for $S_0(t)$ can be reliably fixed for only one value of t, then an exponential model can be imposed. Of course, in some settings, the parametric approach might be the method of choice. As discussed at the end of Section 6.1, a Weibull model might be preferred for trials in infectious diseases or an exponential model for trials in pancreatic cancer.

A Bayesian should decide to proceed to further testing in phase III if the posterior belief that the treatment effect is positive is sufficiently strong. Thus, the value of $\mathbb{P}(\Theta > 0|\boldsymbol{x})$, where Θ denotes the random parameter of which θ is a realisation, should be computed. Before the trial is conducted, a critical value η should be specified (usually as a value close to 1) such that the null hypothesis will be rejected and the experimental treatment taken forward to further testing in phase III provided that $\mathbb{P}(\Theta > 0|\boldsymbol{x}) \geq \eta$. More formally, a rejection region $\mathcal{R}$ can be specified such that if $\boldsymbol{x} \in \mathcal{R}$, the null hypothesis will be rejected and the experimental treatment taken forward to further testing in phase III. If $\boldsymbol{x} \notin \mathcal{R}$, the experimental treatment will be abandoned. The posterior probability satisfies,

$$\mathbb{P}(\Theta > 0|\boldsymbol{x}) \quad \geq \quad \eta \quad \text{for all } \boldsymbol{x} \in \mathcal{R}. \tag{6.2.1}$$

Taking forward the experimental treatment will always be associated with a strong belief that it is more effective than the control treatment, which can be seen to be similar to the frequentist criterion of continuing to further trials if the p-value is sufficiently small.

In determining the number of events required in the trial, a second criterion is defined. This is effectively the Bayesian counterpart of a frequentist power calculation. We specify that the sample size and critical region will be chosen so that the posterior probability satisfies Equation 6.2.1 and

$$\mathbb{P}(\Theta < \theta_1|\boldsymbol{x}) \quad \geq \quad \zeta \quad \text{for all } \boldsymbol{x} \notin \mathcal{R}, \tag{6.2.2}$$

where ζ is a value close to 1. In this case, abandoning the experimental treatment will correspond to being convinced that it does not achieve the pre-specified worthwhile treatment effect, θ_1.

In the Bayesian sample size calculations presented in this chapter, a search procedure is used to identify pairs (m, k) that satisfy Equations 6.2.1 and 6.2.2, where m is the number of events that need to be observed in the trial and k is a critical value defining the rejection region. Clearly m patients could be recruited and followed until they all experience an event. However, it may be more practical to recruit $n > m$ patients and follow them all up until m events are observed.

Once the data have been collected in the trial, any form of analysis, frequentist or Bayesian, could be used. It would be consistent with the sample size determination described in this chapter to use a Bayesian analysis based on the prior for λ_{E} adopted during the design stage and on the proportional hazards model and adopted prior for λ_{C} in the case of a randomised trial. The underlying survival function can be estimated using a Weibull or exponential model if such an assumption was made during the design of the trial. Alternatively, as a substantial amount of data should be available at this stage, parametric modelling might not be necessary or appropriate for the analysis.

A simple Bayesian analysis of the trial, conducted to determine whether to proceed to phase III, requires only calculation of the probability given in Equation 6.2.1. A more thorough analysis should also consider the probability given in Equation 6.2.2 and the full posterior distribution of λ_{E} (and, where appropriate, of λ_{C}). When only one of Equation 6.2.1 or 6.2.2 is satisfied, the conclusion of the trial is clear. If both are

satisfied, then the data show with relative certainty that the experimental treatment is better than control but does not reach the worthwhile treatment effect. In this case, the posterior distribution of λ_{E} and the needs of the treatment area should be used to decide whether to proceed to phase III trials.

The frequentist counterpart to the Bayesian method presented here involves calculating the number of events required in order to control the risk of one-sided type I error α and the power $1-\beta$. The required number of events is derived from the amount of information required, V. In the single-arm case, the number of events required is equal to V, while for a 1:1 randomised trial, $4V$ events are required. With z_ϵ denoting the 100ϵ percentage point of the standard normal distribution, the required information is calculated as $V = [(z_{1-\alpha} + z_{1-\beta})/\theta_1]^2$ (Whitehead, 1997).

6.2.2 A Bayesian Single-arm Trial

In a single-arm trial, all patients are allocated the experimental treatment so that $n = n_{\mathrm{E}}$ and $n_{\mathrm{C}} = 0$. Suppose that the experimental treatment should proceed to further trials if the value of the parameter λ_{E} (introduced in Section 6.2.1) is lower than some pre-specified value λ_0. Denote the random variable representing this parameter by Λ_{E}, and the value representing a clinically worthwhile treatment effect by λ_1 ($\lambda_1 < \lambda_0$). Thus, $\theta_1 = -\log(\lambda_1/\lambda_0)$, is the clinically relevant negative log-hazard ratio. A prior gamma distribution will be taken for Λ_E, with parameters a_{E} and b_{E} so that $f_0(\lambda_{\mathrm{E}}) \propto \lambda_{\mathrm{E}}^{a_{\mathrm{E}}-1} e^{-b_{\mathrm{E}}\lambda_{\mathrm{E}}}$. Given the likelihood, $\mathrm{L}(\lambda_{\mathrm{E}}) \propto \lambda_{\mathrm{E}}^{D_{\mathrm{E}}} e^{-S_{\mathrm{E}}\lambda_{\mathrm{E}}}$, it follows that the resulting posterior density is $h(\lambda_{\mathrm{E}}|\boldsymbol{x}_{\mathrm{E}}) \propto \lambda_{\mathrm{E}}^{a_{\mathrm{E}}+D_{\mathrm{E}}-1} e^{-(b_{\mathrm{E}}+S_{\mathrm{E}})\lambda_{\mathrm{E}}}$ so that $\lambda_{\mathrm{E}}|\boldsymbol{x}_{\mathrm{E}} \sim \mathrm{Gamma}(a_{\mathrm{E}} + D_{\mathrm{E}}, b_{\mathrm{E}} + S_{\mathrm{E}})$.

Since $\mathbb{P}(\Lambda_{\mathrm{E}} < \lambda_0)$ is monotone increasing in S_{E} for given a_{E}, b_{E} and D_{E}, the rejection region $\mathcal{R}$ corresponds to the region where $S_{\mathrm{E}} \geq k$ for a suitable value of k. To identify a suitable number of events and corresponding critical value, note that Equations 6.2.1 and 6.2.2 will be true if borderline data for which $S_{\mathrm{E}} = k$ leads to;

$$\mathbb{P}(\Lambda_{\mathrm{E}} < \lambda_0 | S_{\mathrm{E}} = k) \geq \eta \tag{6.2.3}$$

$$\text{and } \mathbb{P}(\Lambda_{\mathrm{E}} > \lambda_1 | S_{\mathrm{E}} = k) \geq \zeta. \tag{6.2.4}$$

A search procedure is used to identify pairs (m_{E}, k) which satisfy Equations 6.2.3 and 6.2.4, where m_{E} is the number of events which need to be observed on the experimental treatment. Notice that to calculate the number of events required in the trial only proportional hazards are required.

From the value of m_{E}, a suitable sample size can be found as follows: assume that entry into the trial is uniform with P patients recruited per year for Y years. After Y years, recruitment ceases, and all patients are followed up for a further A years. The probability $\pi_{\mathrm{E}}(\lambda_{\mathrm{E}}, S_0)$ of an event during the trial for patient j, given λ_{E}, S_0 and entry to the trial at time u can be expressed as follows:

$$\pi_{\mathrm{E}}(\lambda_{\mathrm{E}}, S_0) = \mathbb{P}(d_{Ej} = 1 | \lambda_{\mathrm{E}}, S_0) = 1 - \frac{1}{Y} \int_0^Y S_0(Y + A - u)^{\lambda_{\mathrm{E}}} du.$$

Letting $v = (Y + A - u)$, this becomes

$$\pi_{\mathrm{E}}(\lambda_{\mathrm{E}}, S_0) = 1 - \frac{1}{Y} \int_A^{Y+A} S_0(v)^{\lambda_{\mathrm{E}}} \, dv. \tag{6.2.5}$$

Integrating over all possible values of λ_{E}, weighted by the prior density of λ_{E}, leads to an expression for the prior predictive probability $\bar{\pi}_{\mathrm{E}}(S_0)$ that a patient will experience an event during the trial:

$$\begin{aligned}\bar{\pi}_{\mathrm{E}}(S_0) = \mathbb{P}(d_{Ej} = 1|S_0) &= 1 - \frac{b_{\mathrm{E}}^{a_{\mathrm{E}}}}{Y\Gamma(a_{\mathrm{E}})}\int_A^{Y+A}\int_0^{\infty} \lambda_{\mathrm{E}}^{a_{\mathrm{E}}-1}\mathrm{e}^{-b_{\mathrm{E}}\lambda_{\mathrm{E}}}S_0(v)^{\lambda_{\mathrm{E}}}\, d\lambda_{\mathrm{E}}\, dv, \\ &= 1 - \frac{b_{\mathrm{E}}^{a_{\mathrm{E}}}}{Y}\int_A^{Y+A}[b_{\mathrm{E}} - \log\{S_0(v)\}]^{-a_{\mathrm{E}}}\, dv. \qquad (6.2.6)\end{aligned}$$

Given sufficient historical data to obtain a Kaplan-Meier estimate $\bar{S}_0$ of S_0 with a suitable degree of accuracy. The Kaplan-Meier estimate can be used to approximate the integrals that appear in Equations 6.2.5 and 6.2.6. Suppose that $\bar{S}_0$ takes the form

$$\bar{S}_0(t) = \bar{s}_h \text{ for } t \in (t_{h-1}, t_h),\ h = 1, 2, ..., H,$$

where $t_0 = 0$. The grid of points $\{t_1, t_2, ..., t_H\}$ comprises all of the uncensored event times and also A and $(Y + A)$. Suppose that it is the $(a+1)^{\text{th}}$ such point, t_a, that is equal to A, and the $(b+1)^{\text{th}}$ such point, t_b, that is equal to $Y + A$. (Note that, unless deaths are recorded at the times A or $Y + A$, then $\bar{s}_a = \bar{s}_{a+1}$ and $\bar{s}_b = \bar{s}_{b+1}$.) Equation 6.2.5 can be approximated from

$$\pi_{\mathrm{E}}(\lambda_{\mathrm{E}}, S_0) \approx 1 - \frac{1}{Y}\sum_{h=a+1}^{b}(t_h - t_{h-1})\bar{s}_h^{\lambda_{\mathrm{E}}},$$

and Equation 6.2.6 from

$$\bar{\pi}_{\mathrm{E}}(S_0) \approx 1 - \frac{b_{\mathrm{E}}^{a_{\mathrm{E}}}}{Y}\sum_{h=a+1}^{b}(t_h - t_{h-1})[b_{\mathrm{E}} - \log\{\bar{s}_h\}]^{-a_{\mathrm{E}}}.$$

When a Weibull model for $S_0(t)$ is preferred or when it has to be used in the absence of sufficient historical data, Equations 6.2.5 and 6.2.6 can be simplified. Taking the parameters of the Weibull model to be ϕ_0 and γ, we have $S_0(t) = \exp(-\phi_0 t^\gamma)$, $t > 0$, so that

$$\begin{aligned} \pi_{\mathrm{E}}(\lambda_{\mathrm{E}}, S_0) &= 1 - \frac{1}{Y\gamma\phi_0^{1/\gamma}} \int_{\phi_0 A^\gamma}^{\phi_0 (Y+A)^\gamma} w^{\frac{1-\gamma}{\gamma}} \mathrm{e}^{-w\lambda_{\mathrm{E}}}\, dw \\ \text{and } \bar{\pi}_{\mathrm{E}}(S_0) &= 1 - \frac{b_{\mathrm{E}}^{a_{\mathrm{E}}}}{Y\gamma\phi_0^{1/\gamma}} \int_{\phi_0 A^\gamma}^{\phi_0 (Y+A)^\gamma} w^{\frac{1-\gamma}{\gamma}} (b_{\mathrm{E}} + w)^{-a_{\mathrm{E}}}\, dw \quad \text{for } a_{\mathrm{E}} > 1, b_{\mathrm{E}} > 0. \end{aligned}$$

In the exponential case (when $\gamma = 1$), these equations have closed forms, but in the general Weibull case they could be calculated using Simpson's rule with a suitably fine grid.

