DOCTORAL THESIS

# Methods to Increase Efficiency in Clinical Trials with Restricted Sample Size


*Author:*
Kristian BROCK

*Supervisor:*
Prof. Lucinda BILLINGHAM
Dr. Christina YAP


*A thesis submitted in fulfillment of the requirements*
*for the degree of Doctor of Philosophy*


Institute of Cancer and Genomic Sciences
College of Medical and Dental Sciences
University of Birmingham

February 19, 2019

# *Abstract*

**Methods to Increase Efficiency in Clinical Trials with Restricted Sample Size**

by Kristian BROCK

Efficiency is a perennial motivation of statistical analysis and clinical trials. This is most pertinent when sample size is constrained. When trials and their analyses are more efficient, results can be more precise, can be disseminated quicker, and impact the clinical pathway faster.

This thesis describes methods developed and investigated by the author in three trials at the Cancer Research UK Clinical Trials Unit. Methods for seamless phase I/II trials that conduct dose-finding by efficacy and toxicity outcomes are studied. A repeated measures analysis in an ultra-rare disease yields a feasible trial where standard approaches do not. Finally, this thesis develops methods for a phase II trial with co-primary outcomes and predictive covariate information.

We conclude that two common goals to increase efficiency are: i) use more outcomes to answer trial questions; and ii) use all available information. In our examples, analysing efficacy and toxicity in dose-finding lets these trials simultaneously achieve phase I and II objectives. However, this thesis highlights operational issues that can impair efficiency. We show that statistical performance is improved by analysing the information in repeated measures and predictive baseline covariates. Methods developed herein help to achieve conventional error rates without prohibitive increases in sample size.

*This thesis is dedicated to my two most enduring loves and my two most reliable diversions, Isabella (aged four) and Elodie (aged two). Girls, I do not imagine you will always love school. But hopefully, in time, you will agree that education is the best thing we have, and that we are never poorer for knowing a little more.*

# *Acknowledgements*

I owe a debt of gratitude to many people for making this body of research possible.

I thank Mhairi Copland, Tim Barrett and Gary Middleton, chief investigators of the Matchpoint, TreatWolfram and PePS2 trials respectively, for allowing me to conduct methodological research on their trials. If it ever struck them that it would be less hassle to just use an off-the-shelf approach, they were generous enough never to show it. I thank Peter Thall, Richard Herrick and Clift Norris of the MD Anderson Cancer Center for their kind advice on EffTox and the official EffTox software. I thank Nolan Wages for sharing R code to run the WT design. Chapter 4 and the TreatWolfram trial would be immeasurably poorer had Tammy Hershey not allowed us to use her patient data.

I thank Andrew Beggs and Alan Girling for conducting internal annual monitoring of this thesis. I thank Peter Thall and Emily Dressler for reviewing the EffTox in Matchpoint manuscript, and providing valuable feedback that improved the research and led to its publication. I extend thanks to the anonymous peer-reviewers who reviewed an earlier incarnation of the PePS2 design, particularly the reviewer who pointed out that it is a special case of Thall, Nguyen & Estey's design.

I undoubtedly owe a debt of gratitude to my supervisors, Lucinda Billingham and Christina Yap, not least for politely ignoring me when I initially declined to study this degree. They have generously given their time to read, critique, and discuss this thesis, and it plainly would not have happened without their dedication. Their reward is that they do not need to read it again (hopefully).

Finally, I thank my wife, my parents, and my wife's parents for entertaining my children whilst I worked. Whenever I was hunched over my laptop in the evening or on the weekend or on holiday, this is what I was doing. When my wife wrote her thesis seven years ago, she thanked me for supporting her. She should be under no doubt that she has more than repaid her debt. I am looking forward to a holiday with her soon where I only bring fiction to read...

# *Publications related to this thesis*

## Papers

**Brock, K.**, Billingham, L., Copland, M., Siddique, S., Sirovica, M., & Yap, C. (2017). Implementing the EffTox dose-finding design in the Matchpoint trial. *BMC Medical Research Methodology*, 17(1), 112. https://doi.org/10.1186/s12874-017-0381-x

## Oral Presentations

**Brock, K.**, Barrett, T., Yap, C., Storey, R. & Billingham, L. Assessing the design of our trial in an ultra-rare condition by the Parmar et al. framework for trials in smaller populations, ISCB Vigo, Jul 2017

**Brock, K.**, Yap, C., Llewellyn, L., Smith, H., McNab, G., Middleton, G. & Billingham, L. A Design for Phase II Clinical Trials in Stratified Medicine with Efficacy and Toxicity Outcomes & Predictive Variables, ISCB Birmingham, Aug 2016

**Brock, K.**, Billingham, L., Copland, M. & Yap, C. Seamless Phase I/II Adaptive Dose-finding Design with Efficacy-Toxicity Trade-offs for Targeted Therapies in Oncology. 36th SCT Arlington, May2015

**Brock, K.**, Billingham, L., Copland, M. & Yap, C. Development of an adaptive dose-finding design with non-monotonic dose-efficacy relationships, motivated by a phase I/II trial in chronic myeloid leukaemia. *Symposium on Early Phase Dose Finding Methodology*, Paris, April 2015.

## Poster Presentations

**Brock, K.**, Billingham, L., Nagy, Z., Hershey, T., Smith, H., Barton, D., & Barrett, T. (2017). Design of a practice-changing trial in the ultra-rare condition of Wolfram Syndrome. 4th ICTMC & 38th SCT Liverpool, May2017; published in *Trials*, p. 175, https://trialsjournal.biomedcentral.com/track/pdf/10.1186/s13063-017-1902-y?site=trialsjournal.biomedcentral.com

**Brock, K.**, Billingham, C., Llewellyn, L., Smith, H., McNab, G., & Middleton, G. (2017). 165: PePS2: Pembrolizumab in Performance Status 2 non-small-cell lung cancer. BTOG, Jan 2017; published in *Lung Cancer*, 103, S76. https://doi.org/10.1016/S0169-5002(17)30215-5

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Aims of this thesis

There are instances in clinical trials where innovative or uncommon methods are used to increase efficiency, particularly when the feasible sample size is smaller than would generally be desirable. This thesis describes in detail a number of methods that have been used and developed by the author in three clinical trials at the Cancer Research UK Clinical Trials Unit of the University of Birmingham.

In this section, we proceed with a brief introduction to clinical trials. We define efficiency in this scientific context, elaborating on the significance of a restricted sample size. We then give a brief overview of the chapters in this thesis, highlighting the novel elements of each.

## 1.2 Clinical trials

Clinical trials are medical experiments on human subjects. They are generally sequential in nature, with treatments typically passing through trials at phases I, II and III before being accepted as part of the standard of care. The collective aim is to learn about clinical interventions so that the conditions that impair our health may be treated and the overall health of the population may be maintained or improved. The objectives of the individual trial phases are specialised to reflect that this onerous task is tackled in stages. The trial phases are not particularly well defined, so considerable heterogeneity is seen. However, a generally accepted pathway to market authorisation for a drug could be described as follows.

Pre-clinical and animal studies yield information on the range of doses that might be tolerable and active in humans. The typical objective of a phase I trial is to select from this dose range the most attractive dose for further investigation in subsequent trials. What constitutes *attractive* varies by scenario, but it usually entails being tolerable to most patients. It might also entail being sufficiently active, in some clinical or pharmacological sense. Sometimes phase I trials are genuinely "first in man" scenarios.

In phase II, a typical objective is to assess early signs of efficacy at the dose selected at phase I. The dose may yet be adjusted as safety data continues to be collected, but the primary focus is to establish the presence of some therapeutic benefit. Trials may or may not be randomised, and the use of shorter-term predictive outcomes is relatively common to make trials quicker. Crucially, this step affords investigators the opportunity to assess whether an expensive and lengthy phase III trial is warranted.

If the treatment looks to hold promise, it is then typically tested at phase III against the standard of care. Phase III trials are usually randomised, and often blinded. Generally, in the case of drugs, the results of phase III trials are used to support applications for marketing authorisation so that the treatment may become part of the standard of care.

## 1.3 Efficiency in Clinical Trials

In their recent review "Improving Clinical Trial Efficiency: Thinking Outside the Box", Mandrekar *et al.*[62] describe some novel approaches to clinical trials that seek to increase efficiency. They identify that efficiency equates to "reduced sample size requirements". This is probably the most common interpretation of efficiency in the clinical trial context. Simon & Maitournam[83] also use this definition in their article evaluating the efficiency of trial designs that seek to allocate patients to treatments based on the presence of molecular targets. If trial design A can expect to arrive at a conclusion subject to given statistical error rates, requiring fewer patients than design B, then A can be said to be more efficient than B.

However, sample size is not the sole resource that is sought to be optimised in the pursuit of increased efficiency in trials. In this thesis, we define efficient methods in clinical trials to be those that reduce the expected resource required for a trial or sequence of trials to achieve their objectives. In addition to patients, two other resources invariably required in clinical trials are time and money. Thus, for instance, an approach that reduces the expected amount of time required to satisfactorily conduct a sequence of trials may also be regarded as efficient.

Efficiency is a perennial motivation in the design and analysis of clinical trials because we want to share the benefits of health research as quickly as possible. The sooner that good treatments are approved, the sooner they may benefit the diseased population. The sooner that unacceptable treatments are discarded, the more resources will be dedicated to investigating alternatives that could provide benefit.

Broadly speaking, efficiency can be imparted on clinical trials in two main ways: through *operational* and *statistical* methods. Operational methods are those that alter the way clinical trials are conducted. Efficiency is garnered when a trial is able to achieve its objective faster, or achieve the objectives of several trials faster than conducting them separately.

PICO, a mnemonic that lists the core defining elements of a comparative clinical trial, stands for *Population, Intervention, Comparison, Outcome*. In this section, it may serve to illustrate how clinical trials have typically been conducted, as a hypothetical traditional trial would have a single response to each item. The article by Mandrekar *et al.*[62] focuses on so-called basket and umbrella trials. These augment the PICO approach by investigating a single novel intervention in several disease populations, and multiple novel interventions in a single population, respectively. The rationale is that several similar questions can likely be more quickly answered in a single larger trial than separate parallel trials, when we consider the fixed costs in time and money required to conduct a clinical trial. The marginal cost of adding to an existing protocol an extra patient population or novel intervention, keeping the other elements of PICO constant, is likely to be less than those required to run a completely new trial. These examples demonstrate how operational efficiency can be achieved.

Another example of operational efficiency comes from achieving the objectives of two consecutive trial phases in a single over-arching *seamless* trial. So-called phase

I/II trials conduct dose-finding whilst assessing efficacy in addition to toxicity in an experimental treatment, thus achieving the objectives of trials at phases I and II. There have been a number of designs proposed in this area[12, 92, 98, 112]. These designs are the subject of Chapters 2 and 3 in this thesis. Further examples of seamless methodology are multi-arm multi-stage (MAMS) trials[77], that fuse the traditional objectives of phases II and III. However, these are not a particular focus of this work.

There are typically many possible ways to analyse a dataset. In the context where a decision must be made on the acceptability of a treatment, the type of decision generally made in trials, statistical methods may be considered efficient when they achieve given error rates with a smaller sample size. In a clinical trial, those *errors* typically involve concluding a novel treatment is superior to a comparator when it truly is not; and rejecting a novel treatment when it truly is superior. Error rates are expected to reduce as sample size is increased and more information is provided to the analysis algorithm. Some algorithms are able to incorporate information from other sources to increase efficiency and outperform alternatives that have no such facility. This is the focus of Chapters 4, 5 and 6.

## 1.4 Restrictions on Sample Size

Sample sizes are constrained in clinical trials for many reasons. The most intuitive is that the number of patients that may be required to conduct a conventional analysis simply does not exist. Obviously, this is particularly pertinent in rare diseases. However, it can be very difficult to recruit to trials in relatively common diseases like lung cancer and leukaemia, if a particular disease subtype or patient characteristic is sought. In this regard, every disease has the capacity to suffer from a small recruitment pool.

Elsewhere, the feasible recruitment level in a trial can be constrained for reasons other than patient availability. For instance, with novel therapeutics that are difficult or expensive to manufacture, it may transpire that only a small number of patients can be treated. If research funds are constrained, as they so often are, only a small trial may be possible. Finally, time is often a limited commodity. There is a strong motivation to conduct phase I and II trials quickly so that the time novel treatments

spend in trials can be reduced and effective new treatments delivered to patients in a timely manner.

There are many reasons that sample sizes are restricted and we will encounter several of them in this thesis. When sample sizes are restricted, trialists have strong motivation to use efficient methods to make the most of the available information.

## 1.5 Chapters in this thesis

It is preferable to pass through the trial phases as quickly as possible. A seamless phase I/II trial that achieves both the traditional objectives of separate phase I and II trials is efficient if it is faster and cheaper than separate trials. Seamless phase I/II trials are the focus of Chapters 2 and 3.

In Chapter 2, we describe our experiences using the EffTox dose-finding design[92] in the Matchpoint trial. The design estimates the rates of binary efficacy and toxicity events at a range of different doses using logit models. The probability model uses six parameters in total. A feature of dose-finding trial designs is that they must make inferences when very few patient outcomes are observed. For example, in typical dose-finding scenarios, patients are treated in cohorts of three. Once the first cohort is treated and assessed, the trial design advises the dose for the second cohort based on the outcomes observed hitherto. In this scenario with the EffTox model, there are fewer patients than parameters so inferences are subject to great amounts of uncertainty. We introduce the phenomenon of *dose ambivalence* where the design can recommend different doses in response to identical outcomes, and advocate a simulation-based method to overcome the ensuant uncertainty. We also describe the challenge arising from outcome ambiguity, and advocate a practical solution using dose-transition pathways in the phase I/II setting. Our methods promote efficiency by aiding the selection of the optimal dose and overcoming delays in the assessment of outcomes.

Chapter 3 introduces a novel design for seamless phase I/II clinical trials. It fuses elements from EffTox and an alternative design by Wages & Tait[98] that uses adaptive randomisation to explore the doses under investigation. Wages & Tait's method

uses a simpler probability model that requires fewer parameters than EffTox. However, randomisation brings potential operational complexity. Our motivation was to create a hybrid design that uses fewer parameters than EffTox in the hope that this would maintain statistical efficiency, whilst abrogating the need for randomisation and reducing possible administrative inefficiency. We present the design and investigate performance in Matchpoint scenarios in Chapter 3.

The desire for efficiency is not unique to early phase trials. It is important whenever sample size is constrained, and felt particularly acutely in rare diseases in pivotal settings. This is the topic of Chapter 4, where we describe a design for a randomised controlled trial in an ultra-rare disease. A traditional experiment where outcomes are compared at the end of an intervention period would require a sample size that exceeds the total number of patients in the UK to achieve conventional error rates. We describe an approach using repeated measures, linear hierarchical models, and simulation to design a trial that is feasible and defensible. Key to achieving this was using all of the information in the repeated measures. The flexibility of the simulation method allowed us to examine expected power under different patterns of missing data. Our motivation for this level of scrutiny was the severely constrained sample size. Our literature review shows that this approach is novel in clinical trials of visual acuity, and indicative of other scenarios where repeated measures analyses are possible but generally not conducted.

Patient numbers might be constrained in otherwise common diseases because of specific eligibility criteria. This is the focus of Chapter 5 where we introduce a novel adaptation to the design of Thall, Nguyen & Estey[89] to assess an immunotherapy drug in a specific subgroup of non-small-cell lung cancer patients. Lung cancer is a regrettably common disease but subgroups can be quite small once molecular stratifiers are included. Our design incorporates baseline predictive information to increase efficiency when simultaneously assessing efficacy and toxicity outcomes in a phase II trial. It achieves statistical operating performance superior to benchmark designs that assess treatment cohort-by-cohort. One of the goals of stratified medicine is to tailor treatments by patient subgroup. We use predictive categorical variables, presented and validated in a trial of a related patient group, to increase efficiency in estimating the event rates of our co-primary outcomes. The design

satisfies an otherwise unmet need that will become more common as biomarker-associated therapies are further investigated.