The number of events, D_{E} observed in n_{E} patients treated with the experimental treatment is binomially distributed with parameters n_{E}, and $\pi_{\mathrm{E}}(\lambda_{\mathrm{E}}, S_0)$ when λ_{E} is assumed known and $\bar{\pi}(S_0)$ otherwise. We propose the following three methods for deducing the sample size:

Method 1: Sample until m_{E} events have been observed. This method will always satisfy the required Bayesian criteria and requires only proportional hazards (and, in the Weibull case, it does not depend on the value of γ). However, even when this method is used, estimates of patient numbers and trial duration are likely to be of interest for planning. Either of Methods 2 or 3 that follow could be used to do this, and an expression for $S_0(t)$ is required for their operation.

Method 2: Identify a combination (P, Y, A), and corresponding sample size n_{E} such that the expected number of events, $n_{\mathrm{E}}\pi_{\mathrm{E}}(\lambda_1, S_0)$ is equal to m_{E}, using Equation 6.2.5 to evaluate $\pi_{\mathrm{E}}(\lambda_1, S_0)$. This method matches the frequentist approach and is straightforward to use. However, it does not guarantee conditions in Equations 6.2.3 and 6.2.4 with any degree of certainty.

Method 3: Identify a combination (P, Y, A), and corresponding sample size n_{E}, such that Equations 6.2.3 and 6.2.4 are satisfied with high probability; ensuring that $\mathbb{P}(D_{\mathrm{E}} \geq m_{\mathrm{E}}) = 1 - \mathrm{F}_{D_{\mathrm{E}}}(m_{\mathrm{E}} - 1) \geq \xi$ for large ξ, where $\mathrm{F}_{D_{\mathrm{E}}}$ is the cumulative distribution function of D_{E} that is binomially distributed with parameters n_{E} and $\bar{\pi}_{\mathrm{E}}(S_0)$. The sample size calculated under this method accounts for uncertainty in λ_{E}.

Both Methods 2 and 3 require knowledge of the underlying survival function to enable calculation of the integrals in Equations 6.2.5 and 6.2.6. In the Weibull case for both of these methods, decreasing the value of γ leads to an increased sample size. This fact can be used to calculate a conservative sample size, perhaps based on a percentile of the prior distribution of γ in this case.

The value of m_{E} found in the search procedure described at the start of Section 6.2.2 is dependent upon the amount of prior information. The choice of the prior parameters a_{E} and b_{E} can be made by expressing b_{E} as a_{E}/λ^*, so that the prior mean of Λ_E is λ^*, and its standard deviation is $\lambda^*/\sqrt{a_{\mathrm{E}}}$. A weighted average of λ_0 and λ_1, such as $(\lambda_0 + \lambda_1)/2$, might be chosen for λ^*, where $\lambda_1 = \lambda_0 e^{-\theta_1}$. Then a_{E} is chosen to determine the strength of the prior, informed after consideration of properties such

as the prior 95% credibility interval for the median survival time on the experimental treatment. Prior determination will be discussed in more detail in the context of an example in Section 6.3.

6.2.3 A Bayesian Randomised Trial

Now consider a trial in which patients are randomised either to treatment E or treatment C. We have $n_{\mathrm{E}}, n_{\mathrm{C}} \geq 0$ and $\lambda_{\mathrm{E}}, \lambda_{\mathrm{C}}$ unknown. Independent prior gamma distributions with parameters $(a_{\mathrm{E}}, b_{\mathrm{E}})$ and $(a_{\mathrm{C}}, b_{\mathrm{C}})$ will be assumed for the random rate parameters Λ_{E} and Λ_{C}. The corresponding posterior distributions will be independent and gamma with parameters $(a_{\mathrm{E}} + D_{\mathrm{E}}, b_{\mathrm{E}} + S_{\mathrm{E}})$ and $(a_{\mathrm{C}} + D_{\mathrm{C}}, b_{\mathrm{C}} + S_{\mathrm{C}})$ respectively.

Let $\Theta = -\log(\Lambda_E/\Lambda_C)$. Upon collecting data, $(b_i + S_i)\Lambda_i \sim$ Gamma$(a_i + D_i, 1)$. It follows that

$$\frac{(b_{\mathrm{C}} + S_{\mathrm{C}})\Lambda_{\mathrm{C}}}{(b_{\mathrm{C}} + S_{\mathrm{C}})\Lambda_{\mathrm{C}} + (b_{\mathrm{E}} + S_{\mathrm{E}})\Lambda_{\mathrm{E}}} \sim \text{ Beta}(a_{\mathrm{C}} + D_{\mathrm{C}}, a_{\mathrm{E}} + D_{\mathrm{E}}),$$

which can be re-written as $\Theta = \log\{(b_{\mathrm{E}} + S_{\mathrm{E}})Z\}/\{(b_{\mathrm{C}} + S_{\mathrm{C}})(1 - Z)\}$, where $Z \sim$ Beta$(a_{\mathrm{C}} + D_{\mathrm{C}}, a_{\mathrm{E}} + D_{\mathrm{E}})$ (see, for example, Chapter V, Example 25 in Mood et al., 1974). Putting $T = (b_{\mathrm{E}} + S_{\mathrm{E}})/(b_{\mathrm{C}} + S_{\mathrm{C}})$, it follows that

$$\mathbb{P}\left(\Theta > 0 | \boldsymbol{x}\right) = \mathbb{P}\left\{\log\left(T\frac{Z}{1-Z}\right) > 0 \middle| \boldsymbol{x}\right\} = \mathbb{P}\left(Z > \frac{1}{1+T}\middle|\boldsymbol{x}\right),$$

which is monotone increasing in T. Hence, we will reject the null hypothesis if $T \geq k$ for a suitable critical value k. Equations 6.2.1 and 6.2.2 will be true if borderline

data for which $T = k$ leads to

$$\mathbb{P}(Z > 1/(1+T)|T = k) \geq \eta \tag{6.2.7}$$

$$\text{and } \mathbb{P}(Z < 1/\{1+\mathrm{e}^{-\theta_1}T\}|T = k) \geq \zeta, \tag{6.2.8}$$

where $\mathbb{P}(\Theta < \theta_1) = \mathbb{P}\{Z < (1+\mathrm{e}^{-\theta_1}k)^{-1}\}$. As with the single-arm trial, these equations can be used to identify pairs $(\boldsymbol{m}, k)$ which satisfy trial requirements, where $\boldsymbol{m} = (m_{\mathrm{E}}, m_{\mathrm{C}})$ is the number of events observed in the trial on experimental and control treatments respectively.

Three methods of sample size calculation that are parallel to those presented for the single-arm trial are available for the randomised case. For Method 1, a conservative sample size choice that minimises the number of events observed in the trial can be obtained by specifying that $m_{\mathrm{E}} = m_{\mathrm{C}}$ are both equal to some common value m. In that case $Z \sim \text{Beta}(a_C + m, a_E + m)$, and the value of m can be deduced from Equations 6.2.7 and 6.2.8. Under the same constraint, it is relatively straightforward to extend Methods 2 and 3 presented for the single-arm case to the randomised setting. However, because survival times are expected to be greater, and so the rate of events lower, on the experimental than control treatment, setting $m_{\mathrm{E}} = m_{\mathrm{C}}$ will not generally yield the smallest sample size for a given combination (P, Y, A).

It may be preferable to minimise the sample size. Suppose that Y and A are fixed, and that the allocation of patients between the experimental and control treatments is to be in an R:1 ratio. In the case of Method 2, Equation 6.2.5 can be used to determine $\pi_{\mathrm{E}}(\lambda_1, S_0)$ and $\pi_{\mathrm{C}}(\lambda_0, S_0)$ and the expected numbers of events on each treatment can

be deduced for any patient entry rate P. The value of P can be adjusted so that when these values are used for D_{E} and D_{C} respectively in the beta distribution of Z, Equations 6.2.7 and 6.2.8 are valid. Finally, the process can be repeated for a variety of feasible allocation ratios R and a value chosen that gives rise to the smallest total sample size.

For Method 3, Equation 6.2.6 can be used to find $\bar{\pi}_{\mathrm{E}}(S_0)$ and $\bar{\pi}_{\mathrm{C}}(S_0)$ and thus determine the corresponding independent binomial distributions for the numbers of events on each treatment for any value of P. Thus, for any pair $(m_{\mathrm{E}}, m_{\mathrm{C}})$ of numbers of events in the two treatment groups, we have the probability of their joint occurrence and, from Equations 6.2.7 and 6.2.8, an indicator of whether they lead to a distribution for Z in which the Bayesian criteria will be satisfied. The minimum value of P is then sought for which, summing these probabilities over the cases in which the criteria are achieved leads to a total that is greater than ξ. Once more, the process can be repeated for a variety of values of R.

For a randomised trial, priors must be specified on both Λ_{E} and Λ_{C}. It seems logical for the prior on Λ_{E} to be specified as in the single-arm trial because the available information concerning this parameter is unchanged. In a similar way, take $b_{\mathrm{C}} = a_{\mathrm{C}}/\lambda_0$ so that the prior mean of Λ_{C} is λ_0 and its standard deviation is $\lambda_0/\sqrt{a_{\mathrm{C}}}$. The prior mean is therefore equal to that observed in historical, conventionally treated patients, and a_{C} is chosen to determine the strength of the prior.

6.3 Illustrative Sample Size Comparisons Based on a Weibull Assumption

Various Weibull settings (including the special case exponential distribution) are investigated in detail, for the single-arm and randomised trial cases. They are illustrated in this section under the assumption that the anticipated probability of survival past 3 years for conventionally treated patients is 0.530. Suppose that increasing this probability to 0.683 is considered to be of clinical importance. Such an improvement represents a hazard ratio relative to the conventional treatment of 0.6. We take $\lambda_C = \lambda_0 = 1$ so that $\lambda_1 = 0.6$ and $\theta_1 = -\log[0.6] = 0.511$.

In order to be confident in the conclusions drawn from the trial, we choose $\eta = 0.95$ and $\zeta = 0.90$ in Equations 6.2.1 and 6.2.2. Although the interpretations of these parameters differ greatly from their frequentist counterparts ($1-\alpha$ and $1-\beta$ respectively), in practice their numerical values are likely to be chosen to coincide, partly because of lack of experience in choosing the Bayesian values. When Method 3 is used, we choose $\xi = 0.95$ to ensure a high probability of observing enough events within the duration of the trial. In all methods considered, we assume constant recruitment of P patients per year for 4 years and conduct an analysis at 6 years. The total sample size is, thus, $n = 4P$. Comparisons are made in terms of the total sample size required to meet the sample size criterion, and searches are taken to the nearest integer value.

Under the frequentist methodology, $V = \{(1.645 + 1.282)/0.511\}^2 = 32.81$, so at least 33 events need to be observed in the single-arm trial in order to satisfy the power

requirements. In the frequentist case, λ is not treated as a random variable and we use $\lambda = \lambda_1$ to find the probability of an event using Equation 6.2.5 for a given γ. For a single-arm trial, the sample size is then the smallest value of n such that the expected number of events in the trial $n\pi_{\mathrm{E}}(\lambda_1, S_0) \geq 33$. Setting $\lambda = \lambda_1$ in this calculation is an arbitrary choice, which could be varied. A frequentist approach to a randomised trial assuming exponential survival times gives a sample size to obtain a required number of events equivalent to that obtained from calculation based on the logrank test. A detailed account of this approach can be found in Whitehead (1997). As in the single-arm approach, the information required is $V = 32.81$. For a 1:1 allocation ratio, this necessitates observing $m = 132$ events.

6.3.1 A Bayesian Single-arm Trial

We take the prior for Λ_{E} to be gamma with parameters a_{E} and b_{E}, where $b_{\mathrm{E}} = 2a_{\mathrm{E}}/(\lambda_0 + \lambda_1)$. Assume that survival times of conventionally treated patients are Weibull distributed with parameters ϕ_0 and γ. Under the transformation of survival times to the exponential distribution, Φ_{E}, the random variable corresponding to the parameter ϕ_{E}, is equal to $\phi_0\Lambda_{\mathrm{E}}$ with $\Phi_0 \sim \mathrm{Gamma}(a_{\mathrm{E}}, b_{\mathrm{E}}/\phi_0)$. As $S_0(3)$ is assumed to be equal to 0.530, $\phi_0 = -\log(0.530)/3^{\gamma}$. For a clinically relevant treatment effect, $\phi_1 = 0.6\phi_0$. With $\gamma = 1$, we have $\phi_0 = 0.212$, the prior mean of Φ_{E} is $\phi_0(\lambda_0 + \lambda_1)/2 = 0.170$, and its standard deviation is $\phi_0(\lambda_0+\lambda_1)/2\sqrt{a_{\mathrm{E}}} = 0.170/\sqrt{a_{\mathrm{E}}}$. Increasing values of a_{E} therefore correspond to increasing levels of certainty in the prior estimate of Λ_{E}. Table 6.3.1 gives the number of events required for Method 1 and the corresponding sample sizes calculated under Methods 2 and 3 for a range of priors and values of γ

consistent with $S_0(3) = 0.530$. Considerations of the interpretation of the prior on λ in terms of median survival is discussed in Section 6.3.2 in relation to the prior on λ_C.

First consider the exponential case when $\gamma = 1$. Weak prior belief, $a_E = 2$, requires observation of 31 events and gives sample sizes of 80 and 89 under Methods 2 and 3 respectively. These are close to the frequentist alternative that requires 85 patients to observe 33 events. With an informative prior, the sample size calculated by Method 2 is always less than the frequentist counterpart because fewer events are required. This is down to the use of the prior. It is also less than that by Method 3 because in the latter approach, uncertainty in the estimate of λ is accounted for. As the precision of prior information increases, the number of events required and corresponding sample sizes from Methods 2 and 3 decrease.

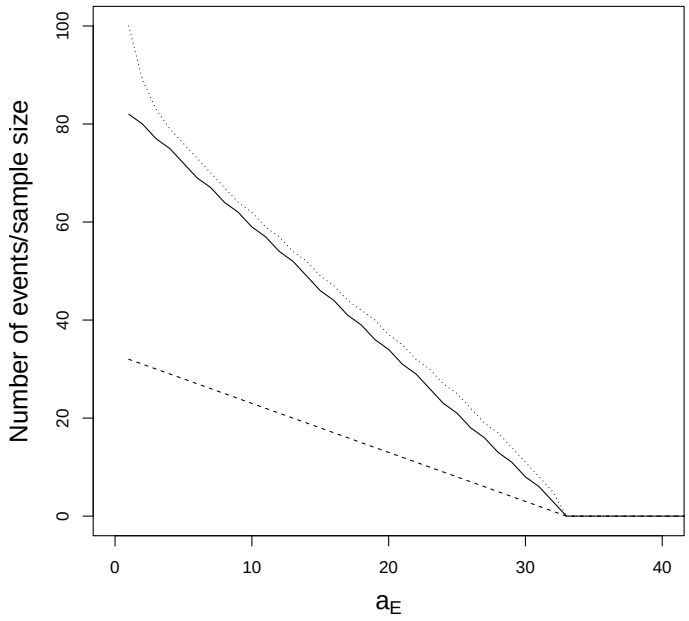

Figure 6.3.1: For varying strength of prior opinion a_E, the number of events m_E (dashed line), required to satisfy trial requirements along with the corresponding single-arm sample size n, calculated using Method 2 (solid line) and Method 3 (dotted line), assuming exponentially distributed survival times.

Figure 6.3.1 plots the number of events required to satisfy the error controls and corresponding sample sizes under Methods 2 and 3, assuming exponentially distributed survival times. It can be seen that a_{E} can be interpreted as being equivalent to a number of imaginary events observed on the experimental treatment prior to the trial. This can also be deduced directly from the gamma posterior with parameters $a_{\mathrm{E}} + D_{\mathrm{E}}$ and $b_{\mathrm{E}} + S_{\mathrm{E}}$. Hence, with prior evidence of $a_{\mathrm{E}} \geq 33$, we obtain $n_{\mathrm{E}} = 0$: no phase II trial is needed. Under this prior setting, for $a_{\mathrm{E}} \geq 33$ both error constraints are satisfied. When this is the case, prior data alone provides convincing evidence that the experimental treatment is better than no treatment and also that it is not better than the control treatment. This leads to a decision about conducting a phase III trial based on the needs of the treatment area, for which available information should be sufficient.

Table 6.3.1 presents numbers of events and sample sizes for a range of values of γ with their corresponding values of ϕ_0 and ϕ_1 for priors constructed as described in the first paragraph of this section. It can be seen that, as the value of γ increases, the required sample size decreases until some point. This is because the rate of events increases with time and so fewer patients are required to observe a fixed number of events. As expected, the greater the value of a_{E}, the more information is contained in the prior and the greater the reduction in the required sample size.

For the situations portrayed in Figure 6.3.1, the difference in sample sizes between Methods 2 and 3 is relatively small (about three patients). Reducing ξ to 0.9 makes the sample sizes from the two methods almost indistinguishable, while a value of $\xi < 0.9$ leads to smaller sample sizes from Method 3 than from Method 2. This is

due to the value of λ being fixed at λ_1 for Method 2 in the example, corresponding to a desirable, low rate of events. In comparison, Method 3 involves integrating over all possible values of λ, including those with high rates of events.

Suppose that previous data led to an estimate and corresponding 95% confidence interval for γ of 1.30 and $(1.14, 1.46)$. Under Method 3, with a moderate prior of $a_{\mathrm{E}} = 5$, an optimistic sample size calculation for a single-arm trial using the estimate of γ requires 71 patients. Alternatively, a conservative calculation using the lower limit of the 95% confidence interval for γ gives a sample size of 74. Both of these are lower than the corresponding frequentist alternatives of 79 and 82, respectively, because they utilise prior information.

Frequentist properties corresponding to the Bayesian sample sizes can be calculated. The frequentist power calculation from observing the number of events required by the Bayesian methods for a given prior to achieve one-sided significance at level 0.05, is given in the penultimate column of Table 6.3.1. As expected, with weak prior belief that $a_{\mathrm{E}} = 2$, the power is high at 0.885. With strengthening prior belief, the frequentist power of the Bayesian design decreases. For comparison, Bayesian properties of the frequentist designs can be found. The final column of Table 6.3.1 shows values of ζ, computed using the Bayesian prior specified but based on the frequentist sample size. With observation of 33 events, as required by the frequentist design, the value of ζ is inflated to 0.954 for moderate prior belief in the novel treatment with $a_{\mathrm{E}} = 10$. A frequentist would be concerned that the Bayesian designs achieved too little power, whereas a Bayesian would feel that the frequentist designs were adding an unnecessary amount of information to the prior opinion already held.

			Frequentist	Bayesian			Frequentist power arising	Bayesian ζ arising
a_E	γ	ϕ_0	n_{E}	m_{E}	n_{E} (Method 2)	n_{E} (Method 3)	from Bayesian m_{E}	from frequentist m_{E}
	0.1	0.57	103	31	96	104		
	0.3	0.46	98	31	93	101		
	0.5	0.37	94	31	89	97		
	0.7	0.29	90	31	85	94		
2	**1.0**	**0.21**	**85**	**31**	**80**	**89**	**0.885**	**0.917**
	1.3	0.15	79	31	75	84		
	1.5	0.12	76	31	72	81		
	1.7	0.10	73	31	69	78		
	1.9	0.08	70	31	66	76		
	0.1	0.57	103	28	87	90		
	0.3	0.46	98	28	84	87		
	0.5	0.37	94	28	80	83		
	0.7	0.29	90	28	77	80		
5	**1.0**	**0.21**	**85**	**28**	**72**	**76**	**0.855**	**0.933**
	1.3	0.15	79	28	67	71		
	1.5	0.12	76	28	65	69		
	1.7	0.10	73	28	62	66		
	1.9	0.08	70	28	60	64		
	0.1	0.57	103	23	72	74		
	0.3	0.46	98	23	69	71		
	0.5	0.37	94	23	66	68		
	0.7	0.29	90	23	63	66		
10	**1.0**	**0.21**	**85**	**23**	**59**	**62**	**0.790**	**0.954**
	1.3	0.15	79	23	55	58		
	1.5	0.12	76	23	53	56		
	1.7	0.10	73	23	51	54		
	1.9	0.08	70	23	49	52		
	0.1	0.57	103	13	41	45		
	0.3	0.46	98	13	39	43		
	0.5	0.37	94	13	37	41		
	0.7	0.29	90	13	36	40		
20	**1.0**	**0.21**	**85**	**13**	**34**	**37**	**0.578**	**0.979**
	1.3	0.15	79	13	32	35		
	1.5	0.12	76	13	30	34		
	1.7	0.10	73	13	29	32		
	1.9	0.08	70	13	28	31		

Table 6.3.1: Sample sizes for single-arm frequentist and Bayesian designs for a range of prior settings and values of γ. The final two columns give some operating characteristics of the design: The frequentist power calculation for the corresponding Bayesian number of events to achieve one-sided significance at level 0.05 and the Bayesian ζ of the frequentist required number of events.

6.3.2 A Bayesian Randomised Trial

As mentioned in Section 6.2.1, we assume that γ is known and equal for the control and experimental treatments. The properties of Bayesian randomised trial designs depend on the priors taken for Λ_{E} and Λ_{C} and on the ratio of allocations to the two treatments. We take these to be gamma with parameters a_{C} and $b_{\mathrm{C}} = a_{\mathrm{C}}/\lambda_0$, and a_{E} and $b_{\mathrm{E}} =$

$2a_{\mathrm{E}}/(\lambda_0 + \lambda_1)$, respectively. Assuming Weibull distributed survival times, Table 6.3.2 gives sample sizes for a range of allocation ratios and γ values under Method 3 with ξ set at 0.95. Method 3 is chosen here because it explicitly quantifies the uncertainty about the sample size calculation (aside from the uncertainty associated with the value of γ). The values of ϕ_0 and ϕ_1 corresponding to each value of γ are the same as for the single-arm setting presented in Table 6.3.1.

Recall that we are assuming that, for conventionally treated patients, the probability of survival past 3 years is 0.530; hence, setting $\gamma = 1$ implies that $\phi_0 = 0.212$. When $a_{\mathrm{C}} = 100$, then $b_{\mathrm{C}} = 472.5$. Thus, Φ_{C} has a 95% prior credibility interval of $(0.172, 0.255)$, corresponding to a 95% prior credibility interval of $(2.7, 4.0)$ for median survival time (in years) on C. Such prior certainty is not unreasonable for a standard therapy that has been widely used. By contrast, taking $a_{\mathrm{C}} = 2$ gives $b_{\mathrm{C}} = 9.5$. This leads to respective 95% prior credibility intervals of $(0.026, 0.590)$ and $(1.2, 27.1)$ for Φ_{C} and for the median survival time on C. This really does virtually represent no knowledge of the control treatment.