In Chapter 6, we extend further the scenario in Chapter 5 by considering alternative model specifications. Our chosen model forms were motivated by the available literature and our feasible sample size. However, the implicit assumptions were potentially undesirable. We research the implications of more complex model forms, and discuss the trade-off of a more flexible model with the attendant greater resource requirements. We describe how marginal further efficiencies are available at relatively little marginal cost.

Finally, in Chapter 7 we discuss the broad themes spanned by the topics in the contained chapters, and highlight some motivations for further work.

## Chapter 2

# Implementing EffTox in the Matchpoint Trial

**Background:** Methods for phase I/II dose-finding use efficacy outcomes in addition to toxicity outcomes to identify the most attractive dose. EffTox is one of the earliest and best-known. The Matchpoint trial uses EffTox to search for an effective and tolerable dose of ponatinib to combine with FLAG-IDA chemotherapy.

**Notable methods in this chapter:** We describe a nomenclature for succinctly describing outcomes in phase I/II dose-finding trials. We use *dose-transition pathways* in the phase I/II setting, where doses are calculated for each feasible set of outcomes in future cohorts. We introduce the phenomenon of *dose ambivalence*, where EffTox can recommend different doses after observing the same outcomes. We also describe our experiences with *outcome ambiguity*, where the categorical evaluation of some primary outcomes is temporarily delayed.

**The implications on efficiency:** Phase I/II trials are efficient because they allow the objectives of two trial phases to be addressed at once. However, phenomena like dose ambivalence and outcome ambiguity stand to erode that efficiency by allowing sub-optimal doses to be selected and causing delays in the assessment of outcomes. The methods we introduce show how those complications can be managed and overcome. Furthermore, our methods facilitate efficient trial planning and conduct.

## 2.1 Introduction

The introduction of BCR-ABL tyrosine kinase inhibitors (TKIs; imatinib, dasatinib, nilotinib, bosutinib and ponatinib) has revolutionised the treatment of chronic myeloid leukaemia (CML). The great majority of patients with chronic phase (CP)-CML obtain a durable complete cytogenetic response and the rate of progression to blast phase (BP) is 1 to 2% per annum in the first few years after diagnosis, falling sharply when major molecular response is obtained[34, 47, 58]. A minority of patients (<10%) present with *de novo* BP-CML and of these two-thirds are myeloid and one-third lymphoid BP[43]. Despite the use of TKIs, median survival after the diagnosis of BP-CML is between 6.5 and 11 months[33, 41, 71, 78], with the majority of long-term survivors being recipients of allogeneic stem cell transplant in second chronic phase of disease[48]. This poor survival is often due to patients developing new mutations, most frequently within the BCR-ABL kinase domain, resulting in resistance to TKIs and further rapid disease progression[85]. Therefore, novel therapies to improve and prolong therapeutic responses in BP-CML are urgently sought.

In the Matchpoint trial (EudraCT 2012-005629-65) we plan to simultaneously assess co-primary safety and efficacy outcomes for the combination of a novel TKI, ponatinib, with conventional FLAG-IDA chemotherapy. We believe this to be the first such study in blastic phase CML. It is envisaged that the data will be the first step to improve the treatment of this difficult clinical problem.

Historically, dose-finding trials in oncology have sought to find the *maximum tolerable dose* (MTD) of a treatment under the *cytotoxic* assumption. Rule-based designs like 3+3 change dose based on the number of dose-limiting toxicities (DLTs) observed. Using a model-based design like the seminal continual reassessment method[69] (CRM), dosage is increased to find the dose with an associated probability of toxicity that is less than (or close to) a pre-specified threshold. The rate of efficacy does not directly affect the dose selection decision. Instead, it is assumed to increase monotonically with the probability of toxicity and dose. This has been a valid assumption in treatments like chemotherapy, that kill diseased and non-diseased cells alike. A notable advantage of the cytotoxic assumption is that it simplifies the mathematics when calculating the ideal dose.

Increasingly, modern treatments like molecularly targeted agents and immunotherapies are being investigated for their therapeutic effects in oncology. Targeted therapies work by altering the behaviour of cells at a molecular level to slow or stop the malignant proliferation. Immunotherapies work by instigating a response from the patient's immune system to fight disease. A positive outcome like longer survival may be achieved whilst containing the aggregate disease burden, rather than reducing it. With each of these classes of treatment, the cytotoxic assumption is not necessarily valid so we can no longer assume that the most toxic dose is the most efficacious dose. This presents a methodological challenge to investigators in dose-finding trials. The goal here is to find the *optimal* dose rather than merely the maximum tolerable dose. We may regard the optimal dose as that which provides the most attractive trade-off between the probabilities of efficacy and toxicity, or that which offers maximal chance of efficacy with the chance of toxicity less than some critical value. Generally, these targeted therapies and immunotherapies are less toxic than cytotoxic therapies, so the optimal dose may be much lower than the MTD[1]. In the so-called *cytostatic* setting, dosing decisions should be guided by patients' outcomes with regard to efficacy and toxicity, yielding designs for joint phase I/II trials.

Published clinical trial designs in this arena include extensions of CRM. Braun's bivariate CRM[12] models separate toxicity and disease progression events. Zhang *et al.*'s variant of CRM[112] uses an ordered trinary outcome that incorporates response and toxicity. More recently, Wages & Tait[98] introduced a method that uses a latent CRM model to monitor toxicity and selects amongst candidate efficacy models using Bayes factors. Amongst non-CRM alternatives, Wang & Day [99] introduce a utility-maximising approach that assumes responses and toxicity occur in patients according to log-normally distributed patient thresholds.

Thall & Cook[92] introduced EffTox, the method we chose to use in Matchpoint. EffTox is a Bayesian adaptive dose-finding trial design that models correlated binary efficacy and toxicity outcomes. A search of PubMed on 17th October 2016 for articles that have cited Thall & Cook[92] returned 54 items. Of these, 36 were methodological in nature, detailing extensions or alternative designs. A further 14 were review articles. Only four articles pertained to the design or reporting of a specific clinical

trial. Three of these used the EffTox design[4, 23, 81]. The first author is based at the MD Anderson Cancer Center for two of these papers[4, 81], and at the University of Washington for the third[23]. The fourth trial article[13] cites the EffTox paper but uses a randomised trial design. It is not our intention to give a full systematic review but this scoping search suggests that EffTox is not widely used, and scarcely used at all outside the USA. Thall[88] himself admitted that "[Bayesian models for early phase clinical trials] have seen limited use in clinical practice". In describing our experience using this important dose-finding clinical trial design, we hope to encourage others to use it too. Our proposed solutions to the problems we encountered will expedite the trial design process.

In Section 2.2 we recap the EffTox design. Section 2.3 details our rationale for choosing EffTox and our experience using it in Matchpoint, the problems we faced and the solutions we proposed. We provide some discussion in Section 2.4, culminating in some conclusions on the impact on clinical trial efficiency in Section 2.5.

## 2.2 The EffTox Design

Thall & Cook[92] introduced the adaptive Bayesian design *EffTox* to facilitate seamless phase I/II dose-finding. EffTox uses logit models for the marginal probabilities of efficacy and toxicity at each dose and utility contours to measure the attractiveness of each dose based on the posterior probabilities of efficacy and toxicity.

Let $\boldsymbol{y} = (y_1, ..., y_n)$ be the $n$ doses under investigation. Thall & Cook use the codified doses $\boldsymbol{x} = (x_1, ..., x_n)$:

$$x_i = \log y_i - \sum_{j=1}^{n} \frac{\log y_j}{n} \tag{2.1}$$

For example, a trial of 4 doses, 10mg, 20mg, 30mg and 50mg, would have $\boldsymbol{y} = (10, 20, 30, 50)$, and $\boldsymbol{x} = (-0.85, -0.16, 0.25, 0.76)$. These values are used as explanatory variables so it is desirable that they are centralised and relatively small in magnitude.

Using the notation of Thall & Cook[92], let $\boldsymbol{Y} = (Y_E, Y_T)$ be indicators of binary efficacy and toxicity events. Let $\pi_{a,b}(x, \theta) = \Pr(Y_E = a, Y_T = b|x, \theta)$ for $a, b \in \{0, 1\}$.

The marginal probabilities of efficacy and toxicity at dose $x$ are given by

$$\text{logit}\, \pi_E(x, \theta) = \mu_E + \beta_{E,1} x + \beta_{E,2} x^2 \tag{2.2}$$

and

$$\text{logit}\, \pi_T(x, \theta) = \mu_T + \beta_T x \tag{2.3}$$

When $\beta_T > 0$, the toxicity probabilities increase monotonically in dose. In contrast, the efficacy curve is not necessarily monotonically increasing. The presence of $\beta_{E,2}$ allows for non-linearity and possibly a turning point.

The joint probability model is

$$\begin{aligned}\pi_{a,b}(x, \theta) = (\pi_E)^a (1 - \pi_E)^{1-a} (\pi_T)^b (1 - \pi_T)^{1-b} \\ + (-1)^{a+b} (\pi_E)(1 - \pi_E)(\pi_T)(1 - \pi_T) \frac{e^\psi - 1}{e^\psi + 1}\end{aligned} \tag{2.4}$$

where $\psi$ is an association parameter and $(x, \theta)$-notation has been suppressed in each function for brevity. Here, $a, b$ are binary patient-specific variables that denote whether efficacy and toxicity events occurred. For a given patient, $a = 1$ means the patient experienced efficacy and $b = 1$ means they experienced toxicity.

The EffTox design requires several pieces of information to be elicited from the investigators. Firstly, the statistician must elicit the prior probability of efficacy and toxicity at each dose. Let us label the vector of efficacy probabilities $\eta_E$, and the toxicity analogue $\eta_T$. The EffTox software[45] published by the MD Anderson Cancer Center will take these prior beliefs and a desired *effective sample size* (ESS) and convert them into univariate normal priors on each component of $\boldsymbol{\theta} = (\mu_T, \beta_T, \mu_E, \beta_{E,1}, \beta_{E,2}, \psi)$. Thall *et al.*[94] detail the algorithm and advise that ESS should be between 0.5 and 1.5. High values for ESS reflect stronger prior information. The preference is for priors that are strong enough to sensibly guide early dosing decisions but weak enough to be overridden by patient outcomes where they diverge from prior beliefs.

Secondly, the statistician must elicit parameters to calculate the utility contours. Thall *et al.* discuss one particular method for this task[92, 93]. The points $(\pi^*_{1,E}, 0)$, $(1, \pi^*_{2,T})$ and $(\pi^*_{3,E}, \pi^*_{3,T})$ are elicited from the investigator such that the pairs are

equally attractive. The quantity $\pi^*_{1,E}$ is the minimum required probability of efficacy when toxicity is impossible. The quantity $\pi^*_{2,T}$ is the maximum permissible probability of toxicity when efficacy is guaranteed. The point $(\pi^*_{3,E}, \pi^*_{3,T})$ is chosen in the first quadrant (i.e. not lying on the $x$- or $y$-axis), representing a pair of probabilities for efficacy and toxicity that are of equal attractiveness as the two other points.

EffTox originally used inverse quadratic functions to model the utility contours but, after observing that the design was reticent was escalate to more efficacious doses, the authors later advocated using $L^p$ norms[27]. An $L^p$ norm is a mathematical tool for generally measuring the distance between two points. The best known is $L^2$, the Euclidean norm, that measures the length of a hypotenuse $c$ in a right triangle to satisfy $c^2 = a^2 + b^2$, where $a$ and $b$ are the lengths of the other two sides.

Thall *et al.*[94] stressed the importance of using contours that are steep enough to encourage the design to accept slightly higher probabilities of toxicity when they are compensated with materially higher probabilities of efficacy. This point was developed in detail in Yuan *et al.*[111]. In Figure 2.1, we see that the neutral utility contour in bold is practically vertical when the probability of toxicity belongs to $(0, 0.2)$, illustrating what we mean by a *steep* contour. Here, an equal absolute percentage increase in the probabilities of efficacy and toxicity will increase the utility score. In contrast, the neutral utility contour is *flatter*, or more horizontal, where the probability of efficacy belongs to $(0.8, 1.0)$. Here, an identical increase in the probabilities of efficacy and toxicity results in a decrease in utility. When the contours are too flat, pathological behaviour can manifest where the design becomes stuck at a sub-optimal dose. This point was unfortunately missed in earlier publications on EffTox[92, 93]. Furthermore, the illustrative example in the original EffTox paper[92] inadvertently uses a family of contours that exhibit pathological behaviour. In order to achieve a design with good properties, Thall advocates selecting three equivalent points that yield a reasonably steep contour, and not trying to elicit points of equal utility from clinicians. Fundamentally, trialists should note that EffTox has evolved since its original 2004 publication[92].

The utility of a dose with associated posterior efficacy and toxicity probabilities $\pi_E$ and $\pi_T$ is

$$u(\pi_E, \pi_T) = 1 - \left( \left( \frac{1 - \pi_E}{1 - \pi^*_{1,E}} \right)^p + \left( \frac{\pi_T}{\pi^*_{2,T}} \right)^p \right)^{\frac{1}{p}} \tag{2.5}$$

In (2.5), $p$ determines the extent of the curvature of the utility contours. For $p > 1$, the contours are convex and for $p = 1$, the contours are simply straight lines[27]. The value for $p$ is calculated by the EffTox software so that the neutral utility curve intersects $(\pi^*_{1,E}, 0)$, $(1, \pi^*_{2,T})$ and $(\pi^*_{3,E}, \pi^*_{3,T})$.

EffTox uses decision criteria to determine the set of admissible doses based on posterior beliefs. Given trial data for the first $j$ patients, $\mathcal{D} = \{(x_1, a_1, b_1), ..., (x_j, a_j, b_j)\}$, dose $x$ is admissible if

$$\Pr\{\pi_E(x, \boldsymbol{\theta}) > \underline{\pi}_E | \mathcal{D}\} > p_E \tag{2.6}$$

and

$$\Pr\{\pi_T(x, \boldsymbol{\theta}) < \overline{\pi}_T | \mathcal{D}\} > p_T \tag{2.7}$$

where $\underline{\pi}_E$ is a lower bound on the acceptable efficacy rate and $\overline{\pi}_T$ an upper bound on the toxicity rate. In order to resolve (2.6) and (2.7), a prior-to-posterior analysis must be carried out to combine the investigators' priors with $\mathcal{D}$. This involves solving a six-dimensional integral. The details are given in Thall *et al.*[92].

The investigators provide values for $\underline{\pi}_E$, $\overline{\pi}_T$, $p_E$ and $p_T$. The set of doses that are admissible is said to be the admissible set. When a dose selection decision is required (e.g. at the end of a cohort), the admissible set is recalculated. If no dose is admissible, the trial stops and no dose is selected for further research. This may occur if all of the doses are too toxic or insufficiently efficacious, or both. If the admissible set is non-empty, the dose with maximal utility, subject to rules about not skipping untested doses, is recommended to be given to the next cohort or patient.

This iterative process is repeated until the maximum sample size or some pre-defined stopping criteria is reached. The dose recommended after all patients have been treated and evaluated is the dose selected for further research in a later phase trial.

## 2.3 EffTox in the Matchpoint trial

The EffTox design was originally selected for use in Matchpoint by Christina Yap (CY), working with the chief investigator and co-investigators. She was aided in early trial design work by Josephine Khan (JK). The trialists chose to use a seamless phase I/II dose-finding design in Matchpoint because it would be more efficient that running separate trials in phases I and II. We wanted the observed efficacy events to influence the doses selected because there was clinical justification in suspecting cytostatic behaviour with respect to the experimental agent, discussed below. CY chose to use EffTox because of the readily-available MD Anderson software[45] with which to conduct a trial using the EffTox design. Critically, the software performs simulation studies, allowing trialists to hone parameter choices.

This section details the parameters chosen for EffTox in the Matchpoint trial, the practical issues we faced and how we surmounted them. A summary of our parameter choices appears in Table 2.1. These are discussed further below.

TABLE 2.1: EffTox parameters chosen in the Matchpoint trial. These are discussed in the main text.

| Notation | Interpretation | Value |
|---|---|---|
| $N$ | Total number of patients | 30 |
| $m$ | Cohort size | 3 |
| $p_E$ | Certainty required to infer dose is threshold efficable | 0.03 |
| $p_T$ | Certainty required to infer dose is threshold tolerable | 0.05 |
| $\underline{\pi}_E$ | Minimum efficacy threshold | 0.45 |
| $\overline{\pi}_T$ | Maximum toxicity threshold | 0.40 |
| $\pi^*_{1,E}$ | Required efficacy probability if toxicity is impossible | 0.40 |
| $\pi^*_{2,T}$ | Permissible toxicity probability if efficacy guaranteed | 0.70 |

The values $p_E$ and $p_T$ may seem unconventionally small. Recall that their function is to define a list of doses appropriate for experimentation. It may be more intuitive to interpret $1 - p_\cdot$ as the posterior certainty required to omit a dose from consideration. In their original demonstration, Thall & Cook[92] used values $p_E = p_T = 0.1$.

In Matchpoint, the binary efficacy event is achieved when patients experience at least a minor cytogenetic response (i.e. <65% Philadelphia chromosome-positive cells), or haematological response with platelets $> 50 \times 10^9$/L, neutrophils $> 1.0 \times 10^9$/L and blasts $< 5\%$ in the peripheral blood and bone marrow. The binary toxicity

outcome is defined by the occurrence of a range of pre-specified adverse events, including any grade 3 or 4 clinically significant non-haematological adverse event, related to ponatinib, that cannot be managed with optimal medical care and likely to endanger the life of the patient or result in long term effects. Both co-primary outcomes are assessed over the eight-week period following the commencement of the first cycle of treatment. The first cycle lasts for 28 to 56 days, depending on how long it takes for blood counts to recover.

Of practical importance when using a seamless phase I/II design is that the co-primary outcomes can be assessed over a similar time horizon. It was felt that responses to treatment could be expected after just one cycle if the treatment could be successfully administered to patients. If toxicity was frequent and treatment discontinuation common, the capacity for response is diminished. In a scenario where outcomes are assessed over materially different horizons, the trial would proceed at the speed determined by the outcome with the longest assessment period, increasing the risk of incomplete data and eroding the scope for operational efficiency.

### 2.3.1 Parameters

We investigate four doses of ponatinib: 15mg every second day, 15mg daily, 30mg daily and 45mg daily, referenced as dose-levels 1, 2, 3, 4 respectively, as shown in Table 2.2. For a tractable analysis, we use $\boldsymbol{y} = (7.5, 15, 30, 45)$, and thus $\boldsymbol{x} = (-0.97, -0.27, 0.42, 0.82)$.

Generally, the clinicians were comfortable providing their prior beliefs on the probability of efficacy and toxicity. These were elicited by CY and JK. The clinicians believed a-priori that all doses would be tolerable. There was some debate about the extent to which the probability of efficacy would improve when moving from the third to the highest dose. On balance, it was felt that efficacy would be low at the lowest doses, increase with dose throughout but begin to level-off at the highest dose. This yielded the priors shown in Table 2.2.

The clinicians were also comfortable specifying $\underline{\pi}_E$ and $\overline{\pi}_T$. Conventional chemotherapy regimens like FLAG-IDA can induce complete cytogenetic responses in 20-40% of patients who have progressed to blastic phase[97]. Cortes *et al.*[29] gave 45mg of ponatinib daily as a monotherapy to CML patients and observed major cytogenetic

TABLE 2.2: Doses under investigation in Matchpoint and the investigators' prior beliefs on rates of efficacy and toxicity. Note, the ponatinib dose labelled 7.5mg per day is actually 15mg every other day.

| Dose-level | Daily ponatinib dose (mg) | Prior Pr(Eff), $\eta_E$ | Prior Pr(Tox), $\eta_T$ |
|---|---|---|---|
| 1 | 7.5 | 0.2 | 0.025 |
| 2 | 15 | 0.3 | 0.05 |
| 3 (start dose) | 30 | 0.5 | 0.1 |
| 4 | 45 | 0.6 | 0.25 |

response in 23% of 62 patients in blast transformation phase. They also observed very good response rates in chronic phase patients. By combining the treatments, we hope to observe a response rate in excess of 45% so we used $\underline{\pi}_E = 0.45$. It was the clinicians' prior belief that only the highest two doses would exceed the minimum efficacy threshold. To achieve this level of efficacy, it was felt that a toxicity rate up to 40% would be acceptable thus we set $\overline{\pi}_T = 0.40$.

The first cohort will receive dose-level 3 (30mg) because this is the lowest dose believed a-priori to be sufficiently active. From here, there is scope to escalate or de-escalate dose as the outcomes dictate.

The values of $p_E$ and $p_T$ in (2.6) and (2.7) determine the posterior confidence required to admit the doses as worthy of investigation. Low values are chosen so that even relatively weak beliefs will render doses worthy of investigation in this early phase clinical trial. CY and JK initially proposed using the values $p_E = p_T = 0.05$ but later, after the author (KB) became involved in this trial, this was altered this to $p_E = 0.03$. The process of refining these values is described in Section 2.3.2.6.

In contrast, the clinicians found it rather more challenging to specify $(\pi^*_{1,E}, 0)$ and $(1, \pi^*_{2,T})$ because of the practical impossibility of a treatment where efficacy is guaranteed or toxicity impossible. Instead, KB and CY elicited $(\pi^*_{3,E}, \pi^*_{3,T})$, $(\pi^*_{4,E}, \pi^*_{4,T})$ and $(\pi^*_{5,E}, \pi^*_{5,T})$, three points in the general efficacy-toxicity space (i.e. not at the extremes) such that the points had equal utility. Using an $L^p$ norm to fit a curve like (2.5) to these points requires $u(\pi^*_{3,E}, \pi^*_{3,T}) = u(\pi^*_{4,E}, \pi^*_{4,T}) = u(\pi^*_{5,E}, \pi^*_{5,T}) = 0$. Thus, we have three simultaneous, non-linear equations with three unknown values:et: $\pi^*_{1,E}$, $\pi^*_{2,T}$ and $p$. We used the multi-variate solver `multiroot` in the R[76] package `rootSolve`[84] to find the simultaneous solution to these equations. The curve fitted to the elicited points (50%, 40%), (45%, 30%) and (70%, 60%) intercepts the axes

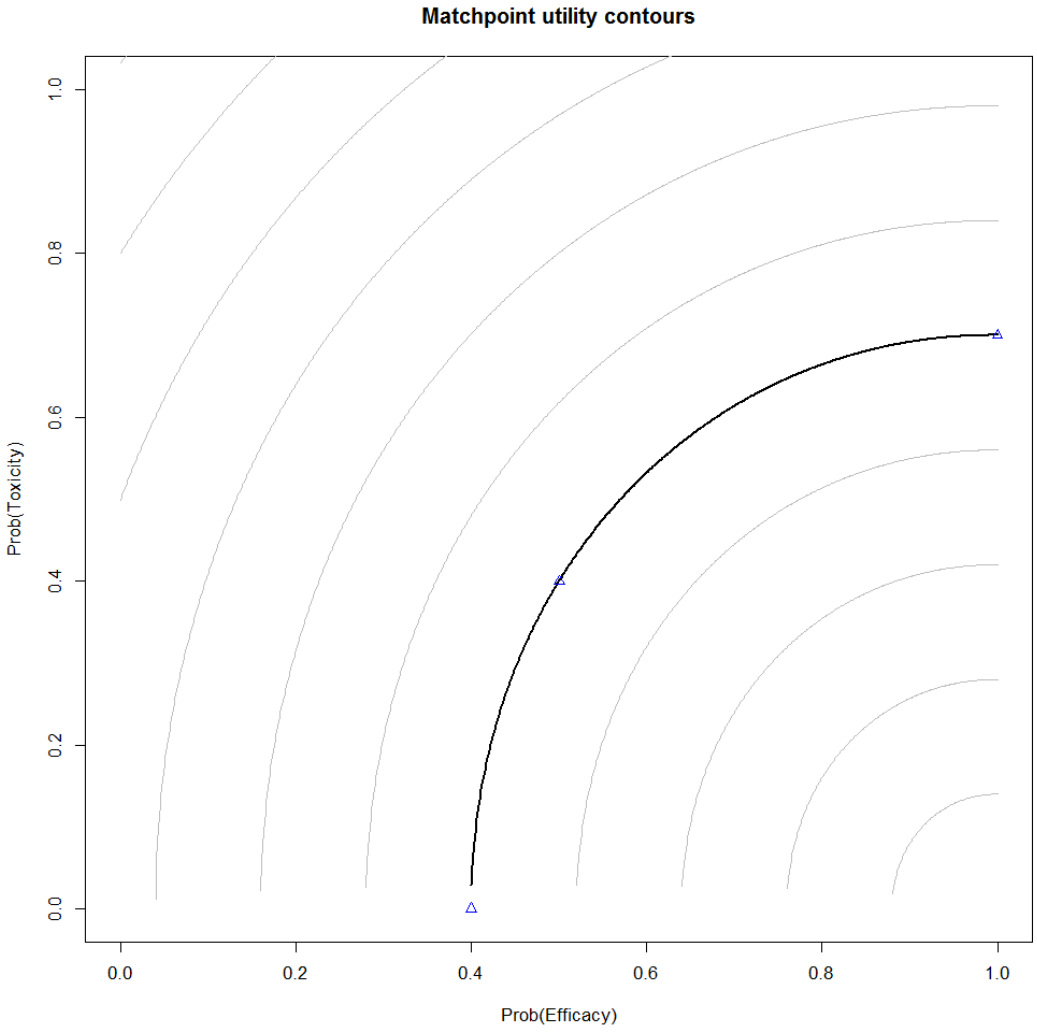



FIGURE 2.1: Utility contours in the Matchpoint trial. The neutral utility contour in bold joins (0.4, 0) to (1.0, 0.7). Points inside this contour have positive utility, increasing as they approach (1.0, 0.0).

at (39.6%, 0%) and (100%, 67.9%). We rounded to take $\pi^*_{1,E} = 0.40$ and $\pi^*_{2,T} = 0.70$. The revised curve actually intersects the points (50%, 40%), (45%, 29.3%) and (70%, 61.4%), as illustrated in Figure 2.1.

Finally, the value of ESS was chosen by trial-and-error. Thall, Cook and Estey[93] advise a value in the range (0.5, 1.5). Increasing ESS generally improves the performance of the design in scenarios that broadly agree with the prior beliefs, and vice-versa. Statisticians and investigators should, however, be mindful of the necessity for the data to override the prior in the event that the priors are wrong. This advocates exercising caution when using inflated ESS values. We arrived at ESS=1.3 because it yielded attractive simulated operating characteristics and sensible dose

transitions, as described in the following sections.

### 2.3.2 Trial Conduct

#### 2.3.2.1 Nomenclature for Describing Outcomes in Phase I/II Trials

To expedite the discussion of phase I/II clinical trial conduct, we introduce some nomenclature, created by KB for succinct interim reporting. Each patient may experience one of four specific outcomes: efficacy without toxicity (E); toxicity without efficacy (T); both (B); or neither (N). Let us string these symbols behind a numerical dose-level to denote the outcomes of cohorts of patients. For instance, 2EET denotes a cohort of three patients that were given dose-level 2, two of whom experienced efficacy only and one who experienced toxicity. These strings can be concatenated to describe the outcomes of several cohorts consecutively. For example the path 2EET 3EBB extends our previous scenario. After the first cohort, the trial escalated to dose-level 3. The next cohort of three were treated at this dose and all three patients experienced efficacy. Unfortunately, two of them also experienced toxicities. Using our notation, this information is unambiguously and efficiently conveyed in 8 characters.

In phase I/II, it is inadvisable to reduce patients' outcomes to simple tallies of efficacy and toxicity events because of the complication that patients may experience both events or neither. For instance, the design may recommend a different dose after observing NTE than it would after observing NNB, even though both cohorts contain a single efficacy event and a single toxicity event. In the first example, the events are experienced by different patients whereas in the latter, they are experienced by the same patient. The distinction is especially pertinent in EffTox because the $\psi$ parameter models the association between efficacy and toxicity.

The described notation combines simple codification of dose-levels and patient outcomes to succinctly and unambiguously describe pathways through phase I/II dose-finding trials. We use it in the next section to define dose transition pathways, and in following sections to discuss the potential problems of outcome ambiguity and dose ambivalence, and to aid trial planning.

#### 2.3.2.2 Dose Transition Pathways

We found it greatly beneficial to prospectively analyse how our dose-finding design would behave with respect to cohorts by supposing each feasible set of future patient outcomes and calculating the model advice in each. From a given starting point, we look to identify the conditions under which the design would escalate dose, stay at a dose, de-escalate dose, or recommend that the trial stops.

*Dose-transition pathways* (DTPs) were introduced by Yap *et al.*[108] in the context of traditional phase I trials with DLT outcomes. A DTP is a single feasible pathway through a dose-finding trial. It reflects the dose selections that a model would make in response to given hypothetical future outcomes. We introduce here the novel extension of Yap *et al.*'s idea to phase I/II trials with efficacy and toxicity outcomes.

The example in Table 2.3 shows the complete set of DTPs for cohort 2 having observed 3TTT in cohort 1. We see that after observing 3TTT, the design unsurprisingly

TABLE 2.3: DTPs after observing 3TTT in cohort 1. Cohort 2 is recommended to receive dose-level 2. The dose recommended for cohort 3 depends on the outcomes in cohort 2, as depicted by this table.

| Cohort 2 Outcomes | Dose for Cohort 3 |
|---|---|
| 2NNN | 3 |
| 2NNE | 1 |
| 2NNT | Stop trial |
| 2NNB | 1 |
| 2NEE | 1 |
| 2NET | 1 |
| 2NEB | 1 |
| 2NTT | Stop trial |
| 2NTB | 1 |
| 2NBB | 1 |
| 2EEE | 1 |
| 2EET | 1 |
| 2EEB | 1 |
| 2ETT | 1 |
| 2ETB | 1 |
| 2EBB | 1 |
| 2TTT | Stop trial |
| 2TTB | 1 |
| 2TBB | 1 |
| 2BBB | 1 |

de-escalates to dose-level 2. If a mix of only N and T events is observed in cohort 2, or three T events, the design recommends no dose for cohort 3, choosing to stop

the trial due to excess toxicity and lack of efficacy. If 2NNN is observed, the design chooses to re-escalate. In contrast to toxicity-only dose-finding methods, observing *no change* is a bad outcome, and the lack of response motivates escalation. In every other path, the design chooses to de-escalate to dose-level 1 in cohort 3. The EffTox design is prevented from skipping doses. The effect of the level of toxicity observed in cohort 1 endures to warrant further de-escalation in the majority of paths to seek a tolerable dose. After 3TTT 2NNN, the design has simultaneously observed excess toxicity and a complete absence of response. It is torn between the completing claims of seeking efficacy and avoiding toxicity. In this particular path, the design chooses to re-escalate to dose-level 3. After observing 3TTT, even before commencing cohort 2, we know from Table 2.3 that if the trial makes proceeds to cohort 3, it will probably be at dose-level 1.