The lack of prior knowledge is reflected in prior scenarios 1, 10 and 19 for which $a_{\mathrm{C}} = 2$. For a 1:1 treatment allocation ratio under Method 3, the sample size is greater than the corresponding frequentist sample sizes that are 296, 280 and 264 for $\gamma = 0.7$, 1 and 1.3, respectively. As in the single-arm case, the Bayesian estimate is greater due to the additional uncertainty in the estimate of λ_{E} and λ_{C}. For a randomised study with $a_{\mathrm{C}} = 2$, unequal allocation ratios lead to an increase in the estimated sample size, as they would for frequentist calculations. The priors are essentially providing little useful information. The priors in scenarios 6, 15 and 24 (for which $a_{\mathrm{C}} = 5$),

contain more information than in scenarios 1, 10 and 19, but again, assume equal amounts of prior information on the two treatments. This is reflected in the smaller sample sizes, just less than the frequentist counterparts.

Prior scenario	γ	Prior parameters		Estimated total sample size, n, for R:1 allocation ratio on E:C				
		a_{E}	a_{C}	$R=1$	$R=2$	$R=3$	$R=4$	$R=\infty$
1	0.7	2	2	328	366	436	510	∞
2	0.7	2	5	312	336	392	455	∞
3	0.7	2	20	286	279	304	335	∞
4	0.7	2	50	258	222	216	215	∞
5	0.7	2	100	234	186	172	165	141
6	0.7	5	5	292	324	384	450	∞
7	0.7	5	20	264	264	292	325	∞
8	0.7	5	50	232	201	196	200	∞
9	0.7	5	100	206	165	152	145	125
10	1.0	2	2	310	348	412	485	∞
11	1.0	2	5	296	318	372	435	∞
12	1.0	2	20	272	264	288	320	∞
13	1.0	2	50	244	210	204	205	∞
14	1.0	2	100	220	177	160	155	134
15	1.0	5	5	278	309	364	425	∞
16	1.0	5	20	250	249	276	305	∞
17	1.0	5	50	220	192	188	185	∞
18	1.0	5	100	194	156	144	135	118
19	1.3	2	2	296	330	396	465	∞
20	1.3	2	5	282	303	356	415	∞
21	1.3	2	20	258	252	272	300	∞
22	1.3	2	50	232	201	192	195	∞
23	1.3	2	100	210	168	152	145	127
24	1.3	5	5	264	294	348	405	∞
25	1.3	5	20	236	237	260	290	∞
26	1.3	5	50	208	180	176	180	∞
27	1.3	5	100	184	147	136	130	112

Table 6.3.2: Sample sizes for randomised Bayesian designs under Method 3 for various prior and allocation ratio settings. The blocks of results correspond to sample sizes for $\gamma = 0.7$, 1 and 1.3 respectively. Results for $R = \infty$ are chosen as those observed with an allocation ratio of $R_{\mathrm{E}} = 30$.

Looking at the exponential case when $\gamma = 1$, when a_{C} is larger than a_{E} (as in prior scenarios 11-14 and 16-18), indicating that more is known about the control than about the experimental treatment, there can be advantage in taking more observations on the experimental than the control treatment. For prior scenario 12, a 2:1 allocation ratio is worthwhile, while for prior scenarios 14 and 18, it is better not to take any observations on the control. Even when no observations are taken on control, this analysis differs from the single-arm case of Section 6.3.1 as uncertainty about the effect of control is being allowed for. In scenarios 10-13 and 15-17 when no patients are allocated to control, no recruitment rate is satisfactory: the prior information about the control is insufficient to allow a suitable design to be found. Similar patterns are seen when $\gamma \neq 1$. It can also be seen, as in the single-arm case, that decreasing the value of γ leads to an increase in sample size.

For a randomised trial with prior mean estimate $\hat{\gamma} = 1.30$, it can be seen that the sample sizes are not really feasible for a standard phase II trial. An option is to relax the Bayesian criteria in Equations 6.2.3 and 6.2.4 and/or the value of ξ. Table 6.3.3 presents the sample sizes for $\eta = 0.85$, $\zeta = 0.75$ and $\xi = 0.90$. The resulting sample sizes are much more consistent with practice in phase II, where a confirmatory phase III trial will follow if suitable evidence of efficacy is apparent in phase II.

Prior scenario	γ	Prior parameters		Estimated sample size, n, for R:1 allocation ratio on E:C				
		a_{E}	a_{C}	$R=1$	$R=2$	$R=3$	$R=4$	$R=\infty$
1	1.3	2	2	184	201	240	280	∞
2	1.3	2	5	174	183	208	240	∞
3	1.3	2	20	154	141	144	150	∞
4	1.3	2	50	136	111	104	100	89
5	1.3	2	100	124	96	88	85	69
6	1.3	5	5	160	174	204	235	∞
7	1.3	5	20	136	129	132	140	∞
8	1.3	5	50	116	96	88	85	77
9	1.3	5	100	104	78	72	70	56

Table 6.3.3: Sample sizes for randomised Bayesian designs under Method 3 for $\gamma = 1.3$ with relaxed Bayesian criteria such that $\eta = 0.85$, $\zeta = 0.75$ and $\xi = 0.90$.

6.4 Evaluation of Therapy for Uveal Melanoma

In this section, the methods described in this chapter are illustrated for the design of a potential trial in high-risk, early ocular melanoma - a rare and serious condition for which there are few treatment options. Throughout this section, we take $\lambda_{\mathrm{C}} = \lambda_0 = 1$ so that for a hazard ratio of the novel treatment relative to control treatment of 0.6, we have $\lambda_1 = 0.6$ and $\theta_1 = -\log(0.6) = 0.511$. Choosing $\eta = 0.95$ and $\zeta = 0.90$ in Equations 6.2.1 and 6.2.2, this information can be used to calculate that 28 events are required to be observed in a single-arm trial. Assume a constant recruitment of P patients per year for 4 years and conduct an analysis at 6 years. Method 3 is the preferred method of sample size calculation and will be used for calculations with ξ equal to 0.95. Take a vague prior on the experimental treatment with $a_{\mathrm{E}} = 5$ and $b_{\mathrm{E}} = 2a_{\mathrm{E}}/(\lambda_0 + \lambda_1)$.

If the only reliable information on the survival of conventionally treated patients available at the time of trial design was that the survival probability past 1.5 years is equal to 0.85, then sample size calculation can be based upon an exponential assumption for survival times. Suppose that increasing this probability to 0.907 is considered to be of clinical importance. Such an improvement represents a hazard ratio relative to the conventional treatment of 0.6. We calculate from the exponential survival function that $\phi_0 = -\log(0.85)/1.5 = 0.11$, giving an estimate of $\bar{\pi}_{\mathrm{E}}(S_0) = 0.281$. Hence, a total of 128 patients would be required to observe the required 28 events under the given trial specification by Method 3.

Next, suppose that clinicians doubt the suitability of an exponential model but suggest that the probability of survival past 6 years for conventionally treated patients is 0.3. Assuming survival times of conventially treated patients follow a Weibull(ϕ_0, γ) distribution, we have $S_0(t_p) = \exp(-\phi_0 t_p^{\gamma})$ for time-points $t_p = 1.5, 6$. Solving simultaneously we have $\hat{\gamma} = 1.44$ and $\hat{\phi}_0 = 0.09$ using

$$\gamma = \frac{\log[\log\{S_0(t_{1.5})\}/\log\{S_0(t_6)\}]}{\log(t_{1.5}/t_6)} \qquad \text{and} \qquad \phi_0 = -\frac{\log\{S_0(t_{1.5})\}}{t_{1.5}^{\gamma}}.$$

These values can be used to calculate a sample size of 90 for a single-arm trial using Method 3. This is derived from an estimated probability of event of $\bar{\pi}_{\mathrm{E}}(S_0) = 0.393$. Note that, as discussed in the previous section, it may be desirable to relax the error constraints - a relatively small change can have an appreciable effect on the sample size calculation. For example, decreasing ξ to 0.9 leads to a sample size of 85. Alternatively, η or ζ could be relaxed depending on the requirements of the trial.

Finally, suppose that sufficient survival data on conventionally treated patients is available to remove the need for a distributional assumption. This is the case for the uveal melanoma data presented in Figure 6.4.1 that shows the survival pattern for 264 high-risk patients with uveal melanoma and is based on records from the database of the Liverpool Ocular Oncology Service. Patients were selected if resident in mainland Britain and diagnosed clinically or histologically with uveal melanoma with a tumour involving the choroid and a metasizing uveal melanoma exceeding 15mm in diameter. They were excluded if they had bilateral uveal melanoma, the genetic tumour type was not identified or the basal tumour diameter was not recorded. These patients have also been used to illustrate a different approach in Whitehead et al. (2012).

Instead of using a distributional assumption in this case, the Kaplan-Meier curve can be used to estimate $\bar{\pi}_{\mathrm{E}}(S_0) = 0.447$ following the method described in Section 6.2.2. This leads to a sample size of 78 patients for a single-arm trial according to Method 3. In this example, the sample size calculated using a Weibull approximation is seen to be conservative. This comes from the fact that, in the area of interest (between $A = 2$ and $Y + A = 6$ years), the Kaplan-Meier curve lies below the best-fitting Weibull curve, as can be seen in Figure 6.4.1.

These calculations can be extended to the randomised setting. Suppose that there is moderate prior information corresponding to $a_{\mathrm{C}} = 20$ available on the control treatment. We take $b_{\mathrm{C}} = a_{\mathrm{E}}/\lambda_0$, a 2:1 allocation ratio and the remaining parameters as described for the single-arm setting. The total sample sizes calculated under exponential and Weibull assumptions are 414 and 294, respectively, and based on the Kaplan-Meier estimate of the survival curve, the total sample size is 258. Re-calculating these

sample sizes based on relaxed error constraints with $\eta = 0.85$, $\zeta = 0.75$ and $\xi = 0.9$ leads to sample sizes of 96, 69 and 60 respectively. This demonstrates the large effect that altering the error constraints can have and why their choice should be carefully considered. As in the single-agent case, the Weibull estimate is conservative.

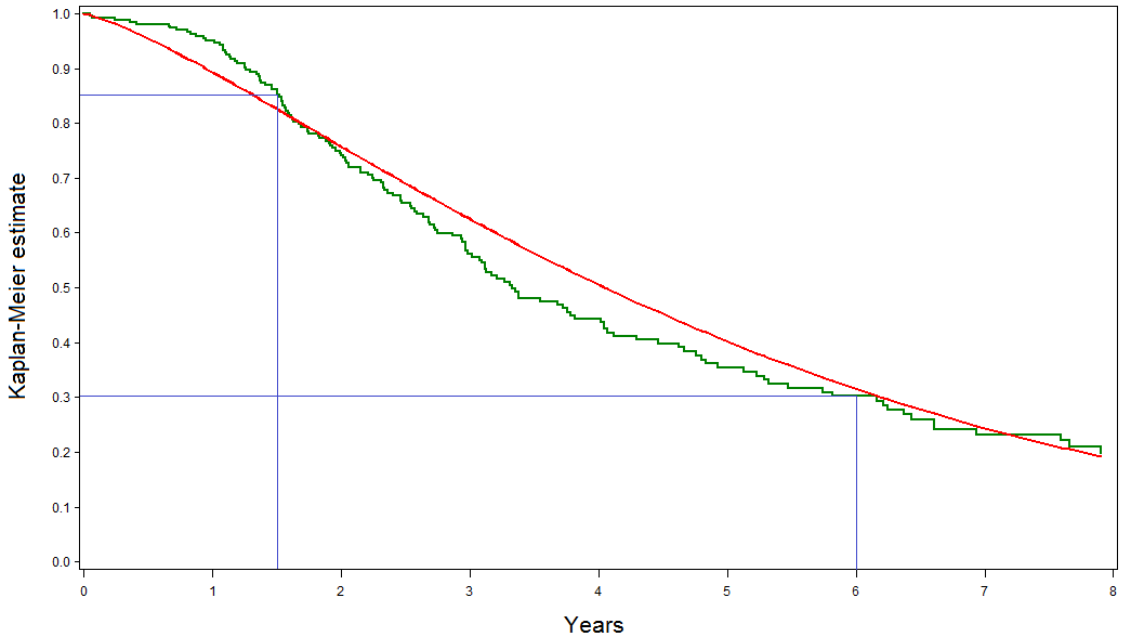

Figure 6.4.1: For varying strength of prior opinion a_{E}, the number of events m_{E} (dashed line), required to satisfy trial requirements along with the corresponding single-arm sample size n, calculated using Method 2 (solid line) and Method 3 (dotted line), assuming exponentially distributed survival times.