Table 2.3 shows DTPs for a single future cohort but that need not be a constraint. We use DTPs in Matchpoint to analyse every feasible outcome of the next few cohorts. DTPs can be calculated for several subsequent cohorts, or even an entire trial. However, the number of possible paths grows geometrically with the number of cohorts being considered. Each evaluable patient will experience exactly one of E, T, N or B, independent of the other patients. With cohorts of three, the number of distinct outcomes for a single cohort is 20, as shown in Table 2.3, hence the number of feasible DTPs for the next two and three cohorts are $20^2$ and $20^3$ respectively. Thus, the limitation of what can be depicted on printed pages tends to limit our DTP analysis to no more than the next two cohorts of three patients.

Our frequent use of DTPs contributed to the efficient conduct of the Matchpoint trial. We were particularly interested to learn the outcomes that would have to manifest to change dose or stop the trial and DTPs allowed us to make timely preparations. Furthermore, the method allowed us to convey to investigators and the monitoring committee the behaviour of our design in many hypothetical scenarios, emulating the familiar transparency of rule-based designs like 3+3.

#### 2.3.2.3 Outcome Ambiguity

Patients in the blastic transformation phase of CML under study in the Matchpoint trial are particularly sick. The FLAG-IDA regimen is toxic and the addition of ponatinib only increases the potential for toxic adverse events. Periodic dose-selection meetings are a feature of dose-finding studies, where early safety and efficacy outcomes are reviewed and a new dose for the next patient or cohort is selected. Sometimes, because of the frail nature of the patients, efficacy assessments might be temporarily delayed. This *outcome ambiguity* presents a challenge for dose-selection because the decision seemingly requires that full patient outcomes be available. However, we have already seen that this is not necessarily the case. From Table 2.3, we know that at least one E or B event in cohort 2 is enough to know with certainty that the trial will proceed to cohort 3 using dose-level 1. If one of the patients experiences E or B in cohort 2, the dose-decision is independent of the other two patients so it does not matter, purely from a dose-decision perspective, if some of the outcome information is temporarily missing for the other patients in cohort 2.

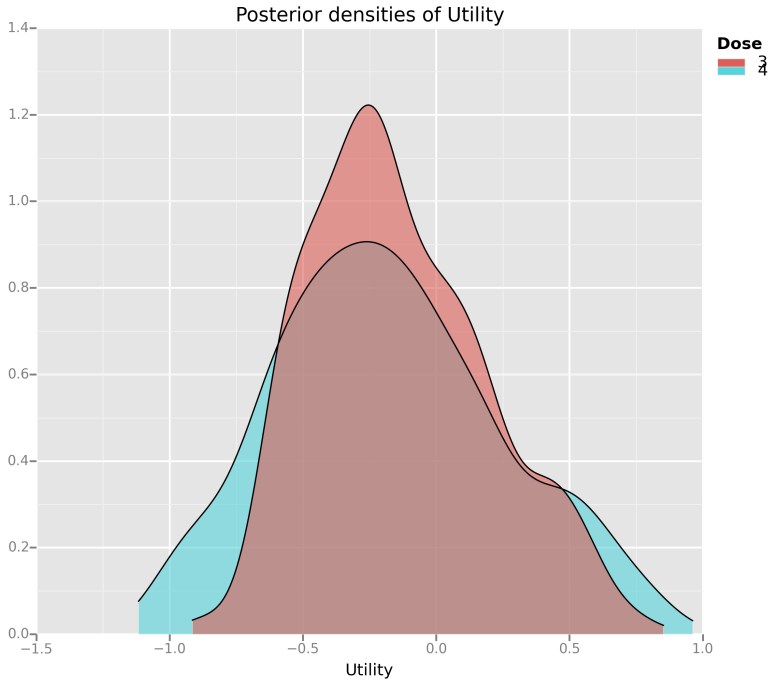
Naturally, this phenomenon does not always occur and there are many occasions when every patient's outcomes will be required promptly to know the course of action in the subsequent cohort. Furthermore, it is important that outcomes for cohort 2 are finalised before trying to establish doses for cohorts after cohort 3, for example, because all patient outcomes affect the dosing decision in model-based dose-finding designs. The described method merely offers short-term respite in *some* occasions if a small number of data-points are *temporarily* missing.

#### 2.3.2.4 Posterior Utility

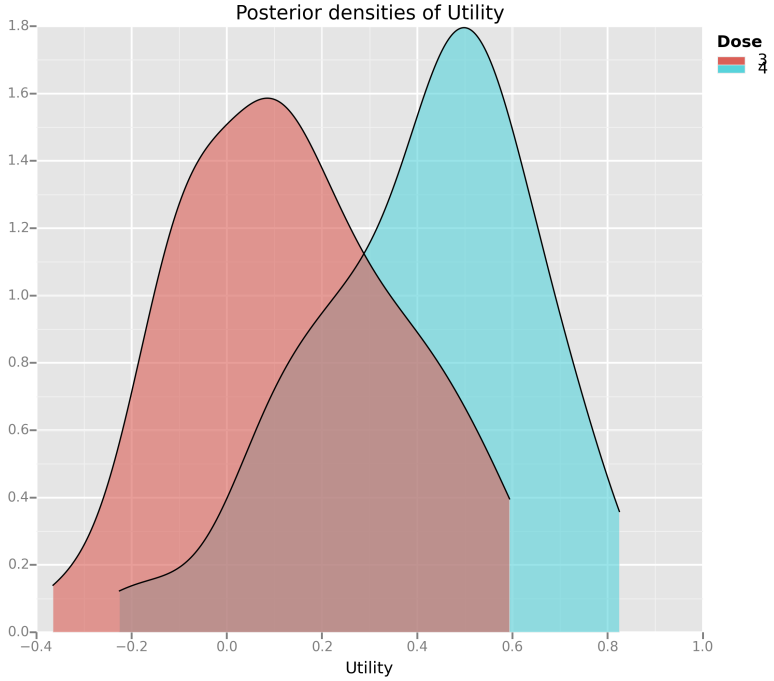
Thall & Cook work with utility as a function of the mean posterior efficacy and toxicity probabilities of the doses. In contrast, we consider here the posterior distribution of utility scores. For example, the posterior mean utility of dose $x$ is

$$\hat{u}(\pi_E(x,\boldsymbol{\theta}),\pi_T(x,\boldsymbol{\theta}))|\mathcal{D}) = \frac{\int u(\pi_E(x,\boldsymbol{\theta}),\pi_T(x,\boldsymbol{\theta}))\mathcal{L}(\boldsymbol{\theta}|\mathcal{D})f(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int \mathcal{L}(\boldsymbol{\theta}|\mathcal{D})f(\boldsymbol{\theta})d\boldsymbol{\theta}} \tag{2.8}$$

where $\mathcal{L}$ is the likelihood function given in EffTox[92] and $f(\boldsymbol{\theta})$ is the parameter prior distribution.




(A) After 3 patients with outcomes 3NTE




(B) After 15 patients with outcomes 2NNN 3ENN 4EBE 3TEE 4NEE

FIGURE 2.2: Posterior densities of utility at dose-levels 3 and 4. After only three patients, the densities largely occupy the same space and dose ambivalence is likely. However, after 15 patients, they are quite distinct and dose ambivalence is much less likely.

The posterior utility density curves in Figure 2.2 demonstrates the difficulty a utility-maximising design like EffTox faces when few patient outcomes have been observed and two doses have very similar utility scores. Figure 2.2a depicts the posterior beliefs on the utilities of dose-levels 3 and 4 after observing 3NTE in cohort 1. For clarity in illustration, dose-levels 1 and 2 are not shown. Figure 2.2a shows that the distribution for dose 4 has slightly greater variability but that the two utilities have approximately equal mode. When the posterior utilities for the two doses are so similar, it is difficult for the design to reliably choose between them and dose ambivalence (expanded below) is the likely result. Figure 2.2b shows similar curves after 15 patients with outcomes 2NNN 3ENN 4EBE 3TEE 4NEE. In contrast, the posterior utilities are now quite distinct and a consistent dose decision is almost guaranteed. The EffTox implementation in the `trialr`[16] package offers functionality to plot posterior utility densities.

#### 2.3.2.5 Dose Ambivalence

The EffTox probability model has six parameters for which prior distributions are specified. After patient outcomes are observed, posterior estimates of efficacy and toxicity come from evaluating a six-dimensional integral, one dimension for each parameter. Such integrals are approximated numerically rather than solved analytically and this leads to estimation error. Typically, early phase clinical trials do not use a large number of patients so the amount of information in the trial will usually be quite low, i.e. the number of patients divided by the number of parameters being estimated will be lower than a typical phase II or III trial. The combined effect of these two sources of variability is that the EffTox outputs are subject to quite a lot of uncertainty, especially in early cohorts. An unwelcome consequence of this uncertainty is that the model can make different dose recommendations based on the same patient outcomes. This is obviously undesirable in a clinical trial where a categorical course of action is sought. KB identified this phenomenon and named it *dose ambivalence*.

Consider, for example, our Matchpoint parameterisation and the posterior utilities depicted in Figure 2.2a. After observing outcomes 3NTE in the first cohort, the design sometimes recommends dose-level 3 for the next cohort, and sometimes

dose-level 4. This ambiguity manifests because the two doses are both admissible, have similar utility scores, and the Bayesian update integral is imperfectly calculated. This happens when using the MD Anderson implementation of the EffTox software[45] that uses the spherical radial method of Monahan and Genz[65] to estimate the posterior integrals, and our own implementation that uses Monte Carlo Markov Chain methods.

It is possible to calculate the integral more precisely by increasing the number of posterior samples or integration points but this risks missing an important message. If the dose-recommendation is not consistent when calculated to a reasonable numerical precision, the design is telling us that it is difficult to pick between the doses. It could be that several doses have similar utility scores, as we have seen. Alternatively, it could be that a dose is very close to the boundary for inclusion in the admissible set. For instance, purely by chance, repeated invocations of the imperfect statistical analysis may alternatively include or exclude a dose from the admissible set. In the former, the dose is available for selection. In the latter, it is not. In these circumstances, the dose recommended is likely to vary. When the design is ambivalent about a dose, rather than rely on one invocation of the dose update decision, it is more appropriate in our opinion to calculate the dose recommendation many times (say, 1,000) using reasonable precision and analysing the distribution of the selections. This presents the uncertainty of the dose recommended.

We give a further example in Table 2.4. Our design with $p_E = 0.05$ was ambivalent on the preferred action after observing 3TTT in the first cohort. This uncovered a flaw in our design. In that particular instance, the reticence to take the logical action and de-escalate motivated us to re-parameterise the model to $p_E = 0.03$, as described in the following section.

#### 2.3.2.6 Changing $p_E$ to avoid premature stopping

We commenced the trial with both $p_E$ and $p_T$ set to 0.05 so that the design only had to be at least 5% sure that a dose was efficacious and safe to include it in the admissible set.

Table 2.4 summarises the decision in two dose transition scenarios for the first

TABLE 2.4: Outcomes of 1,000 replicates for two possible dose-paths in the first cohort receiving dose-level 3, calculated using EffTox parameterisations with different values for $p_E$. Pr(Stop) is the probability that the design recommends stopping and Pr($i$) is the probability that the design recommends selecting dose-level $i$ for the next cohort. Decision is the dose level most frequently recommended for the next cohort in the replicates.

| | Using $p_E = 0.05$ | | | | | | Using $p_E = 0.03$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Path | Pr(Stop) | Pr(1) | Pr(2) | Pr(3) | Pr(4) | Decision | Pr(Stop) | Pr(1) | Pr(2) | Pr(3) | Pr(4) | Decision |
| 3NEE | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 4 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 4 |
| 3TTT | 0.45 | 0.00 | 0.18 | 0.37 | 0.00 | Stop | 0.09 | 0.00 | 0.85 | 0.06 | 0.00 | 2 |

Matchpoint cohort using two slightly different parameterisations of the EffTox design. The columns on the left show the behaviour of a design with $p_E = 0.05$; the right a design with $p_E = 0.03$. Each row summarises 1,000 replicates of the dose-transition decision. *Pr(Stop)* is the probability that the design recommends stopping and *Pr(i)* is the probability that the design recommends selecting dose-level $i$ for the next cohort. *Decision* is the dose level most frequently recommended for the next cohort in the replicates.

TABLE 2.5: EffTox posterior beliefs after observing 3TTT in cohort 1. The values for $p_E$ and $p_T$ determine the admissible doses.

| | Dose 1 | Dose 2 | Dose 3 | Dose 4 |
|---|---|---|---|---|
| Utility | -0.489 | -0.534 | -0.777 | -0.817 |
| Pr($\pi_E > \underline{\pi}_E$) | 0.079 | 0.037 | 0.060 | 0.200 |
| Pr($\pi_T < \overline{\pi}_T$) | 0.919 | 0.758 | 0.051 | 0.005 |
| Admissible under $p_E = 0.05$, $p_T = 0.05$ | 1 | 0 | 1 | 0 |
| Admissible under $p_E = 0.03$, $p_T = 0.05$ | 1 | 1 | 1 | 0 |

After observing 3NEE, both designs recommend escalating to dose-level 4 in all iterations. This is sensible and consistent behaviour. In contrast, after observing 3TTT, the designs take different courses. The design using $p_E = 0.05$ seems unsure of its preferred behaviour. In approximately half of the iterations, it recommends stopping and in the other half, it proposes to select dose-level 2 or 3. This is another manifestation of the ambivalence previously described. The design using $p_E = 0.03$ is rather more consistent because it recommends de-escalating to dose-level 2 in the great majority of replicates.

Output from the official EffTox software in Table 2.5 reveals the cause. Having observed three toxicities at dose-level 3, all doses are believed to be unattractive, hence the negative utility scores. The most attractive dose is actually dose-level 1,

so the design would like to de-escalate. However, the design cannot go straight to dose-level 1. The restriction to not skip untried doses requires that dose-level 2 is tested first. The software does not allow this feature to be turned off. However, with $p_E = 0.05$, dose-level 2 is actually inadmissible so the design cannot de-escalate.

The problem is potentially exacerbated by the fact that $\Pr(\pi_T < \overline{\pi}_T)$ is very close to the value $p_T = 0.05$ for dose-level 3. If this probability is estimated to be slightly less than 0.05, as is possible with just 3 data-points and a six-dimensional Bayesian integral solved numerically, then dose-level 3 becomes inadmissible also. Under these circumstances, with dose-level 4 inadmissible too on account of excess toxicity, the design cannot recommend a dose so it advocates stopping. This accounts for the relatively large probability of stopping under $p_E = 0.05$ in Table 2.4.

By reducing $p_E$ to 0.03, we made it much more reliable that the design would de-escalate after 3TTT rather than stop. Observing three toxicities in the first three patients is clearly a grave situation. However, we should be mindful of the play of chance and the extent of our knowledge on the event rates. The lower bound of the 95% confidence interval for a binomial proportion having observed three events in three trials is 29.2% using the exact method, and 43.9% using the Wilson method. This implies that the true toxicity rate could plausibly be much lower than the 100% rate observed with 3TTT. Also, we have no direct knowledge of the toxicity rates at the other dose-levels, only the information extrapolated by our model from the toxicities observed at dose-level 3. To stop the trial before even trying the lower dose-levels seems hasty and wasteful. We had not noted the restriction of no-skipping, and its implications, until we examined the DTPs more closely. We chose $p_E = 0.03$ over $p_E = 0.05$ so that our design would be more willing to de-escalate at early trial stages when toxicities are observed. We advise fellow trialists to study DTPs routinely, especially in early cohorts, to spot undesirable behaviour.