6.5 Discussion and Conclusions

In this chapter, we have presented novel Bayesian methods of sample size estimation based on a proportional hazards model for time-to-event outcomes. The method allows the number of events required in a trial to be calculated without knowledge of the survival function. With no knowledge of the survival function, under only the assumption of proportional hazards, Method 1 could be used and the trial run until the required m events had been observed. To predict whether this number of events will be

observed within a specified time requires knowledge concerning the survival function; be this a distributional assumption or suitable information to reliably estimate the survival function. The fact that, under a Weibull assumption, lower values of γ lead to increased sample sizes can be used to obtain conservative sample size calculations. In addition, when the survival curve lies above the fitted Weibull model, sample size estimates using the Weibull assumption will be greater than those based on the survival function itself. An alternative and more reliable approach involves conducting an interim analysis leading to sample-size re-assessment.

For the final analysis of the trial data, the quantities in Equations 6.2.1 and 6.2.2 can be estimated using the posterior estimate of γ or by a Bayesian analysis which allows for a prior on γ. By Methods 2 and 3, achievement of one of the criteria is then very likely but no longer guaranteed. The critical value corresponding to the actual number of events observed in the trial and observed γ should be used in analysis. Within the Bayesian framework, Equations 6.2.1 and 6.2.2 can be applied at interim analyses without penalty - enabling early stopping as soon as the experimental treatment shows sufficient promise, or lack of it.

The Bayesian methods of sample size calculation presented involve calculation of the number of events required in the trial. This is recorded as Method 1 and results in fewer events being required in the trial than its frequentist counterpart with type I error rate and power equated to $1 - \eta$ and ζ, respectively. Similarly, the nature of Method 2 leads to a sample size smaller than that from a frequentist calculation. Method 3 does not always produce sample sizes smaller than the frequentist approach because it is a more thorough calculation as uncertainty in λ_{E} (and λ_{C} in

the randomised setting) is accounted for rather than anticipated values being treated as known. Once a suitable number of events have been observed in the trial, the trial can be stopped regardless of whether the calculated sample size has been reached. By accounting for uncertainty in the parameter estimates, the Bayesian designs should provide more accurate sample size calculations than a frequentist counterpart for valid assumptions concerning the survival distribution.

In the methodology and examples presented in this chapter for the single-arm case, λ_0 was taken to equal λ_C, which in turn was set at the value 1. This need not be the case. Actually λ_0 is a value that we wish to show that the experimental treatment improves upon (is lower than) by some margin. It may therefore be the case that, instead of being derived from historical control data, λ_0 is a standard or agreed value. Frequentist sample sizes, and those from Method 2, could be made more conservative by choosing less arbitrary values than λ_1 and λ_0 in place of λ in calculating Equation 6.2.5 for the experimental and, when relevant, control treatments. Similarly one could increase the value of ξ in Method 3. The calculation presented for Methods 2 and 3 can be extended to cater for non-constant recruitment rates. For example, the recruitment period could be split into three time blocks; $[0, Y_1)$, $[Y_1, Y_2)$ and $[Y_2, Y]$ with corresponding recruitment rates P_1, P_2 and P_3. The expected number of events by analysis time A can be found for each time block and the sum of these is used to identify the required sample size (Whitehead, 2001).

A Bayesian analysis must incorporate the prior according to Bayes theorem: any other approach departs from the Bayesian principle. If the results of the trial are contrary to prior optimism, which was used to create a well-formulated prior favour-

ing the novel treatment, then there might be some concern in the interpretation. Nevertheless, if the prior expressed confidence in the new treatment (but not enough to warrant direct progress to phase III), then unfavourable trial results will lead to weakened posterior confidence in the new treatment. Investigators will therefore be less ready to proceed to phase III than they were before the phase II trial begun.

In the examples of Sections 6.3 and 6.4, the prior mean on the experimental treatment was selected as a value that was a compromise between the survival rate on control treatment and the hypothesised clinically relevant rate. The effect on the sample size of altering the standard deviation was then investigated by altering a_{E}. Alternative values of the prior mean and indeed prior formulation, in terms of the relationship between the prior parameters, are possible. In a real implementation, one should determine expert clinicians' views on the performance of the novel treatment to elicit a more appropriate prior. This could be achieved through a formal elicitation meeting (Hampson et al., 2014) in which details of the sample size calculation could then be re-worked according to a variety of priors. However, this process cannot be illustrated in an abstract example such as that presented in this chapter.

By formally expressing what is known about survival patterns on the trial treatments as priors, it is possible to address directly questions of whether more patients should be allocated to the experimental treatment, whether there should be a control arm and whether a trial should be conducted at all. The prior information substitutes for observed data, and so it has to be based on reliable considerations. Provided the study is a true phase II investigation, that is one that, if successful, will be confirmed by a conventional, comparative, frequentist phase III study with strict control

of type I and type II error; then it would appear to be reasonable to make use of prior knowledge.

Ratain and Sargent (2009), among others, argue in favour of randomising between control and experimental treatments. However, it seems natural that there will be situations where, given the limited number of patients available for phase II and the experience already available about use of the control treatment, the optimal strategy is to allocate all patients to the experimental treatment. If a small control group is recruited, and their outcomes prove to be at odds with previous experience, it is unlikely that they would be taken at face value anyway. The Bayesian approach also allows a middle way between a single-arm and a randomised study. That is, a study in which uncertainty about responses to the control treatment is expressed as a Bayesian prior, but no new control patients are studied in the trial. While the experimental prior is updated from new data, the control prior is not: nevertheless, allowance is made for uncertainty in the predictions for the control. For this reason, in the Weibull setting Bayesian randomised sample sizes with $R = \infty$ are greater than those from a single-arm trial for the same value of γ.

The sample sizes calculated for a Bayesian single-arm trial assuming Weibull distributed survival times given in Table 6.3.1 are reasonable for a phase II trial. Those for a randomised phase II trial given in Table 6.3.2 may not be. However, an alternative that involves relaxing the Bayesian criteria gave more reasonable sample sizes (Table 6.3.3). The compromise for obtaining these feasible, randomised sample sizes is a loss in the eventual certainty concerning the specified hypotheses. Depending upon the situation, this may still be preferable to a single-arm trial where the treat-

ment comparison is based purely on historical control data which may not be strictly comparable to the data collected in the trial. This dilemma is present in most phase II trials where resources are not sufficient to conduct a definitive trial.

Acknowledgments

The authors thank Professor Bertil Damato of the Royal Liverpool Hospital and Dr Ernie Marshall of the Clatterbridge Centre of Oncology for motivating conversations about the design of clinical trials in uveal melanoma.

We would also like to thank the anonymous reviewer for helpful comments which prompted us to look into the case of non-parametric sample size calculation.

Chapter 7

Summary and Further Work

7.1 Summary

Three manuscripts concerning novel methods for early phase clinical trials are contained in this thesis. These span two important stages in early phase clinical trial design. First, dose-escalation trials were considered; in the single-agent setting in Chapter 3 and in the dual-agent setting in Chapter 4. The focus changed in Chapter 6 to non-confirmatory phase II trials with time-to-event endpoints. Bayesian methods were used in all approaches and each approach was illustrated using real trial data.

The dose-escalation trial designs presented in Chapters 3 and 4 were based on standard model-based dose-escalation procedures for oncology trials. The set-up and running of the proposed designs is therefore similar to that of methods currently used in practice. The proposed designs are practical in that they utilise data which is feasibly available within the constraints of a dose-escalation trial. The stopping rules employed in each case show that the designs are also practical in terms of sample size.

In addition, the flexibility of standard model-based dose-escalation trial designs (for example, their ability to use varied cohort sizes, account for delayed availablility of pharmacokinetic data and for a clinical team to over-ride model decisions if required) is maintained.

Instead of the Whitehead and Williamson (1998) method being used as the underlying dose-escalation trial design in Chapter 3, that of Neuenschwander et al. (2008) (demonstrated with regard to the inclusion of pharmacokinetic exposure data in dose-escalation in Chapter 4) could be used, with escalation and stopping rules updated accordingly. The reverse case is also possible. The proposed dose-escalation trial designs are transferable to therapeutic areas aside from oncology but adaptions may be required to meet the specific needs of patients involved in the trial. The escalation and stopping rules can also be altered based on knowledge of the planned trial in order to make them better suited.

A limitation of all model-based dose-escalation trial designs is the assumption on the form of the underlying dose-toxicity relationship. Alternative monotonic dose-toxicity models (the power model used by O'Quigley et al., 1990, in the CRM or the copula regression model used by Thall et al., 2003 for dual-agent dose-escalation, for example) could be substituted in the dose-escalation methods presented in Chapters 3 and 4. Assuming a non-monotonic dose-toxicity relationship may require additional alterations to the design but in such a case, methods which account for efficacy as well as toxicity are likely to be more suitable than those considered in this thesis.

In Chapter 3, pre-existing knowledge concerning a biomarker is utilised. The selected biomarker is indicative of patient membership of one of two subgroups within

the population, between which the reaction to treatment is expected to differ. Dose-escalation methods which accounted for a potential subgroup effect were considered with the aim of recommending a different dose in each subgroup, for use in future trials, when this was necessary. Through a simulation study, the use of a hypothesis test for identifying whether a subgroup effect had been observed was compared to a design using spike and slab priors. The hypothesis test was found to be low powered in the sample sizes feasible for a dose-escalation trial but the method using spike and slab priors performed better. The potential benefits of correctly identifying different recommended doses in each subgroup greatly outweighs the risks of a missed, or masked, treatment effect; which is more likely when the trial population is treated as being homogeneous.

The small sample sizes considered in dose-escalation trials mean that conclusive evidence on the presence of a subgroup effect cannot be obtained. That is why it is required for the biomarker of interest to be identified before the trial commences. Unfortunately, this means that a subgroup effect can only potentially be identified if the biomarker considered in the trial identifies the relevant subgroups. In addition, only the case of two potential subgroups was considered but there may be more than two relevant, distinct subgroups in a population. This could arise in a paediatric trial for example, with subgroup membership defined by age.

Dual-agent dose-escalation was the subject of Chapter 4. In this setting, single-agent trials of the experimental treatments will have been completed. This single-agent data was used to develop decision rules for escalation, and stopping, which were based on both toxicity and pharmacokinetic exposure data. Simulations showed that

the use of pharmacokinetic data leads to more informed escalation decisions, increasing the safety of the trial, and more importantly, the consistency of the recommended dose-pair. Formal incorporation of pharmacokinetic data removes the inconsistent and subjective use of this data in current practice. Stopping rules can be based upon toxicity and/or pharmacokinetic data, enabling early stopping of dose-escalation having identified the recommended dose-pair with suitable accuracy.

To observe the greatest benefits from utilising exposure data in dose-escalation, the data sould be available in a short time span. Although modern technology makes this possible, higher priority needs to be placed on this data than is currently standard to achieve this in practice. This design could also encounter problems in uptake by medical teams stemming from a reluctance to formally base escalation decisions on pharmacokinetic data. Methodological limitations of the dual-agent dose-escalation design presented surround the assumed dose-response models. Since pre-clinical data on drug-drug interactions does not transfer reliably to the clinical setting, specification of a suitable model for a combination treatment is especially difficult.

The setting of Chapter 6 is that of a phase II clinical trial with a time-to-event endpoint. Bayesian sample size calculations for a single-arm and a randomised trial were presented. The calculation of the number of events required in order to achieve specified Bayesian error constraints was calculated based soley on an assumption of proportional hazards. To obtain the more useful estimate of the number of patients required to achieve this number of events in a given time-frame, further information on the expected time to the event of interest of patients was required. Depending on the amount of relevant historical time-to-event data available, this additional information

could take the form of a parametric assumption. Alternatively, the historical time-to-event data could be used directly. The flexibility of this method makes it applicable in a range of situations. In the randomised setting, the method also enables identification of an optimum allocation ratio which minimises the total sample size of the trial. Consequently, this method can be used to make a decision over whether a control group is indeed required in the trial. That is, when the optimum allocation ratio of patients on experimental and control treatments is ∞:1 suggesting that suitable information is already available on control treatment.

Bayesian methods of sample size calculation require a reduced number of events to be observed than their frequentist counterparts to achieve comparable error controls over trial conclusions. When a confirmatory frequentist trial will follow the Bayesian phase II trial, this is often acceptable. Reducing the size of the trial by using historical data can enable trials to commence which, due to funding restrictions, otherwise could not. Similar arguments arise over the use of single-arm, as opposed to randomised, trials. The subjectivity of Bayesian methods, and the lack of a concurrent control in single-arm trials, when using a time-to-event endpoint can also be considered a reasonable compromise over a potential alternative; using a short-term binary endpoint which is not a good predictor of the long-term time-to-event endpoint of interest.

The proposed sample size calculations are applicable in a range of therapeutic areas when a relevant time-to-event endpoint, which is likely to be observed in a relatively short time-frame, can be identified. Another situation where such a calculation may be relevant is that when there is no short-term binary endpoint which reliably predicts the long-term response of interest. When this is the case, a phase II trial design

using a time-to-event endpoint could be beneficial in reducing the chance of wrongly progressing to phase III trials. However, if the time to observation of the time-to-event endpoint of interest is too great, then it may be difficult to obtain funding for such a trial, even though it could have resource savings in the long-term. In the setting where there is a short-term binary endpoint, which is known to be a reliable predictor of the long-term endpoint of interest, then our design is unlikely to be applied. In such a situation, it is likely that a phase II trial with a binary endpoint could be carried out significantly quicker (potentially leading to earlier progression of the treatment to confirmatory phase III trials) than one based on a time-to-event endpoint. The conclusive nature of phase III trials means that the proposed Bayesian sample size calculation is unlikely to have application in the phase III setting.

The main limitation of the proposed Bayesian sample size calculation is the assumption of proportional hazards, between the experimental treatment and a control treatment, which is used to calculate the number of events required in the trial. The calculation also assumes that observations are independent, as is any potential censoring which occurs in the trial, and there is no missing data in the trial. In addition, it is assumed that the control data upon which calculations are based is similar to that which could be obtained in the trial. The validity of the trial assumptions should be checked to ensure that the calculation is valid and as accurate as possible. An interim analysis which checks these assumptions and updates calculations based upon the updated assumptions could be beneficial.

In calculating the number of patients needed to observe the required number of events in a given time-frame, further assumptions concerning the survival time of

patients on the control treatment are required. When Kaplan-Meier estimates are used for sample size calculation, then high levels of censoring and/or small samples can lead to increased variance in estimates of the survivor function. In these situations, the assumtion that survival is constant between observed events also potentially fails, decreasing the accuracy of Kaplan-Meier estimates. If a parametric assumption (such as Weibull distributed survival times) is utilised, then a conservative sample size estimate can be obtained based on properties of the calculation. This is not ideal and if the parametric assumption is not justified, then the sample size calculation could be inaccurate. However, in either of these cases, if the proportional hazards assumption holds and the trial continues until the calculated number of events are obtained, then the error constraints on trial conclusions will be maintained.

7.2 Further Work

It would be interesting to develop frequentist (based on maximum likelihood estimates) and curve-free alternatives to the Bayesian model-based dose-escalation trial designs proposed in Chapters 3 and 4. This has been done in the single-agent setting for the CRM, as published in O'Quigley and Shen (1996) and O'Quigley (2002), respectively. Another extension of the dose-escalation designs could be to healthy volunteer studies. In this situation, the same patient may receive multiple doses of the treatment over the course of the trial. The dose-response models therefore need to account for inter- as well as intra-patient variability. Whitehead et al. (2001) demonstrate this for a single-agent trial which incorporates pharmacokinetic data.

The dose-escalation trial design presented in Chapter 3, which accounts for a potential subgroup effect, could be made more flexible by extending the design to allow for more than two subgroups within the patient population. Ordinal responses could also be considered. This could be achieved in a similar manner to that in which Tighioutart et al. (2012) extend the EWOC design to allow for ordinal toxicity grading. The possibility that every patient cannot be definitively classified into a subgroup based on their biomarker status could be considered. This would require consideration of how best to dose the patient with unknown subgroup membership and how to utilise the resulting data for the benefit of future patients.

Development of a phase I/II design which continues the investigation of a potential subgroup effect beyond dose-escalation would be beneficial to obtain more confirmatory evidence of a subgroup effect. In this setting, the chance of detecting variability between subgroups (which materialises in different reactions to the treatment via toxicity and/or efficacy outcomes) would be increased. If the sample size of such a design were great enough, it may become possible to consider multiple biomarkers which divide the patient population in different ways.

Now consider the dual-agent dose-escalation trial design in Chapter 4 which utilised pharmacokinetic data in the trial decision rules. The issue of a reluctance of uptake of a dose-escalation design which formally incorporates pharmacokinetic data could potentially be reduced by using a hierarchical model for dose-exposure-toxicity relationship. In using this model, escalation decisions could be based on toxicity alone, whilst still being heavily influenced by the observed exposures. However, such a design is expected to be prone to model mis-specification and so an in-depth investigation

would be required into the sensitivity of the hierarchical model to mis-specification.

Looking into alternative formulations for the case of no interaction in the dose-response models used in this design would increase its flexiblity, enabling the most suitable model to be used for the particular drug combination of interest. In addition, this design could be extended to allow for a three (or more) agent combination. Again, the main difficulty in this arises in defining models for the dose-toxicity and dose-exposure relationships which have enough flexibity to be able to model the wide range of drug-drug interaction which could arise.

The main down-fall of the sample size calculation presented in Chapter 6 arises if the proportional hazards assumption is not valid. Royston and Parmar (2011) have proposed a method of analysis for a randomised trial with a time-to-event endpoint which does not require a proportional hazards assumption. It may be possible to develop a Bayesian method of sample size calculation based upon such a method. A potentially more straight-forward extension of the calculation is to ordinal data, based on the binary calculation presented by Whitehead et al. (2008).

Bibliography

M. Adamina, G. Tomlinson, and U. Guller. Bayesian statistics in oncology: A guide for the clinical investigator. *Cancer*, 115(23):5371–5381, 2009.

K. Amikura, M. Kobari, and S. Matsuno. The time of occurrence of liver metastasis in carcinoma of the pancreas. *International Journal of Pancreatology*, 17(2):139–146, 1995.

J. Arrowsmith and P. Millar. Phase II and phase III attrition rates 2011-2012. *Nature Reviews: Drug Discovery*, 12:569, 2013.

J. Babb and A. Rogatko. Patient specific dosing in a cancer phase I clinical trial. *Statistics in Medicine*, 20:2079–2090, 2001.

J. Babb, A. Rogatko, and S. Zacks. Cancer phase I clinical trials: Efficient dose escalation with overdose control. *Statistics in Medicine*, 17:1103–1120, 1998.

S. Bailey, B. Neuenschwander, G. Laird, and M. Branson. A Bayesian case study in oncology phase I combination dose-finding using logistic regression with covariates. *Journal of Biopharmaceutical Statistics*, 19:469–484, 2009.

T. Bayes and R. Price. An essay towards solving a problem in the doctrine of chance.

By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1763.

JM Bernardo, MJ Bayarri, JO Berger, AP Dawid, D Heckerman, AFM Smith, and M West. Bayesian variable selection for random intercept modeling of Gaussian and non-Gaussian data. *Bayesian Statistics*, 9:165–185, 2011.

S. Biswas, D. Liu, J. Lee, and D. Berry. Bayesian clinical trial at the University of Texas M.D. *Clinical Trials*, 6(3):205–216, 2009.

W. Brannath, E. Zuber, M. Branson, F. Bretz, P. Gallo, M. Posch, and A. Racine-Poon. Confirmatory adaptive designs with Bayesian decision tools for targeted therapy in oncology. *Statistics in Medicine*, 28(10):1445–1463, 2009.

T. Braun and S. Wang. A hierarchical Bayesian design for phase I trials of novel combinations of cancer therapeutic agents. *Biometrics*, 66:805–812, 2010.

Bristol-Myers Squibb. Phase I study of oral Ixabepilone in subjects with advanced cancer (NLM Identifier: NCT00422097 in: ClinicalTrials.gov). Technical report, Bethesda (MD): National Library of Medicine (US), [cited February 2013] 2007-2011. URL `http://ClinicalTrials.gov/show/NCT00422097`.

S. Carter. Study design principles for the clinical evaluation of new drugs as developed by the chemotherapy programme of the National Cancer Institute. *The Design of Clinical Trials in Cancer Therapy*, 1:242–289, 1973.

C. Carvalho, J. Chang, J. Lucas, J. Nevins, Q. Wang, and M. West. High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association*, 103(484):1438–1456, 2008.

L. Case and T. Morgan. Design of phase II cancer trials evaluating survival probabilities. *BMC Medical Research Methodology*, 3(1):6, 2003.

CDER/CBER. Guidance for industry : E6(R1) Guideline for good clinical practice. Technical report, FDA, 1996.

CDER/CBER. Guidance for industry : E16 biomarkers related to drug or biotechnology product development: Context, structure, and format of qualification submissions. Technical report, FDA, 2011.

C. Chen and R. Beckman. Hypothesis testing in a confirmatory phase III trial with a possible subset effect. *Statistics in Biopharmaceutical Research*, 1(4):431–439, 2009.

J. Cheng, J. Babb, C. Langer, S. Aamdal, F. Robert, L. Engelhardt, O. Fernberg, J. Schiller, G. Forsberg, R. Alpaugh, L. Weiner, and A. Rogatko. Individualized patient dosing in phase I clinical trials: The role of escalation with overdose control in PNU-214936. *Journal of Clinical Oncology*, 22(4):602–609, 2004.

CHMP et al. Guideline on clinical trials in small populations. *European Medicines Agency, London*, 3, 2006.

S-C. Chow and J-P. Liu. *Design and analysis of clinical trials: Concepts and methodologies*, volume 979. John Wiley & Sons, 2014.

P. Clark, M. Slevin, S. Joel, R. Osborne, D. Talbot, P. Johnson, R. Reznek, T. Masud, W. Gregory, and P. Wrigley. A randomized trial of two Etoposide schedules in small-cell lung cancer: The influence of pharmacokinetics on efficacy and toxicity. *Journal of Clinical Oncology*, 12(7):1427–1435, 1994.

D. Collett. *Modelling of survival data in medical research*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 3 edition, 2014.

E. Colosimo, F. Ferreira, M. Oliveira, and C. Sousa. Empirical comparisons between Kaplan-Meier and Nelson-Aalen survival function estimators. *Journal of Statistical Computation and Simulation*, 72(4):299–308, 2002.