In the scenario in question, it is sensible to ask why we are tweaking a parameter that pertains to efficacy when toxicity is the problem. The succinct answer is that it was the posterior prediction of efficacy that rendered the doses inadmissible after invocation of (2.6). In addition to observing the *presence* of toxicity, the design simultaneously observed the *absence* of response. Decreasing $p_E$ was one solution but was likely not the only one. The same ends could have perhaps been achieved by

increasing the effective sample size to give more weight to our priors, thus overriding the low efficacy and high toxicity rates observed. This seemed a less satisfactory alteration to us.

It is important that investigators are aware of the circumstances under which their design would recommend stopping because the official EffTox software (v4.0.12) will not allow further patients to be added once the stop point has been reached. If investigators are relying on this software, they could find themselves constrained by a hitherto unknown feature of their design. It is better to address these issues in set-up rather than when the trial is in progress. Nevertheless, KB developed an open source implementation of EffTox[16]for research purposes only that will continue to accept new patient outcomes even after the stop point has been reached.

This section has described a flaw in our EffTox parameterisation that was not initially evident to us, that could have led to undesirable behaviour and disruption in our trial, that we discovered through novel analysis of dose transition pathways and repeated simulation, and resolved via a minor adjustment to the parameterisation. Efficient trial conduct is maintained when flaws like this are uncovered before they become critical, or avoided altogether.

### 2.3.3 Operating Characteristics

Once a complete set of parameters has been proposed, we learn how the design performs by simulation.

Blastic transformation phase CML is relatively rare. It was felt that 30 patients was the feasible limit to recruit in a reasonable time frame. The trialists, including CY and JK but not KB, selected 30 patients as the target sample size. This was chosen to maximise the expected probability of identifying the optimal dose. The trialists also chose to use cohorts of three, thus re-evaluating the recommended dose after every third patient. Evaluating dose after every patient would allow maximum capacity for adaptation. However, it would also require the maximum number of interim analyses. Under the procedures of the trials unit, each analysis would be associated with a monitoring visit, data cleaning and a meeting of the independent data monitoring committee. It was hoped that 10 dose decisions (i.e. cohorts of three) would provide enough opportunity for dose adaptation whilst not coercing an undesirable

administrative burden on those persons running and monitoring the trial. We investigated by simulation study the operating performance that could be expected with 30 patients treated cohorts of three.

The EffTox software provides the ability to simulate the outcome of thousands of trials using the chosen design and some assumed true efficacy and toxicity curves. In practice, of course, the true curves are unknown. We choose a variety of scenarios for simulation that will provide pertinent information on the behaviour of the design in real usage.

Trialists should assess performance in a set of clinically relevant scenarios. One of the scenarios should closely resemble the investigators' prior beliefs, as this represents the anticipated outcome. We would expect the probability of correctly selecting the best dose to be high in the scenario that matches the investigators' priors. The setting for any clinical trial is that we are unsure of the truth so the range of scenarios in which our design performs well should reflect our ignorance. We considered how the design would perform if adverse circumstances prevailed. To these ends, we advocate analysing performance when (i) no doses are tolerable, and (ii) no doses are efficacious. In these scenarios, the desirable behaviour is to stop. As the clinical scenario dictates, we might also advocate analysing scenarios where the true efficacy and toxicity curves are not monotonically increasing.

In Table 2.6 we analyse in six indicative scenarios the performance of our chosen EffTox model in Matchpoint, labelled ESS=1.3. We have also given the performance of two other models with priors on $\boldsymbol{\theta}$ recalibrated using the EffTox software to give ESS set to 0.5 and 1.5, being the recommended lower and upper limits on ESS advised by Thall & Cook[92]. These convey the feasible range of performance, holding all other parameters constant. In every other regard, the three models are exactly the same. 10,000 replicates were simulated for each model in each scenario. The Monte Carlo standard error for probabilities estimated by simulation with this number of replicates is up to $\sqrt{0.5 \times 0.5/10000} = 0.5\%$ so that selection probabilities that differ by more than 1% probably differ by more than can be regarded merely as simulation error.

We naturally seek a design that selects the optimal dose most reliably. The optimal selection is shown in bold in Table 2.6. However, the design must reliably

TABLE 2.6: Final dose-selection and stopping probabilities of EffTox designs with 30 patients in cohorts of 3 and ESS=0.5, 1.3 and 1.5. In rows pertaining to design performance, the optimal decision is in bold and the admissible decisions are underlined. The EffTox software gives selection probabilities to the nearest whole percent.

| Scenario | | Dose 1 | Dose 2 | Dose 3 | Dose 4 | Stop |
|---|---|---|---|---|---|---|
| 1 | Pr(Eff) | 0.20 | 0.30 | 0.50 | 0.60 | |
| | Pr(Tox) | 0.03 | 0.05 | 0.10 | 0.30 | |
| | Utility | -0.33 | -0.17 | 0.16 | **0.22** | |
| | ESS=0.5 | 0.01 | 0.01 | 034 | **0.63** | 0.01 |
| | ESS=1.3 | <0.01 | <0.01 | 0.22 | **0.76** | <0.01 |
| | ESS=1.5 | <0.01 | <0.01 | 0.22 | **0.77** | <0.01 |
| 2 | Pr(Eff) | 0.40 | 0.60 | 0.75 | 0.79 | |
| | Pr(Tox) | 0.10 | 0.25 | 0.55 | 0.60 | |
| | Utility | -0.01 | **0.25** | 0.12 | 0.08 | |
| | ESS=0.5 | 0.06 | **0.59** | 0.32 | <0.01 | 0.03 |
| | ESS=1.3 | 0.03 | **0.60** | 0.35 | <0.01 | 0.01 |
| | ESS=1.5 | 0.03 | **0.57** | 0.39 | <0.01 | 0.01 |
| 3 | Pr(Eff) | 0.25 | 0.40 | 0.60 | 0.60 | |
| | Pr(Tox) | 0.10 | 0.20 | 0.38 | 0.42 | |
| | Utility | -0.26 | 0.04 | **0.15** | 0.12 | |
| | ESS=0.5 | 0.03 | 0.10 | **0.70** | 0.13 | 0.04 |
| | ESS=1.3 | 0.01 | 0.10 | **0.73** | 0.13 | 0.02 |
| | ESS=1.5 | 0.01 | 0.09 | **0.73** | 0.15 | 0.02 |
| 4 | Pr(Eff) | 0.50 | 0.60 | 0.70 | 0.80 | |
| | Pr(Tox) | 0.20 | 0.20 | 0.20 | 0.20 | |
| | Utility | 0.12 | 0.28 | 0.43 | **0.57** | |
| | ESS=0.5 | 0.02 | 0.03 | 0.61 | **0.34** | <0.01 |
| | ESS=1.3 | <0.01 | 0.02 | 0.47 | **0.50** | <0.01 |
| | ESS=1.5 | <0.01 | 0.01 | 0.47 | **0.51** | <0.01 |
| 5 | Pr(Eff) | 0.05 | 0.08 | 0.20 | 0.25 | |
| | Pr(Tox) | 0.05 | 0.08 | 0.12 | 0.14 | |
| | Utility | -0.58 | -0.54 | -0.34 | -0.26 | |
| | ESS=0.5 | 0.06 | 0.03 | 0.01 | 0.37 | **0.53** |
| | ESS=1.3 | 0.06 | 0.07 | 0.02 | 0.34 | **0.51** |
| | ESS=1.5 | 0.07 | 0.08 | 0.02 | 0.36 | **0.48** |
| 6 | Pr(Eff) | 0.05 | 0.08 | 0.12 | 0.25 | |
| | Pr(Tox) | 0.60 | 0.65 | 0.70 | 0.80 | |
| | Utility | -0.78 | -0.78 | -0.76 | -0.67 | |
| | ESS=0.5 | 0.09 | 0.01 | 0.01 | 0.01 | **0.88** |
| | ESS=1.3 | 0.06 | 0.01 | 0.01 | 0.01 | **0.91** |
| | ESS=1.5 | 0.04 | 0.01 | 0.01 | 0.01 | **0.93** |

stop when no dose satisfies (2.6) and (2.7). When stopping is the correct decision, the stopping probability is shown in bold. Sometimes, there will be many *admissible*

doses that satisfy (2.6) and (2.7), irrespective the fact that one generally dominates all others by our utility metric. Admissible decisions are underlined in Table 2.6. When stopping is the correct decision, stopping is the only admissible decision.

Scenarios 1 and 4 show the benefit of a modestly more informative prior. Through the addition of prior information approximately equivalent to one patient (i.e. increasing the effective sample size of the prior from 0.5 to 1.3 or 1.5), the probability that the design selects the optimal dose is increased by up to 17%. Investigators will naturally ponder the existence of the opposite effect, i.e. an increased propensity to do the wrong thing when the prevailing scenario disagrees with the prior. Scenario 3 shows that this is not necessarily the case. The designs with more informative priors actually perform slightly better, despite a shape of efficacy curve that disagrees with the prior.

TABLE 2.7: Mean probabilities of performing the optimal decision in the scenarios presented in Table 2.6.

| Design variant | Mean Pr(Optimal decision) |
|---|---|
| ESS=0.5 | 0.612 |
| ESS=1.3 | 0.668 |
| ESS=1.5 | 0.650 |

Table 2.7 shows the mean probability that each design variant identifies the optimal decision, given the six scenarios in Table 2.6. We see that our design is the superior of the three presented. The variant with ESS=0.5 has inferior performance, mostly for the reasons discussed. The variant with ESS=1.5 is only modestly inferior but provides no reason to be preferred to our design.

We investigated by simulation the larger sample size $n = 60$. The extra patients greatly improve performance in some scenarios. In scenario 2, the probability of selecting dose 2 increases by 20% to 80%. Similarly, the chances of correctly stopping early in scenario 5 increase by 27% to 78%. In many clinical scenarios, recruiting a higher number of patients is warranted in a phase I-II trial because of the associated improvement in performance and the presence of efficacy assessment that may abrogate a further traditional phase II trial. Phase I-II trials are an opportunity to optimise the delivery of a new agent. In the Matchpoint scenario, unfortunately, higher recruitment was simply not feasible because of the rarity of BP-CML.

We also investigated the impact of using $p_E = p_T = 0.1$. The chances of stopping in scenario 5 are improved by 30%. As expected, the reciprocal effect is that the design stops slightly more frequently in scenarios like 3 where an optimal dose exists.

As well as their propensity to make the correct decision, we also discriminate designs on how they allocate doses to patients. A design that always makes the correct decision but treats every patient at an over dose would not desirable, or indeed ethical. Table 2.8 gives the mean number of patients allocated to each dose in the scenarios presented in Table 2.6.

Of the three designs presented, our chosen design uses the fewest patients in scenarios 5 and 6, where the correct decision is to stop the trial. On the four remaining scenarios, our chosen design allocates the most patients on average to the optimal dose in two scenarios. We see this as further reason to prefer the design with ESS=1.3.

In summary, we have shown by simulation that our selected EffTox parameterisation performs well in six scenarios. We have demonstrated that it stops reliably in situations where all doses are too toxic or inefficacious. We have also shown it to perform well in a scenario that broadly matches our prior, and in scenarios that depart from our prior. Lastly, we have demonstrated that our chosen parameterisation with effective sample size set to 1.3 is superior to alternatives with ESS set to 0.5 and 1.5, in terms of probability of making the correct decision, and in the allocation of patients to favourable doses.

## 2.4 Discussion

Finalising an EffTox design is generally an iterative process. The inferences from analysing dose transition pathways and simulations will naturally lead to re-parameterisation and further testing.

It is our general preference to first hone the dose transitions. For the reasons described, we pay particular attention to the earliest circumstances under which the trial would stop. The investigators should agree that these circumstances are dire enough to warrant closing the trial. We also look for any sign that the design

TABLE 2.8: Mean numbers of patients allocated to each dose, and in total, in the six scenarios and three EffTox variants presented in Table 2.6. Sum $\neq$ 30 when the trial stops early. Patients allocated to the optimal dose are given in bold and admissible doses underlined.

| Scenario | | Dose 1 | Dose 2 | Dose 3 | Dose 4 | Sum |
|---|---|---|---|---|---|---|
| 1 | Pr(Eff) | 0.20 | 0.30 | 0.50 | 0.60 | |
| | Pr(Tox) | 0.03 | 0.05 | 0.10 | 0.30 | |
| | Utility | -0.33 | -0.17 | 0.16 | **0.22** | |
| | ESS=0.5 | 0.7 | 0.6 | 12.1 | **16.4** | 29.8 |
| | ESS=1.3 | 0.2 | 0.2 | 9.8 | **19.6** | 29.8 |
| | ESS=1.5 | 0.1 | 0.1 | 9.5 | **20.1** | 29.8 |
| 2 | Pr(Eff) | 0.40 | 0.60 | 0.75 | 0.79 | |
| | Pr(Tox) | 0.10 | 0.25 | 0.55 | 0.60 | |
| | Utility | -0.01 | **0.25** | 0.12 | 0.08 | |
| | ESS=0.5 | 1.5 | **11.5** | 16.0 | 0.4 | 29.4 |
| | ESS=1.3 | 0.8 | **11.6** | 16.9 | 0.6 | 29.9 |
| | ESS=1.5 | 0.7 | **10.3** | 18.2 | 0.7 | 29.9 |
| 3 | Pr(Eff) | 0.25 | 0.40 | 0.60 | 0.60 | |
| | Pr(Tox) | 0.10 | 0.20 | 0.38 | 0.42 | |
| | Utility | -0.26 | 0.04 | **0.15** | 0.12 | |
| | ESS=0.5 | 1.1 | 2.8 | **21.8** | 3.7 | 29.4 |
| | ESS=1.3 | 0.5 | 2.5 | **22.2** | 4.4 | 29.6 |
| | ESS=1.5 | 0.4 | 2.0 | **22.1** | 5.2 | 29.7 |
| 4 | Pr(Eff) | 0.50 | 0.60 | 0.70 | 0.80 | |
| | Pr(Tox) | 0.20 | 0.20 | 0.20 | 0.20 | |
| | Utility | 0.12 | 0.28 | 0.43 | **0.57** | |
| | ESS=0.5 | 0.5 | 1.0 | 19.3 | **9.2** | 30.0 |
| | ESS=1.3 | 0.1 | 0.7 | 15.9 | **13.3** | 30.0 |
| | ESS=1.5 | 0.1 | 0.3 | 15.9 | **13.7** | 30.0 |
| 5 | Pr(Eff) | 0.05 | 0.08 | 0.20 | 0.25 | |
| | Pr(Tox) | 0.05 | 0.08 | 0.12 | 0.14 | |
| | Utility | -0.58 | -0.54 | -0.34 | -0.26 | |
| | ESS=0.5 | 2.4 | 2.4 | 4.7 | 14.1 | 23.6 |
| | ESS=1.3 | 1.5 | 1.9 | 4.7 | 15.3 | 23.4 |
| | ESS=1.5 | 1.4 | 1.7 | 4.5 | 16.0 | 23.6 |
| 6 | Pr(Eff) | 0.05 | 0.08 | 0.12 | 0.25 | |
| | Pr(Tox) | 0.60 | 0.65 | 0.70 | 0.80 | |
| | Utility | -0.78 | -0.78 | -0.76 | -0.67 | |
| | ESS=0.5 | 2.6 | 3.0 | 4.0 | 0.9 | 10.5 |
| | ESS=1.3 | 1.1 | 2.8 | 5.2 | 0.8 | 9.9 |
| | ESS=1.5 | 1.0 | 2.8 | 5.3 | 0.9 | 10.0 |

seems reluctant to select a dose. This could suggest unsuitable priors or inappropriate parameter choices. It should be stressed, however, that EffTox exists to guide our sequential selection of doses based on patient outcomes. The trialists should

not stipulate every conceivable dose path and select parameters that replicate their choices. This approach would preclude the use of a model at all. Rather, in our opinion, the parameters should be selected for generally acceptable behaviour, with particular consideration given to the extremes.