A. Cotterill and J. Whitehead. Bayesian methods for setting sample sizes and choosing allocation ratios in phase II clinical trials with time-to-event endpoints. *Statistics in Medicine*, 2015; Early view DOI: 101002/sim6426.

A. Cotterill, D. Lorand, J. Wang, and T. Jaki. A practical design for a dual-agent dose-escalation trial that incorporates pharmacokinetic data. *Statistics in Medicine*, 2015; Early view DOI: 101002/sim6482.

D. Cox. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34:187–220, 1972.

D. Cox and D. Oakes. *Analysis of survival data*. Monographs on statistics and applied probability. Chapman & Hall, 1996.

S. Dahlberg, G. Shapiro, J. Clark, and B. Johnson. Evaluation of statistical designs

in phase I expansion cohorts: The Dana-Farber/Harvard cancer center experience. *Journal of National Cancer Institute*, 106(7):dju163, 2014.

D. Dejardin, E. Lesaffre, P. Hamberg, and J. Verweij. A randomized phase I Bayesian dose escalation design for the combination of anti-cancer drugs. *Pharmaceutical Statistics*, 13:196–207, 2014.

H. Dette, F. Bretz, A. Pepelyshev, and J. Pinheiro. Optimal designs for dose-finding studies. *Journal of the American Statistical Association*, 103(483):1225–1237, 2008.

M. Deza and E. Deza. *Encyclopedia of Distances*. Springer Berlin Heidelberg, 2 edition, 2009.

K-M. Do, Z. Qin, and M. Vannucci. *Advances in statistical bioinformatics: Models and integrative inference for high-throughput data*. Cambridge University Press, 2013.

G. Dong, W. Shih, D. Moore, H. Quan, and S. Marcella. A Bayesian-frequentist two-stage single-arm phase II clinical trial design. *Statistics in Medicine*, 31(19): 2055–2067, 2012.

V. Dragalin, V. Fedorov, and Y. Wu. Adaptive designs for selecting drug combinations based on efficacy-toxicity response. *Journal of Statistical Planning and Inference*, 138(2):352–373, 2008.

P. Ellis, Q. Chu, N. Leighl, S. Laurie, H. Fritsch, B. Gaschler-Markefski, S. Gyorffy, and G. Munzert. A phase I open-label dose-escalation study of intravenous BI 2536

together with Pemetrexed in previously treated patients with non-small-cell lung cancer. *Clinical lung cancer*, 14(1):19–27, 2013.

C. Evans and S. Ildstad. *Small clinical trials: Issues and challenges*. National Academy Press Washington DC, 2001.

FDA et al. Guidance for industry: Clinical trial endpoints for the approval of cancer drugs and biologics. *Washington, DC, US Food and Drug Administration*, pages 1–19, 2007.

A. Fox, L. Hoa, C. Simmons, M. Wolbers, H. Wertheim, P. Khuong, T. Ninh, T. Lien, N. Lien, N. Trung, H. Hien, J. Farrar, P. Horby, W. Taylor, and N. Kinh. Immunological and viral determinants of dengue severity in hospitalized adults in Ha Noi, Viet Nam. *PLOS Neglected Tropical Diseases*, 5(3):e967, 2011.

M. Gasparini and J. Eisele. A curve-free method for phase I clinical trials. *Biometrics*, 56(2):609–615, 2000.

A. Gelfand and B. Mallick. Bayesian analysis of proportional hazards models built from monotone functions. *Biometrics*, 51(3):843–852, 1995.

A. Genkin, D. Lewis, and D. Madigan. Sparse logistic regression for text categorization. *DIMACS Working Group on Monitoring Message Streams, Project Report*, 2005.

E. George and R. McCulloch. Approaches to Bayesian variable selection. *Statistica Sinica*, 7:339–373, 1997.

O. Gerke and H. Siedentop. Optimal phase I dose-escalation trial designs in oncology - A simulation study. *Statistics in Medicine*, 27:5329–5344, 2008.

J. Gittins and H. Pezeshk. How large should a clinical trial be? *The Statistician*, 49: 177–187, 2000.

S. Goodman, M. Zahurak, and S. Piantadosi. Some practical improvements in the continual reassessment method for phase I studies. *Statistics in Medicine*, 14(11): 1149–1161, 1995.

W. Greco, G. Barvo, and J. Parsons. The search for synergy: A critical review from a response curve perspective. *Pharmacological Review*, 47(2):331–385, 1995.

W. Greco, H. Faessel, and L. Levasseur. The Search for cytotoxic synergy between anticancer agents: a case of Dorothy and the ruby slippers? *Journal of the National Cancer Institute*, 88(11):699–700, 1996.

L. Haines, I. Perevozskaya, and W. Rosenberger. Bayesian optimal designs for phase I clinical trials. *Biometrics*, 59(3):591–600, 2003.

L. Hampson, J. Whitehead, D. Eleftheriou, and P. Brogan. Bayesian methods for the design and interpretation of clinical trials in very rare diseases. *Statistics in Medicine*, 23(24):4186–4201, 2014.

J. Harrington, G. Wheeler, M. Sweeting, A. Mander, and D. Jodrell. Adaptive designs for dual-agent phase I dose-escalation studies. *Nature Reviews: Clinical Oncology*, 10:277–288, 2013.

F. Hodi, S. O'Day, D. McDermott, R. Weber, J. Sosman, J. Haanen, R. Gonzalez, C. Robert, D. Schadendorf, J. Hassel, W. Akerley, van den Eertwegh A., J. Lutzky, P. Lorigan, J. Vaubel, G. Linette, D. Hogg, C. Ottensmeier, C. Lebbè, C. Peschel, I. Quirt, J. Clark, J. Wolchok, J. Weber, J. Tian, M. Yellin, G. Nichol, A. Hoos, and Urba W. Improved survival with Ipilimumab in patients with metastatic melanoma. *New England Journal of Medicine*, 363:711–723, 2010.

J. Hoeting, D. Madigan, A. Raftery, and C. Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–401, 1999.

P. Hoff. *A first course in Bayesian statistical methods*. Springer Texts in Statistics. Springer Science & Business Media, 2009.

N. Holford. The target concentration approach to clinical drug development. *Clinical Pharmacokinetics*, 29(5):287–291, 1995.

L. Huo, Y. Yuan, and G. Yin. Bayesian dose finding for combined drugs with discrete and continuous doses. *Bayesian Analysis*, 7(4):1035–1052, 2012.

J. Ibrahim and M. Chen. Power prior distributions for regression models. *Statistical Science*, 15:46–60, 2000.

H. Ishwaran and J. Rao. Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, 33(2):730–773, 2005.

K. Jain. *The handbook of biomarkers*. Springer Science & Business Media, 2010.

T. Jaki, S. Clive, and C. Weir. Principles of dose-finding studies in cancer: a comparison of trial designs. *Cancer Chemotherapy and Pharmacology*, 71:1107–1114, 2013.

S. Jambhekar, P. Breen, and Royal Pharmaceutical Society of Great Britain. *Basic pharmacokinetics*. Pharmaceutical Press, 2009.

M. Jenkins, A. Stone, and C. Jennison. An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical Statistics*, 10(4):347–356, 2011.

A. Källén. *Computational Pharmacokinetics*. Chapman & Hall/CRC Biostatistics Series. CRC Press, 2008.

M. Kayahara, T. Nagakawa, K. Ueno, T. Ohta, T. Takeda, and I. Miyazaki. An evaluation of radical resection for pancreatic cancer based on the mode of recurrence as determined by autopsy and diagnostic imaging. *Cancer*, 72(7):2118–2123, 1993.

A. Ko, E. Dito, B. Schilinger, A. Venook, Z. Xu, E. Bergsland, D. Wong, J. Scott, J. Hwang, and M. Tempero. A phase II study evaluating Bevacizumab in combination with fixed-dose rate Gemcitabine and low-dose Cisplatin for metastatic pancreatic cancer: Is an anti-VEGF strategy still applicable? *Investigational New Drugs*, 26:463–471, 2008.

A. Lièvre, J-B. Bachet, D. Le Corre, V. Boige, B. Landi, J-F. Emile, J-F. Côtè, G. Tomasic, C. Penna, M. Ducreux, P. Rougier, F. Penault-Llorca, and P. Laurent-

Puig. KRAS mutation status is predictive of response to Cetuximab therapy in colorectal cancer. *Cancer Research*, 66(8):3992–3995, 2006.

C. Lima, M. Green, R. Rotche, W. Miller Jr, G. Jeffrey, L. Cisar, A. Morganti, N. Orlando, G. Gruia, and L. Miller. Irinotecan plus Gemcitabine results in no survival advantage compared with Gemcitabine monotherapy in patients with locally advanced or metastatic pancreatic cancer despite increased tumor response rate. *Journal of Clinical Oncology*, 22(18):3776–3783, 2004.

S. Mandrekar, Y. Cui, and D. Sargent. An adaptive phase I design for identifying a biologically optimal dose for dual agent drug combinations. *Statistics in Medicine*, 26:2317–2330, 2007.

E. Manolis, S. Rohou, R. Hemmings, T. Salmonson, M. Karlsson, and P. Milligan. The role of modeling and simulation in development and registration of medicinal products: Output from the EFPIA/EMA modeling and simulation workshop. *Pharmacometrics and Systems Pharmacology*, 2(2):e31, 2013.

Merck Sharp & Dohme Corp. Phase I study of MK-1496 in patients with advanced solid tumor (NLM Identifier: NCT00880568 in: ClinicalTrials.gov). Technical report, Merck Bethesda (MD): National Library of Medicine (US), [cited February 2013] 2009-2012. URL `http://ClinicalTrials.gov/show/NCT00880568`.

J. Mestre-Ferrandiz, J. Sussex, and A. Towse. The R&D cost of a new medicine. Technical report, Office of Health Economics, 2013. URL `https://www.ohe.org/news/ohe-study-pharmaceutical-rd-costs-released`.

A. Millar and K. Lynch. Rethinking clinical trials for cytostatic drugs. *Nature Reviews: Cancer*, 3:540–545, 2003.

A. Mood, F. Graybill, and D. Boes. *Introduction to the Theory of Statistics*. McGraw-Hill International Editions. McGraw-Hill, 3 edition, 1974.

NCI. National Cancer Institute Dictionary of Cancer Terms, [cited August 2014] 2014. URL `http://www.cancer.gov/dictionary`.

B. Neuenschwander, M. Branson, and T. Gsponer. Critical aspects of the Bayesian approach to phase I cancer trials. *Statistics in Medicine*, 27:2420–2439, 2008.

B. Neuenschwander, A. Matano, Z. Tang, S. Roychoudhury, S. Wandel, and S. Bailey. *Statistical Methods in Drug Combination Studies; Chapter: A Bayesian industry approach to phase I combination trials in oncology*. 2015.

D. Newell. Pharmacologically based phase I trials in cancer chemotherapy. *Hematology-Oncology Clinics of North America*, 8(2):257–275, 1994.

S. Nicholson, M. Krailo, M. Ames, N. Seibel, J. Reid, W. Liu-Mares, L. Vezina, A. Ettinger, and G. Reaman. Phase I study of Temozolomide in children and adolescents with recurrent solid tumors: A report from the children's cancer group. *Journal of Clinical oncology*, 16(9):3037–3043, 1998.

A. O'Hagan and J. Stevens. Bayesian methods for design and analysis of cost-effectiveness trials in the evaluation of health care technologies. *Statistical Methods in Medical Research*, 11:469–490, 2002.

J. O'Quigley. Curve-free and model-based continual reassessment method designs. *Biometrics*, 58(1):245–249, 2002.

J. O'Quigley and L. Shen. Continual reassessment method: A likelihood approach. *Biometrics*, 52:673–684, 1996.

J. O'Quigley, M. Pepe, and L. Fisher. CRM: A practical design for phase I clinical trials in cancer. *Biometrics*, 46(1):33–48, 1990.

J. O'Quigley, L. Shen, and A. Gamst. Two-sample continual reassessment method. *Journal of Biopharmaceutical Statistics*, 9(1):17–44, 1999.

K. Owzar and S-H. Jung. Designing phase II studies in cancer with time-to-event endpoints. *Clinical Trials*, 5(3):209–221, 2008.

S. Paul, D. Mytelka, C. Dunwiddie, C. Persinger, B. Munos, S. Lindborg, and A. Schacht. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Reviews: Drug Discovery*, 9:203–214, 2010.

Y. Pawitan. *In all likelihood: statistical modelling and inference using likelihood*. Oxford University Press, 2001.

Piantadosi and Liu. Improved designs for dose-escalation studies using pharmacokinetic measurements. *Statistics in Medicine*, 15:1605–1618, 1996.

S. Piantadosi. *Clinical trials: A methodological perspective*. Wiley Series in Probability and Statistics. John Wiley & Sons, 1 edition, 1997.

S. Pocock. *Clinical trials: A practical approach*. John Wiley & Sons, 2004.

R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

M. Ramond, T. Poynard, B. Rueff, P. Mathurin, C. Théodore, J-C. Chaput, and J-P. Benhamou. A randomized trial of Prednisolone in patients with severe alcoholic hepatitis. *The New England Journal of Medicine*, 326(8):507–512, 1992.

M. Ratain and D. Sargent. Optimising the design of phase II oncology trials: The importance of randomisation. *European Journal of Cancer*, 45:275–280, 2009.

E. Reiner, X. Paoletti, and J. O'Quigley. Operating characteristics of the standard phase I clinical trial design. *Computational Statistics and Data Analysis*, 30:303–315, 1999.

C. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Texts in Statistics. Springer New York, 2005.

C. Robert, L. Thomas, I. Bondarenko, S. O'Day, J. Weber, C. Garbe, C. Lebbe, J-F. Baurain, A. Testori, J-J. Grob, N. Davidson, J. Richards, M. Maio, A. Hauschild, W. Miller, P. Gascon, M. Lotem, K. Harmankaya, R. Ibrahim, S. Francis, T-T. Chen, R. Humphrey, A. Hoos, and J. Wolchok. Ipilimumab plus Dacarbazine for previously untreated metastatic melanoma. *New England Journal of Medicine*, 364: 2517–2526, 2011.

A. Rodrigues. *Drug-drug interactions*. Drugs and the pharmaceutical sciences. Informa Healthcare, 2 edition, 2008.

A. Rogatko, J.S. Babb, M. Tighiouart, F.R. Khuri, and G. Hudes. New paradigm in dose finding trials: Patient specific dosing and beyond phase I. *Clinical Cancer Research*, 11:5342–5346, 2005.

A. Rogatko, D. Schoeneck, W. Jonas, M. Tighiouart, F. Khuri, and A. Porter. Translation of innovative designs into phase I trials. *Journal of Cinical Oncology*, 24(31): 4982–4986, 2007.

W. Rosenberger and L. Haines. Competing designs for phase I clinical trials: A review. *Statistics in Medicine*, 21:2757–2770, 2002.

P. Royston and M. Parmar. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in Medicine*, 30(19):2409–2421, 2011.

F. Scheipl. *Bayesian regularization and model choice in structured additive regression.* PhD thesis, Ludwig-Maximilians-Universität München, 2011.

E. Schmid and D. Smith. Is pharmaceutical R&D just a game of chance or can strategy make a difference? *Drug Discovery Today*, 9(1):18–26, 2004.

S. Scott. *BoomSpikeSlab: MCMC for spike and slab regression.*, 2014. URL `http://CRAN.R-project.org/package=BoomSpikeSlab`. R package version 0.4.1.

L. Seymour, S. Ivy, D. Sargent, D. Spriggs, L. Baker, L. Rubinstein, M. Ratain, M. Le Blanc, D. Stewart, D. Crowley, S. Groshen, J. Humphrey, P. West, and D. Berry. The design of phase II clinical trials testing cancer therapeutics: Consensus recommendations from the clinical trial design task force of the National cancer institute

investigational drug steering committee. *Clinical Cancer Research*, 16:1764–1769, 2010.

R. Simon. Clinical trials and sample size considerations: Another prerspective: Comment. *Statistical Science*, 15:103–105, 2000.

K. Sinclair and A. Whitehead. A Bayesian approach to dose-finding studies for cancer therapies: Incorporating later cycles of therapy. *Statistics in Medicine*, 33:2665–2680, 2014.

D. Spiegelhalter, K. Abrams, and J. Myles. *Bayesian approaches to clinical trials and health-care evaluation*. Statistics in Practice. Wiley, 2004.

N. Stallard. *Statistical advances in the biomedical sciences: Clinical trials, epidemiology, survival analysis, and bioinformatics - Chapter 2: Phase II clinical trials*. Wiley Series in Probability and Statistics. Wiley, 2008.

Stan Development Team. *Stan modeling language: User's guide and reference manual*, 2012.

Stan Development Team. *Stan: A C++ library for probability and sampling, version 2.2.0*, 2013.

B. Storer. Design and analysis of phase I clinical trials. *Biometrics*, 45(3):925–937, September 1989.

J. Sühnel. Parallel dose-response curves in combination experiments. *Bulletin of Mathematical Biology*, 60:197–213, 1998.

X. Sun, P. Peng, and D. Tu. Phase II cancer clinical trials with a one-sample log-rank test and its corrections based on the Edgeworth expansion. *Contemporary Clinical Trials*, 32:108–113, 2011.

M. Sweeting and A. Mander. Escalation strategies for combination therapy phase I trials. *Pharmaceutical Statistics*, 11:258–266, 2012. And corresponding presentation slides.

R. Temple. Enrichment designs: Efficiency in development of cancer treatments. *Journal of Clinical Oncology*, 23(22):4838–4839, 2005.

P. Thall and J. Lee. Practical model-based dose-finding in phase I clinical trials: Methods based on toxicity. *International Journal of Gynecological Cancer*, 13(3): 251–261, 2003.

P. Thall, R. Millikan, P. Mueller, and S-J. Lee. Dose-finding with two agents in phase I oncology trials. *Biometrics*, 59:487–496, 2003.

P. Thall, L. Wooten, and N. Tannir. Monitoring event times in early phase clinical trials: Some practical issues. *Clinical Trials*, 2:467–478, 2005.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

M. Tighioutart, G. Cook-Wiens, and A. Rogatko. Escalation with overdose control using ordinal toxicity grades for cancer phase I clinical trials. *Journal of Probability and Statistics*, 2012, 2012.

R. Tsutakawa. *Bayesian inference for bioassay [Technical report No. 52 (University of Missouri-Columbia)].* 1975.

R. Tüchler. Bayesian variable selection for logistic models using auxiliary mixture sampling. *Journal of Computational and Graphical Statistics*, 17:76–94, 2008.

H. Wagner and C. Duller. Bayesian model selection for logistic regression models with random intercept. *Computational statistics and data analysis*, 56:1256–1274, 2012.

K. Wang and A. Ivanova. Two-dimensional dose finding in discrete dose space. *Biometrics*, 61:217–222, 2005.

G. Wheeler, M. Sweeting, and A. Mander. Bayesian penalized D-optimality for dual-agent phase I oncology trials. [cited January 2014] 2014. URL `http://www.mrc-bsu.cam.ac.uk/wp-content/uploads/3_Wheeler.pdf`.

J. Whitehead. *The design and analysis of sequential clinical trials.* Statistics in Practice. John Wiley & Sons, 2 edition, 1997.

J. Whitehead. Predicting the duration of sequential survival studies. *Drug Information Journal*, 35:1387–1400, 2001.

J. Whitehead. One- and two-stage designs for phase II clinical trials with survival endpoints. *Statistics in Medicine*, 33:3830–3843, 2014.

J. Whitehead and D. Williamson. Bayesian decision procedures based on logistic regression models for dose-finding studies. *Journal of Biopharmaceutical Statistics*, 8(3):445–467, 1998.

J. Whitehead, Y. Zhou, S. Patterson, D. Webber, and S. Francis. Easy-to-implement Bayesian methods for dose-escalation studies in healthy volunteers. *Biostatistics*, 2:47–61, 2001.

J. Whitehead, Y. Zhou, A. Mander, S. Ritchie, A. Sabin, and A. Wright. An evaluation of Bayesian designs for dose-escalation studies in healthy volunteers. *Statistics in Medicine*, 25:433–445, 2006a.

J. Whitehead, Y. Zhou, J. Stevens, G. Blackey, J. Price, and J. Leadbetter. Bayesian decision procedures for dose-escalation based on evidence of undesirable events and therapeutic benefit. *Statistics in Medicine*, 25:37–53, 2006b.

J. Whitehead, Y. Zhou, L. Hampson, E. Ledent, and A. Pereira. A Bayesian approach for dose-escalation in a phase I clinical trial incorporating pharmacodynamic endpoints. *Journal of Biopharmaceutical statistics*, 17:1117–1129, 2007.

J. Whitehead, E. Valdes-Marquez, P. Johnson, and G. Graham. Bayesian sample size for exploratory clinical trials incorporating historical data. *Statistics in Medicine*, 27:2307–2327, 2008.

J. Whitehead, H. Thygesen, and A. Whitehead. A Bayesian dose-escalation procedure for phase I clinical trials based only on the assumption of monotonicity. *Statistics in Medicine*, 29:1808–1824, 2010.

J. Whitehead, H. Thygesen, and A. Whitehead. Bayesian procedures for phase I/II clinical trials investigating the safety and efficacy of drug combinations. *Statistics in Medicine*, 30:1952–1970, 2011.

J. Whitehead, S. Tishkovskaya, J. O'Connor, and B. Damato. Devising two-stage and multi-stage phase II studies on systemic adjuvant therapy for uveal melanoma. *Investigative Ophthalmology and Visual Science*, 53:4986–4989, 2012.

IPCS WHO. Environmental health criteria 155, biomarkers and risk assessment: Concept and principles. *World Health Organization, Geneva*, 1993.

M. Wijesinha and S. Piantadosi. Dose-response models with covariates. *Biometrics*, 51:977–987, 1995.

World Medical Association et al. World Medical Association Declaration of Helsinki. Ethical principles for medical research involving human subjects. *The Journal of the American Medical Association*, 310(20):2191, 2013.

G. Yin and Y. Yuan. Bayesian dose finding in oncology for drug combinations by copula regression. *Journal of the Royal Statistical Society, Series C*, 58:211–224, 2009.

Y. Yuan and Y. Guosheng. Sequential continual reassessment method for two-dimensional dose finding. *Statistics in Medicine*, 27(27):5664–5678, 2008.

B. Zaslavsky. Bayesian sample size estimates for one sample test in clinical trials with dichotomous and countable outcomes. *Statistics in Biopharmaceutical Research*, 4: 76–85, 2012.

B. Zaslavsky and J. Whitehead. Letter to the editor. *Statistics in Biopharmaceutical Research*, 4:395, 2012.

W. Zhang, D. Sargent, and S. Mandrekar. An adaptive dose-finding design incorporating both toxicity and efficacy. *Statistics in Medicine*, 25:2365–2383, 2006.

L. Zhao, J. Taylor, and S. Schuetze. Bayesian decision theoretic two-stage design in phase II clinical trials with survival endpoint. *Statistics in Medicine*, 31:1804–1820, 2012.

Y. Zhou, J. Whitehead, P. Korhonen, and M. Mustonen. Implementation of a Bayesian design in a dose-escalation study of an experimental agent in healthy volunteers. *Biometrics*, 64:299–308, 2008.