Once an acceptable parameterisation has been proposed, the performance of the design should be assessed by simulation under a broad range of scenarios. The design should stop sufficiently early and reliably when all doses are too toxic. In scenarios where optimal and/or acceptable doses exist, the design should select those with acceptable probability. Refinements to the parameterisation here will likely require the trial designer to consider how they affect the behaviour of the design in dose transition, and thus the circularity of the challenge is illustrated.

We have considered even steeper contours, as stressed by Thall *et al.*[94] and Yuan *et al.*[111]. They did not lead to superior performance in the particular scenarios we have chosen. This is likely due to the fact that our contours are steep for efficacy probabilities as high as 70%, which we consider to be the clinically plausible scenario in BP-CML. However, the point remains that trade-off contours should be steep to motivate the design to accept higher probabilities of efficacy for acceptably higher probabilities of toxicity.

EffTox is a powerful yet underused statistical design for seamless phase I/II dose-finding clinical trials. Model-based designs are becoming more important as trialists and funders move away from so-called "up-and-down" designs[59] like 3+3. This trend will be further driven as investigators research treatments for which the *maximum-tolerated* dose is unlikely to coincide with the most effective dose. We have described our approach to overcoming some of the obstacles we faced in implementing EffTox in Matchpoint.

We were able to choose EffTox because our co-primary outcomes efficacy and toxicity were assessed over a similar time-frame. EffTox and other dose-finding designs with co-primary outcomes would not have been suitable if one had required a longer assessment period.

A key reason for selecting EffTox was the readily available, free software provided by the MD Anderson Cancer Center for performing dose calculations and simulations of trial operating characteristics. With the many time pressures that come

with working in an academic clinical trials unit, it was a tremendous advantage to have reliable software with which to design and run this trial. One of the drawbacks of using compiled software was our inability to alter or add certain behaviours. For instance, we might have suppressed the no-skipping in de-escalation rule, had that been possible. The desire to routinely calculate dose-transition pathways led us to developing an open-source implementations of EffTox in `clintrials`[15] and `trialr`[16].

## 2.5 Conclusion

Joint phase I/II clinical trials will likely become more common in coming years as we investigate non-cytotoxic treatments and streamline the drug approval process. EffTox is an important trial design because it addresses both of these goals. The Matchpoint trial will yield data on the efficacy and toxicity of the optimum dose of ponatinib to be given with FLAG-IDA chemotherapy. It will allow the research community to decide whether there is sufficient promise to warrant a pivotal trial, effectively shortening the pathway to approval by removing the need for a separate phase II trial. This efficiency is important in a relatively rare disease like BP-CML.

However, EffTox presents its challenges. It requires parameterisation and preliminary calculation. Choices for parameters can potentially have undesirable consequences, and without care the efficiency gains can be eroded. The process of finalising an EffTox design is inherently iterative. We have described our experiences in the hope that it helps trialists implement this design successfully.

We have discussed the parameters we chose and how we selected them. We have stressed the need to look at the dose transition pathways, particularly in the early stages when few outcomes are observed, and at the circumstances that would lead to the trial's termination. We have highlighted the problem of dose ambivalence, illustrated graphically, and suggested a pragmatic solution. We have described the problem of outcome ambiguity, and how dose-transition pathways can mitigate the problem in the short term, allowing the dose-finding trial to proceed whilst clinical evaluation is ongoing. Finally, we have advised on the simulation scenarios that

should be considered. We hope this paper will help other investigators implement this important dose-finding clinical trial design.

We used version 4.0.12 of the official EffTox software, available from the MD Anderson Centre website at
https://biostatistics.mdanderson.org/SoftwareDownload/.

# Chapter 3

# Development of an Adaptive Dose-Finding Design

**Background:** Wages and Tait introduced a method for phase I/II dose-finding as an alternative to established designs like EffTox, described in the previous chapter. Their method uses a simpler probability model than EffTox, and adaptive randomisation to preferably allocate patients to doses estimated to be tolerable and effective. Adaptive randomisation potentially brings an operational cost to clinical trials units if the randomisation probabilities must be independently validated before use. With frequent analyses, as in a dose-finding trial, this administrative burden threatens to impact efficient trial progress.

**Notable methods in this chapter:** We introduce a hybrid trial design that borrows Wages and Tait's probability model, whilst abrogating the need for randomisation by implementing the utility contours used in EffTox. We compare nine variants of the three designs in a simulation study, seven of which do not use randomisation, comparing them using a novel measure borrowed from quantitative finance.

**The implications on efficiency:** A version of Wages & Tait's design that does not use randomisation showed superior statistical performance. This design achieved our operational efficiency objective without compromising statistical efficiency. Our hybrid design achieved the same operational objective but offered slightly inferior statistical performance and greater heterogeneity, whilst allocating marginally fewer patients at attractive doses.

## 3.1 Symbols used

Table 3.1 contains a list of the symbols used in this chapter.

## 3.2 Introduction

This chapter builds on the last by considering alternative designs for phase I/II dose-finding trials. The clinical motivations for conjoining phases I and II were discussed in Chapter 2. For decades, chemotherapy has been one of the cornerstones of cancer treatment and the dose-finding objective has typically been to identify the MTD. However, modern therapies increasingly challenge the validity of the cytotoxic assumption. An immunotherapy example is developed in this thesis at length in Chapter 5. A very brief review of phase I/II methodologies was presented in Chapter 2 and we elaborate on that now.

An early design in this area came from O'Quigley *et al.*[68] conducting dose-finding studies in HIV retroviral therapies, extending the Continual Reassessment Method (CRM)[69] by using relatively simple functions to model the probabilities of efficacy and toxicity at discrete doses. They use a trinomial outcome with categories Toxicity, No Response and Response, and consider situations where efficacy is not, in general, monotonically increasing in dose.

Braun[12] introduced a bivariate generalisation of CRM (bCRM) with competing outcomes for toxicity and response. In their setting, the 'dose' being studied is the amount of time after a stem-cell transplant (SCT) to wait before tapering immunosuppressive (IS) therapy and beginning donor leukocyte infusions (DLI) intended to incite a graft-versus-leukaemia (GVL) effect. If IS therapy is tapered too soon, there is a risk of acute graft-versus-host disease (aGVHD), a common and potentially fatal complication with SCTs. In contrast, if DLI are given too late, there is a risk of disease progression. Treating occurrence of aGVHD as the toxicity event and absence of disease progression as the efficacy event, it is clear that a trade-off between the two events must be sought to provide the best outcome for patients. They model the probabilities of efficacy and toxicity using independent logit models. The two

| Symbol | Definition |
|---|---|
| $d_i$ for $i = 1, ..., n$ | a dose, e.g. 10 might represent 10mg |
| $g(\beta)$ | prior distribution for $\beta$ |
| $h(\theta)$ | prior distribution for $\theta$ |
| $n$ | the number of doses under investigation |
| $i$ | index of doses |
| $j$ | index of patients |
| $k$ | index of efficacy skeletons |
| $k^*$ | index of efficacy skeleton with greatest posterior probability |
| $p_i$ | prior probability of toxicity at dose $i$ |
| $p_E$, $p_T$ | certainties required that efficacy / toxicity rate is acceptable |
| $q_{ik}$ | prior probability of efficacy at dose $i$ under skeleton $k$ |
| $u(\pi_E, \pi_T)$ | utility function |
| $w(k, \mathcal{D}_j)$ | posterior probability of efficacy skeleton $k$ in WT and WATU |
| $x_j$ | dose allocated to patient $j$ |
| $y_j$ | toxicity variable for patient $j$, 1 meaning tox; 0 meaning no tox |
| $z_j$ | efficacy variable for patient $j$, 1 meaning eff; 0 meaning no eff |
| $\mathscr{D}$ | the set of doses, $d_i$ |
| $\mathcal{D}_j$ | trial data up to and including patient $j$ |
| ET | EffTox |
| $F(d, \beta)$ | toxicity link function in WT and WATU |
| $G_k(d, \theta)$ | efficacy link function using efficacy skeleton $k$ in WT and WATU |
| $X_j$ | random variable for dose allocated to patient $j$ |
| $Y_j$ | random variable for toxicity presence in patient $j$ |
| $Z_j$ | random variable for efficacy presence in patient $j$ |
| WT | Wages and Tait |
| WATU | Wages and Tait with Utility |
| $\alpha_E$ | significance for deficient efficacy at optimal dose in WT |
| $\alpha_T$ | significance for excess toxicity at lowest dose in WT |
| $\beta$ | toxicity curve parameter in WT and WATU |
| $\hat{\beta}_j$ | posterior estimate of $\beta$ using data for $j$ patients |
| $\beta_{E,1}$ | slope term for dose in EffTox efficacy logit |
| $\beta_{E,2}$ | slope term for dose-squared in EffTox efficacy logit |
| $\beta_T$ | slope term in EffTox toxicity logit |
| $\delta_i$ for $i = 1, ..., n$ | transformed dose, used in EffTox |
| $\boldsymbol{\delta}$ | vector of transformed doses in EffTox, $\delta_i$ |
| $\boldsymbol{\theta}$ | vector of parameters in EffTox model |
| $\theta$ | efficacy curve parameter in WT and WATU |
| $\hat{\theta}_{jk}$ | posterior estimate of $\theta$ using efficacy skeleton $k$ and data for $j$ patients |
| $\lambda$ | curvature parameter in efficacy-toxicity contours |
| $\mu_T$ | intercept term in EffTox toxicity logit |
| $\pi_E$, $\pi_T$ | probability functions for efficacy and toxicity |
| $\hat{\pi}_E$, $\hat{\pi}_T$ | posterior probabilities of efficacy / toxicity in WT and WATU |
| $\pi^*_{1,E}$, $\pi^*_{1,T}$, etc | probabilities of efficacy / toxicity at notable points |
| $\underline{\pi}_E$ | lower threshold for efficacy rate |
| $\overline{\pi}_T$ | upper threshold for toxicity rate |
| $\tau(k)$ | weight assigned to efficacy skeleton $k$ |
| $\psi$ | association parameter in EffTox model |
| $\omega$ | random draw from a normal distribution |

TABLE 3.1: Symbols used in this chapter, in alphabetical order.

outcomes are then combined into a joint likelihood model with an association parameter to handle the tendency for events to co-occur.

Thall and Cook[92] introduced *EffTox* in 2004. We covered this design in great detail in the previous chapter so no further elaboration is warranted here.

Yin *et al.*[110] introduce another dose-finding method for co-primary efficacy and toxicity. Unlike EffTox, they do not specify any functional form for the dose-response curve. Instead, they use a novel class of priors to impose a monotonic constraint on the probabilities of toxicity at increasing doses. The efficacy curve is free from constraint and they choose amongst doses by trading-off the efficacy-to-toxicity odds ratios at the investigated doses.

Wang & Day[99] introduce another Bayesian approach for co-primary efficacy and toxicity dose-finding. They model the patient-level thresholds at which events occur using bivariate log-normal distributions and, like EffTox, select amongst doses for individuals using utility functions. This allows patients and clinicians to potentially influence the delivered dose by reflecting the extent to which toxicity will be risked in pursuit of efficacy.

Zhang *et al.*[112]. introduced a model-based trivariate CRM (TriCRM) design. It uses a continuation ratio logit models to partition the bivariate efficacy and toxicity outcome space into three exclusive and exhaustive outcomes: response with no toxicity; no response and no toxicity; and toxicity (irrespective response). Wang & Day[99] are critical of this approach, pointing out that toxicity with response is clearly preferable to toxicity with no response and that this should be reflected in the model.

Recently, Shimamura *et al.*[82] introduced a two-stage approach for combinations of two agents in phase I/II trials. Their first stage is for identifying the 'most admissible toxicity zone' by varying doses of treatments in the combination. In their second stage, they use adaptive randomisation to collect outcomes under the admissible combinations.

Further, Ananthakrishnan *et al.*[3] extended the modified Toxicity Probability Interval (mTPI) design of Ji & Wang[49] and Toxicity Equivalence Range design (TEQR) of Blanchard & Longmate[9] to include a binary efficacy outcome. They apply isotonic regression to the observed toxicity and efficacy outcomes considering a range dose-response curves to determine the optimal dose.

One of the focuses of this chapter is the method introduced by Wages and Tait[98]

(that we will refer to as WT) for seamless phase I/II dose-finding trials. Their design uses a latent CRM to continually model the probabilities of toxicity. To model the rates of efficacy, they choose amongst pre-specified efficacy skeletons, simple sparse efficacy curves that reflect the plausible shapes of the general dose-efficacy curve. They choose the skeleton that best fits the observed efficacy curve using Bayes factors. The method of choosing amongst these skeletons is described in Section 3.3. The vertical location of the dose-efficacy curve is allowed to vary using a single parameter, much like the common empiric CRM model[24]. Overall, their method requires two parameters, a prior estimate of the dose-toxicity curve and the specification of the efficacy skeletons. They provide guidance in their paper for specifying the recommended $(2n - 1)$ skeletons in a trial that investigates $n$ doses.

WT conducts a dose-finding study in two stages. In the first stage, patients are adaptively randomised amongst the doses believed to be tolerable, with probabilities proportional to the estimated chances of efficacy. Adaptive randomisation creates potential friction within a trials unit where standard operating procedures likely require that randomisation probabilities must be independently validated before use. Conducted infrequently, this is not a great hindrance. However, when a model is updated after each small cohort, as is common in a dose-finding trial, the validation requirement stands to become arduous. If the model is updated after every patient, as Wages & Tait themselves investigated, the validation burden could become prohibitive.

Our pre-occupation in this thesis is an examination of methods that promote trial efficiency. WT is attractive because its probability model is relatively simple, building upon methods used in CRM that are now well understood and widely implemented in software[16, 24, 86]. However, its operational efficiency could be diminished by the use of adaptive randomisation. In this chapter, we propose a fusion of WT and Thall & Cook's *EffTox* (ET) that removes the randomisation requirement. In Section 3.3, we describe the statistical aspects of WT and our motivation for combining this design with ET. We introduce the hybrid method that we call *Wages And Tait with Utility* (WATU) in Section 3.4. In Section 3.5, we compare the performance of the three methods in a simulation study. Finally in Section 3.6, we close with a discussion.

## 3.3 Wages & Tait

The EffTox model was detailed in Chapter 2. In this section we introduce the WT design, after establishing some notation.

We are studying seamless phase I/II dose-finding clinical trials with joint binary efficacy and toxicity outcomes. We denote $\mathscr{D} = \{d_1, ..., d_n\}$ to be the set of $n$ doses under investigation. Each patient will be treated at exactly one dose level and will yield binary outcomes for efficacy and toxicity. Let $X_j$ be the random variable representing the dose allocated to patient $j$, taking values $x_j \in \mathscr{D}$. Let $Y_j$ and $Z_j$ be the random variables representing binary toxicity and efficacy events respectively for patient $j$, taking values $y_j, z_j \in \{0, 1\}$, where 1 denotes the event occurred and 0 that it did not.

Wages and Tait's[98] method estimates the toxicity curve by delegating to the univariate Bayesian variant of the Continual Reassessment Method (CRM)[69]. Let $p$ represent the trialists' prior beliefs on the rate of toxicity at each dose. In a monotonic dose-toxicity scenario, we have $0 < p_1 < ... < p_n < 1$. We will denote the single parameter in the toxicity model $\beta$ and assume it has prior distribution $g(\beta)$. Wages & Tait choose the same $\beta \sim N(0, 1.34)$ prior used by O'Quigley & Shen[70]. Using trial data for the first $j$ patients $\mathcal{D}_j = \{(x_1, y_1, z_1), ..., (x_j, y_j, z_j)\}$ and toxicity probability function $F(d, \beta)$, the likelihood for $\beta$ is

$$\mathcal{L}(\beta|\mathcal{D}_j) = \prod_{l=1}^{j} \{F(x_l, \beta)\}^{y_l} \{1 - F(x_l, \beta)\}^{(1-y_l)} \ , \tag{3.1}$$

the posterior density for $\beta$ is

$$P(\beta|\mathcal{D}_j) = \frac{\mathcal{L}(\beta|\mathcal{D}_j)g(\beta)}{\int_{-\infty}^{\infty} \mathcal{L}(\beta|\mathcal{D}_j)g(\beta)d\beta} \tag{3.2}$$

and the posterior mean[1] is

$$\hat{\beta}_j = \int_{-\infty}^{\infty} \beta P(\beta|\mathcal{D}_j)d\beta \tag{3.3}$$

[1]Note that in contrast to more common usage, the hat symbol here denotes the posterior mean and not the maximum likelihood estimate.

For instance, using the empiric link function $F(d_i, \beta) = p_i^{\exp(\beta)}$, the posterior estimate of the dose-toxicity curve is

$$\hat{\pi}_T(d_i) = F(d_i, \hat{\beta}_j) = p_i^{\exp(\hat{\beta}_j)} \tag{3.4}$$

The authors use the values of $\hat{\pi}_T(d_i)$ to define an *acceptable* set of doses, an object analogous to the admissible set in ET. Henceforth, we will use the term *admissible* throughout for consistency. A dose is admissible in WT if the estimated rate of toxicity is less than the maximum acceptable rate. More formally, the admissible set after evaluation of patient $j$ is:

$$\mathcal{A}_j = \{d_i : \hat{\pi}_T(d_i) < \overline{\pi}_T; i = 1, ..., n\} \tag{3.5}$$

where $\overline{\pi}_T$ is the maximum acceptable toxicity rate.

To model the efficacy probabilities, they use order restricted inference and Bayesian model selection by specifying a set of working models, or *skeletons*, that describe the plausible shapes of the dose-efficacy curve and iteratively choosing the skeleton that best fits the observed efficacy data. The authors describe a general method for identifying $2n - 1$ skeletons when the dose-efficacy curve might be monotonically increasing, unimodal (i.e. initially increasing and then decreasing) or plateaued. Naturally, if the situation demands it, more or fewer skeletons could be considered. A monotonically increasing efficacy curve would have $\pi_E(d_1) < ... < \pi_E(d_n)$. In contrast, an efficacy curve that plateaus at the penultimate dose would have $\pi_E(d_1) < ... < \pi_E(d_{n-1}) = \pi_E(d_n)$.

Let $K$ denote the number of efficacy skeletons under consideration in a trial. For skeleton $k$, let the probabilities of efficacy at the $n$ doses be $(q_{1k}, ..., q_{nk})$ for $k = 1, ..., K$. Under skeleton $k$, the authors model $\pi_E(d_i) = Pr(Z_j = 1|d_i) \approx G_k(d_i, \theta) = q_{ik}^{\exp(\theta)}$. Once again, the empiric link function is used.

The parameter $\theta$ controls the vertical location of the efficacy curve. Let $\theta$ have prior distribution $h(\theta)$. After $j$ patients have been treated and assessed on the study,

the likelihood under model $k$ is

$$\mathcal{L}_k(\theta|\mathcal{D}_j) = \prod_{l=1}^{j} \{G_k(x_l, \theta)\}^{z_l} \{1 - G_k(x_l, \theta)\}^{(1-z_l)} \ , \tag{3.6}$$

the posterior density for $\theta$ is

$$P_k(\theta|\mathcal{D}_j) = \frac{\mathcal{L}_k(\theta|\mathcal{D}_j)h(\theta)}{\int_{-\infty}^{\infty} \mathcal{L}_k(\theta|\mathcal{D}_j)h(\theta)d\theta} \tag{3.7}$$

and the posterior mean under skeleton $k$ is

$$\hat{\theta}_{jk} = \int_{-\infty}^{\infty} \theta P_k(\theta|\mathcal{D}_j)d\theta \tag{3.8}$$

At each dose-update decision, the authors select the skeleton with the highest posterior probability. Their method allows investigators to express their prior beliefs on which skeletons are more likely via a weight function $\tau(k)$, scaled so that $\sum_{k=1}^{K} \tau(k) = 1$. If the investigators believe the skeletons to be equally likely, they set $\tau(k) = \frac{1}{K}$ for $k = 1, ..., K$. Then, the posterior model probabilities are

$$w(k|\mathcal{D}_j) = \frac{\tau(k) \int_{-\infty}^{\infty} L_k(\theta|\mathcal{D}_j)h(\theta)d\theta}{\sum_{k=1}^{K} \tau(k) \int_{-\infty}^{\infty} L_k(\theta|\mathcal{D}_j)h(\theta)d\theta} \tag{3.9}$$

and the skeleton chosen to model the dose-efficacy curve at the dose-update decision, $k^*$, is that with greatest posterior model probability, i.e.

$$k^* = \arg\max_k w(k, \mathcal{D}_j) \tag{3.10}$$

for $k = 1, ..., K$. As the authors note, "the more the data support model $k$, the greater its posterior probability will be". Having selected the best skeleton, the posterior probabilities of efficacy are estimated as

$$\hat{\pi}_E(d_i) = G_{k^*}(d_i, \hat{\theta}_{jk^*}) \tag{3.11}$$

Wages & Tait propose two stages to their design and these differ in the way the next dose is selected for the next patient or cohort.

During the first stage, the so-called *randomisation stage*, the next dose is randomly

selected from the admissible set with probability proportional to $\hat{\pi}_E(d_i)$ so that doses with the high estimated efficacy are preferably selected. The adaptive randomisation probability of dose $i$ in $\mathcal{A}_j$ is

$$R_i = \frac{\hat{\pi}_E(d_i)}{\sum_{d_i \in \mathcal{A}_j} \hat{\pi}_E(d_i)} \tag{3.12}$$

During the second stage, the admissible dose with maximal $\hat{\pi}_E(d_i)$ is selected. This is called the *maximisation stage*. At each stage, dose transition may be restricted to avoid skipping untried doses in escalation and/or de-escalation. If at either stage the admissible set is empty, the trial proposes no dose and the trial stops.

The trialists can set the size of the stages, or even choose to only use one particular stage by setting the size of the other to 0, an option we explore below. Wages & Tait investigate different mixes and show that a 50:50 split works quite well.

As with ET, there are stopping rules in WT. The *safety* stopping rule is applied at each dose update decision. The exact binomial quantile is calculated for the observed rate of toxicity at the lowest dose, $d_1$ using significance $\alpha_T$. If the lower bound exceeds the maximum acceptable toxicity rate, $\overline{\pi}_T$, the treatment is understood to be too toxic at all doses and the trial is stopped. If the lowest dose has not been given, the lower bound for the confidence interval is effectively 0 and the safety rule does not fire. Wages & Tait advocate using $\alpha_T = 0.05$.

There is also a *futility* stopping rule to stop investigators pursuing a treatment unduly when the observed efficacy rate is too low. This rule is applied at dose update decisions only in the maximisation stage. It is not invoked during the randomisation stage. This rule uses a similar method, calculating the exact binomial confidence interval for the observed rate of efficacy at the proposed dose using significance $\alpha_E$. If the upper bound is less than the minimum acceptable efficacy rate, $\underline{\pi}_E$, the treatment is understood to be inefficacious at all doses and the trial is stopped. Wages & Tait use $\alpha_E = 0.05$.

### 3.3.1 The Rationale for Combining WT and ET

EffTox is a powerful yet complicated trial design. It requires the specification of many parameters and a degree of familiarity to ensure that those parameters work

harmoniously to yield an effective trial design. We demonstrated this in Chapter 2.

The Bayesian update integral is six-dimensional, a not-inconsiderable challenge that must be resolved at each dose decision. MD Anderson provides software for conducting and simulating EffTox[45] trials but not source-code. Naturally, some desirable features may be missing. We wrote an implementation of EffTox to produce dose-transition pathways for future cohorts. This was instrumental in our investigation into dose-ambivalence. To our knowledge, we have published the only open-source EffTox implementations, in the Python package *clintrials*[15] and the R package *trialr*[16].

Wages & Tait's design is simpler and has several benefits. It delegates to a well-known design in CRM to perform a key trial role. The method for modelling efficacy is intuitive and tractable. At dose decisions, the method requires only that one-dimensional integrals be solved. It is not particularly onerous to implement the design in a new programming language.

In our opinion, WT has potential drawbacks too. The design randomises between doses in the first stage. This requires that the trialists be willing to give any dose selected. Wages & Tait recommend that the design be constrained to avoid skipping untested doses in escalation. In a similar vein, the design could select a low dose that the trialists believe to be sub-therapeutic. To combat this, it would be relatively simple to prevent skipping in de-escalation too. However, whilst it is the job of clinical trials to provide objective evidence that may confirm *or refute* trialists' beliefs, we feel that in some scenarios, investigators might prefer a method that changes doses deterministically rather than randomly. It is possible in WT to skip the randomisation mechanism by moving straight to the maximisation stage. We investigate this option in the simulation study.

Our main motivation to alter WT however is to remove the adaptive randomisation component. Under the present operating procedures of our trials unit, randomisation methods must be validated by the trial statistician. If the randomisation mechanism changes, it has to be validated again. In the case of WT, this would at least necessitate that each set of randomisation probabilities is independently replicated. We believe this requirement would be present under the standard procedures

of many trials units. Small but frequent hurdles such as this will reduce the operational efficiency of the method and could make it less attractive.

We feel that WT has many advantages that should be available to trialists in all dose-finding scenarios that require the joint consideration of efficacy and toxicity, particularly ease of use. In this spirit, we seek to adapt Wages & Tait's design to transition dose using the ET principle of utility maximisation, rather than adaptive randomisation.

All seamless phase I/II methods can be considered efficient because they perform two trial functions at once. The efficiency of a dose-finding method could be inferred from its performance. A design that identifies the optimum dose to a given threshold reliability using fewer patients can be considered more efficient. To these ends, we investigate the performance of WT, ET and their hybrid with a simulation study.

## 3.4 WATU - A Hybrid Model

We introduce in this section a hybrid of Wages & Tait's design and EffTox, named Wages And Tait with Utility (WATU)[2] The following design was conceived by Christina Yap (CY) and Kristian Brock (KB) on a train journey, returning home from an early phase trials workshop.

Our starting point is to mimic the probability models for both efficacy and toxicity in Wages & Tait. We advocate using a Bayesian CRM model to continually estimate the dose-toxicity curve. This requires the trialists' prior beliefs on the rate of toxicity at each dose, $p_i$ for $i = 1, ..., n$ in a study of $n$ doses. Using a one-parameter CRM model with empiric link function, the posterior probabilities of toxicity are given by (3.4). Other parameterisations and link functions are possible[24].

We also use Wages & Tait's method of choosing amongst efficacy skeletons to estimate the dose-efficacy curve. The posterior probabilities of efficacy are given by (3.11).

[2]According to Wiktionary, watu is a Quechua word, meaning clothesline or spell. It seems appropriate to this author that a dose-finding trial design would be named at the intersection of support structure and incantation.

At this juncture, we depart from Wages & Tait. For calculating the admissible set of doses, we mirror EffTox in equations (2.6) and (2.7) by admitting doses that are threshold efficacious and tolerable according to their posterior distributions. A dose $d$ is admissible in WATU after observing trial data $\mathcal{D}_j$ if

$$\Pr\{\pi_E(d) > \underline{\pi}_E | \mathcal{D}_j\} > p_E \tag{3.13}$$

and

$$\Pr\{\pi_T(d) < \overline{\pi}_T | \mathcal{D}_j\} > p_T \tag{3.14}$$

Equation (3.3) gives the posterior mean for $\beta$. The posterior variance is

$$\text{var}(\beta_j) = \int_{-\infty}^{\infty} \beta^2 P(\beta|\mathcal{D}_j) d\beta - \hat{\beta}_j^2 \tag{3.15}$$

and the posterior distribution for $\beta$ after $j$ patients have been observed will be approximately $N(\hat{\beta}_j, \text{var}(\beta_j))$. It is simpler to sample from this specification of the posterior distribution than from (3.2). We can estimate $\Pr\{\pi_T(d) < \overline{\pi}_T | \mathcal{D}_j\}$ and thus resolve (3.14), for example, by randomly sampling $\omega_1, ..., \omega_M$ from the normal distribution $N(\hat{\beta}_j, \text{var}(\beta_j))$ for suitably large $M$, and calculating

$$\Pr\{\pi_T(d_i) < \overline{\pi}_T | \mathcal{D}_j\} \approx \frac{1}{M} \sum_{m=1}^{M} \mathbb{I}(F(d_i, \omega_m) < \overline{\pi}_T) \tag{3.16}$$

for each dose. Here $\mathbb{I}(A)$ is the indicator function taking value 1 when event $A$ is true, else 0. We infer that dose $d_i$ is tolerable if the quantity on the right-hand side of (3.16) is greater than $p_T$. We can perform a similar calculation to resolve (3.13).

As with ET, the investigators must provide values for $\underline{\pi}_E$, $\overline{\pi}_T$, $p_E$ and $p_T$. If no dose is admissible, the trial stops. The admissible set is recalculated at the end of each cohort.

A further departure from WT comes in the way the next dose is selected. We explained our preference for a similar design to WT that does not randomly assign doses. Having reappropriated in WATU the probability models from WT to estimate

the posterior probabilities of efficacy and toxicity, we can mimic Thall & Cook's approach in EffTox of selecting the admissible dose with the greatest utility. The equation for calculating dose utility is given in (2.5). The process of identifying an $L^p$ norm is the same as that described in Section 2.2 and Thall et al[93]. CY had the idea of porting EffTox's utility contours to abrogate the need for randomisation in WT.

A dose $d_i$ is admissible if it satisfies (3.13) and (3.14). Additionally, the trialists may choose to refine the admissible set by preventing the design from skipping untried doses in escalation and/or de-escalation. Given our preference for deterministic dose-transition and the nature of dose-finding under the competing moderators efficacy *and* toxicity, we believe it is preferable to avoid skipping untried doses in both escalation and de-escalation. At the dose update decision, WATU will select the most attractive dose from the admissible set. What determines 'most attractive' changes in WATU according to the stage of the trial.

### 3.4.1 Trial Conduct - Stage One

The aim of the first stage is to gather information on which doses are safe. We envisage starting at a low dose, albeit not necessarily the lowest dose, and escalating though the doses in a sequential fashion, all the while mindful to not allocate a dose that is believed to be too toxic. To achieve these ends, we propose that dose be guided by the underlying CRM model in the first stage. CRM is calibrated with a target rate of toxicity and iteratively recommends the dose that it believes to have associated toxicity rate closest to the target. The design forecasts the rate of toxicity at each dose by combining observed trial outcomes with prior information. The toxicity target will be set as the clinical situation dictates but will naturally be less than $\overline{\pi}_T$. For instance, in the Matchpoint setting (introduced below), we believe a priori that efficacy will initially increase with dose (and toxicity) but then begin to level off. Hence, we use a toxicity target slightly below $\overline{\pi}_T$ when we use WATU in Matchpoint scenarios.

The following pseudo-code describes the dose-update decision in the first trial stage:

```
1. Select each dose in turn, ordered from estimated closest to
```

```
    furthest from toxicity target:
2.      If the dose is admissible:
3.          If the dose does not violate a no-skipping rule:
4.              Select this as the recommended dose
5.              Exit
6. If no dose is selected, stop the trial.
```

### 3.4.2 Trial Conduct - Stage Two

Stage two seeks to allocate to the optimal dose by trading the probability of toxicity against the probability of efficacy. Each dose is allocated an attractiveness score using the method of Thall & Cook[92], as described. Doses with high attractiveness scores will offer a relatively high rate of efficacy for the rate of toxicity that must be endured.

The following pseudo-code describes the dose-update logic in the second trial stage:

```
1. Select each dose in turn, from highest to lowest utility:
2.      If the dose is admissible:
3.          If the dose does not violate a no-skipping rule:
4.              Select this as the recommended dose
5.              Exit
6. If no dose is selected, stop the trial.
```

### 3.4.3 Sizes of Stage One & Two

Trialists can set the sizes of the stages as they see fit. Stage one allows the design to step through doses and collect efficacy and toxicity information when trial data is very limited and safety is paramount. Naturally, the size of stage one could be set to zero to suppress the CRM-only allocation stage. Stage one need not even be of fixed size. Trialists might prefer stage one to end when some pre-defined event happens, like the moment the first (or $n$th) efficacy event is observed, for instance. We investigate different sizes for stage one below.

## 3.5 Comparing the Designs by Simulation

In this section we compare the performance of ET, WT and WATU in simulations motivated by the Matchpoint scenario, described in Chapter 2. The trial uses an ET design although it could feasibly have used another joint phase I/II design like WT or WATU. Once again, we codify the doses under investigation as 7.5mg, 15mg, 30mg and 45mg each day. The trial and hence our simulations will recruit up to 30 patients in cohorts of three. We will start each iteration of each design at dose-level 3, as we did in the actual study. The doses under investigation and our prior beliefs on efficacy and toxicity are given in Table 2.2.

### 3.5.1 Designs Under Investigation

The main objective of the simulation study is to compare the general performance of the three competing designs, ET, WT and WATU in a real trial situation. However, it is important to consider that each of the designs may be configured in different ways to promote or inhibit certain behaviour. There is no uniquely *correct* parameterisation in any case. Trialists will naturally use all available levers to get the most desirable behaviour from their design. It is likely that a design could be configured is such a way as to replicate the desirable behaviour of another. For instance, WT provides a mechanism for suppressing adaptive randomisation by simply setting the size of the randomisation stage to 0. As such, a secondary objective of the simulation study is to compare the performance of variations of the designs.

Naturally, it is infeasible to analyse all design variants. Specifically, we are interested in how reliably the designs pick the optimum dose. Further to this, we will investigate how the designs perform in monotonic scenarios, where the dose-efficacy curve broadly concurs with our prior beliefs. Additionally, we will analyse performance in non-monotonic efficacy scenarios, where the prior beliefs must be overruled by the designs to accurately model the prevailing clinical scenario. To these ends, we propose to investigate two instances of EffTox (ET1, ET2), three instances of Wages & Tait (WT1, WT2, and WT3) and three instances of Wages And Tait with Utility (WATU1, WATU2 and WATU3), as described in the following sections.

### 3.5.2 Parameterising the Designs

Each of the designs under study requires parameters. These determine the behaviour of the methods and will naturally be selected by the trialists to reflect their expectations and objectives in the clinical trial. We discuss how our expectations on efficacy and toxicity inform the parameterisation below.

In a phase I/II dose-finding setting, trial designs may either select a dose for further study or recommend that the trial stop early with no dose being selected. A trial may be stopped for lack of efficacy or excess toxicity. Highly divergent propensities to stop hinder the comparison of designs. For instance, the performance of a design that never stops will look artificially superior in scenarios where the correct decision is to not stop. The indecision on whether to stop is a burden that must be carried by all designs. ET, WT and WATU use different methods to infer dose admissibility. After discussing efficacy and toxicity parameterisation below, we describe a systematic method for parameterising designs so that they stop with approximately equal probability in a given benchmark scenario. Neutralising this important source of variation aids comparison of the designs in general scenarios.

Our prior estimates for the rates of efficacy and toxicity at the four doses are given in Table 2.2 in the previous chapter. The CRM models in WT and WATU both use these prior toxicity probabilities. Each uses the empiric link function and a $N(0, 1.34)$ prior for $\beta$.

ET requires normal priors on the six elements of $\boldsymbol{\theta}$. In ET1, we use an effective sample size of 1.3 to match the choice we made in the real Matchpoint trial. The EffTox software produces the normal priors shown in Table 3.2. Under parameterisation ET2 we have inflated the standard deviation hyperparameter for $\beta_{E,2}$ fivefold to 1.0, whilst leaving the other elements unchanged, to facilitate non-linearity and turning points in the dose-efficacy curve.

To give validity to the choice of 1.0 as a suitable value for the standard deviation of $\beta_{E,2}$, consider the following example. In Matchpoint, the unimodal efficacy curve (0.17, 0.44, 0.50, 0.40) is fitted by the parameter values $\mu_E = 0$, $\beta_{E,1} = 0.5$ and $\beta_{E,2} = -1.2$. Under the ET1 priors, this value of $\beta_{E,2}$ is 6 prior standard deviations away from the prior mean. Under the ET2 priors, however, this value is only 1.2

TABLE 3.2: Normal priors for the elements of the EffTox parameter vector $\boldsymbol{\theta}$ in the Matchpoint setting. Priors in ET1 are calculated using the MD Anderson EffTox software with ESS = 1.3.

| | ET1 | | ET2 | |
|---|---|---|---|---|
| Parameter | Mean | St Dev | Mean | St Dev |
| $\mu_T$ | -5.4317 | 2.7643 | -5.4317 | 2.7643 |
| $\beta_T$ | 3.1761 | 2.7703 | 3.1761 | 2.7703 |
| $\mu_E$ | -0.8442 | 1.9786 | 0.8442 | 1.9786 |
| $\beta_{E,1}$ | 1.9857 | 1.9820 | 1.9857 | 1.9820 |
| $\beta_{E,2}$ | 0 | 0.2 | 0 | 1 |
| $\psi$ | 0 | 1 | 0 | 1 |

prior standard deviations away. This change increases the probability that ET2 will fit unimodal efficacy curves. Unfortunately, the EffTox software does not reveal the effect of inflating the prior standard deviation of $\beta_{E,2}$ on ESS.

To create the efficacy skeletons in WT and WATU, we permuted the prior efficacy rates, as demonstrated in Table 3.3. It does not matter that skeleton 7, for example, plateaus at a high efficacy probability that we do not necessarily expect to manifest because the $\theta$ parameter adjusts the average height of the curve to best fit the observed efficacy rates. Rather, the ordering of the nodes is important. The efficacy models in WT and WATU use a $N(0, 1.34)$ prior for $\theta$.

TABLE 3.3: Efficacy skeletons for WT and WATU in the Matchpoint trial setting. We consider monotonic, unimodal and plateau skeletons.

| $k$ | $q_{1k}$ | $q_{2k}$ | $q_{3k}$ | $q_{4k}$ |
|---|---|---|---|---|
| 1 | 0.6 | 0.5 | 0.3 | 0.2 |
| 2 | 0.5 | 0.6 | 0.5 | 0.3 |
| 3 | 0.3 | 0.5 | 0.6 | 0.5 |
| 4 | 0.2 | 0.3 | 0.5 | 0.6 |
| 5 | 0.3 | 0.5 | 0.6 | 0.6 |
| 6 | 0.5 | 0.6 | 0.6 | 0.6 |
| 7 | 0.6 | 0.6 | 0.6 | 0.6 |

In WT1, we allocate 15 patients to the randomisation stage, and 15 to the maximisation stage, to give the design equal opportunity to explore the doses and identify the optimal dose. Furthermore, in WT1 we equally weight the efficacy skeletons with $\tau(k) = 1/7$ for $k = 1, .., 7$ so that each efficacy scenario is equally likely, a priori. We parameterise WT2 the same as WT1, with the exception that efficacy skeleton 4 is weighted twice as likely as the other skeletons, to reflect the anticipation of a

monotonic dose-efficacy curve, i.e.

$$\tau(k) = \begin{cases} 1/4, & \text{for } k = 4 \\ 1/8, & \text{for } k = 1, 2, 3, 5, 6, 7 \end{cases} \tag{3.17}$$

We parameterise WT3 the same as WT2 with the exception that the randomisation stage size is set to 0, i.e. WT3 is inclined towards the monotonically-increasing efficacy skeleton and proceeds immediately with the maximisation stage of dose-allocation without randomisation, keeping maximum sample size fixed at 30.

In WATU1, we allocate 15 patients to the first stage and 15 patients to the second stage, as described in Sections 3.4.1 and 3.4.2. In this configuration, the design has equal opportunity to escalate through the doses safely and identify the optimal dose. We also uniformly weight the efficacy skeletons, as with WT1.

In WATU2, we bias the model to prefer the monotonically increasing efficacy skeleton using (3.17), as in WT2. In every other regard, WATU2 matches WATU1. Lastly, we parameterise WATU3 the same as WATU2, with the exception that the size of stage one is set to zero so that the design proceeds immediately with identifying the optimal dose, without the initial CRM-driven safety stage,

We require utility measures for the ET and WATU designs. Following the example in Thall et al[92, 93], we selected neutral utility points $(\pi^*_{1,E}, 0) = (0.4, 0)$, $(1, \pi^*_{2,T}) = (1, 0.7)$ and $(\pi^*_{3,E}, \pi^*_{3,T}) = (0.5, 0.4)$, yielding a family of utility curves with $p = 2.07$. We calculate the utility of doses using (2.5). These match the utility curves we used in Matchpoint.

The design parameterisations we consider in our simulation study are summarised in Table 3.4.

### 3.5.3 Simulation method

To compare the general performance of the designs in Table 3.4, we conducted a simulation study using a wide range of scenarios. Patient outcomes were randomly sampled according to assumed true efficacy and toxicity probabilities, with the events assumed to be independent. In each scenario, we simulate 10,000 trial outcomes using up to 30 patients in each iteration with doses being given in cohorts of three. For

TABLE 3.4: Parameterisations of all designs under study. Only the parameters that vary are shown. Common parameters are given in the text. The efficacy skeleton that is upweighted in WT2, WT3, WATU2 and WATU3 is the one representing a monotonic dose-efficacy curve.

| Design | Parameterisation |
|---|---|
| ET1 | $\beta_{E,2} \sim N(0, 0.2)$ |
| | $p_E = 0.15$ |
| | $p_T = 0.16$ |
| ET2 | $\beta_{E,2} \sim N(0, 1.0)$ |
| | $p_E = 0.16$ |
| | $p_T = 0.16$ |
| WT1 | First stage (randomisation) size = 15 |
| | efficacy skeleton weights = (1,1,1,1,1,1,1) |
| | $\alpha_E = 0.3$ |
| | $\alpha_T = 0.3$ |
| WT2 | First stage (randomisation) size = 15 |
| | efficacy skeleton weights = (1,1,1,2,1,1,1) |
| | $\alpha_E = 0.3$ |
| | $\alpha_T = 0.3$ |
| WT3 | First stage (randomisation) size = 0 |
| | efficacy skeleton weights = (1,1,1,2,1,1,1) |
| | $\alpha_E = 0.32$ |
| | $\alpha_T = 0.32$ |
| WATU1 | First stage (CRM) size = 15 |
| | efficacy skeleton weights = (1,1,1,1,1,1,1) |
| | $p_E = 0.2$ |
| | $p_T = 0.2$ |
| WATU2 | First stage (CRM) size = 15 |
| | efficacy skeleton weights = (1,1,1,2,1,1,1) |
| | $p_E = 0.2$ |
| | $p_T = 0.2$ |
| WATU3 | First stage (CRM) size = 0 |
| | efficacy skeleton weights = (1,1,1,2,1,1,1) |
| | $p_E = 0.22$ |
| | $p_T = 0.22$ |

each design in each iteration, dose decisions were made at the end of each simulated cohort of three and the next cohort treated at the recommended dose, or the trial stopped, as the design advised. ET, WT and WATU were constrained from skipping doses in escalation and de-escalation, as described for each design above. All simulated trials started at dose-level 3, as Matchpoint did.

### 3.5.4 Parameters for Dose Admissibility

We have discussed the concept of dose admissibility and how this is handled in the different designs. An inadmissible dose is not considered for allocation to a cohort of patients. The sets of admissible and inadmissible doses will change as the trial progresses and outcomes are collected. Designs will stop the trial when no dose is admissible.

ET, WT and WATU use different methods to infer admissibility and this will affect performance because a decision to stop is simultaneously a decision to recommend no dose. For instance, a design that stops one third of the time and recommends the correct optimal dose two thirds of the time will look inferior to another design that never stops, selects the optimal dose 70% of the time and a sub-optimal dose 30% of the time. However, if the stopping probability is calibrated in the first design, it is likely that superior performance will be attained. Although this illustrative example is contrived, our early attempts at conducting simulation studies to compare the designs using the original authors' recommended stopping parameters yielded such disparate stopping probabilities that performance was essentially non-comparable. Calibration was necessary.

The aim of this section is to describe a systematic method of parameterising the admissibility components of the designs so that each stops with similar probability in a particular baseline scenario. The choice of baseline scenario will be important and it is expository to consider the process of arriving at a recommendation to stop.

ET, WT and WATU use different statistical methods to model the dose-efficacy and dose-toxicity curves, as described in Sections 2.2, 3.3 and 3.4. As with every statistical model, each is subject to estimation error and this is pertinent to the decision to stop. For instance, when a design recommends stopping, it could have correctly identified a scenario where all doses truly are too toxic. Alternatively, it could have

misclassified a scenario where at least some doses are tolerable. In this instance, the misclassification will stem from imperfect estimation of the dose-toxicity curve. Generally in phase I/II trials, estimates of the efficacy and toxicity curves will have bearing on the decision to stop. When calibrating designs to stop with similar probability, we remove this source of uncertainty using the following simple method.

In the baseline stopping scenario, we assume the toxicity and efficacy curves are flat by setting the probability of efficacy and toxicity to be the same at every dose. This removes the variation stemming from their imperfect estimation of the event curves because any dose selected yields the same event probabilities. The particular dose selected does not affect the decision to stop. Rather, the act of choosing any dose forgoes the opportunity to stop. The decision to stop is focussed as much as possible on the mechanism explicitly introduced to govern stopping and not on imprecise estimation of the dose-toxicity or dose-efficacy curves. We tweak the stopping parameters in the designs until each stops with a pre-determined probability in the baseline scenario.

In Matchpoint, we seek a dose with at least 45% efficacy and at most 40% toxicity thus we set $\underline{\pi}_E = 0.45$ and $\overline{\pi}_T = 0.4$ in each design. These values are chosen for their clinical relevance. WATU requires a toxicity target in the first stage. We set this to be 0.35, slightly below the toxicity limit. WT does not use a toxicity target in its CRM component.

In our baseline stopping scenario we set the true efficacy to 45% throughout so that each dose is borderline sufficiently efficacious. In contrast, we set toxicity to 50% at each dose. At this 10% margin over the maximum allowable toxicity rate, we require that each design stop with 60% probability. Achieving an exact stopping probability is not a realistic goal. At a practical level, we sought a parameterisation that yielded a stopping probability between 60% and 62% in each design. Where this was not possible, the parameterisation giving stopping probability closest to 60% was retained. Using the metric described in the previous section, the utility of the point $(\pi_E, \pi_T) = (0.45, 0.5)$ is -0.15. In summary, in our four dose trial setting, the baseline stopping scenario has true efficacy curve $(0.45, 0.45, 0.45, 0.45)$ and true toxicity curve $(0.5, 0.5, 0.5, 0.5)$.

Having fixed the threshold values for the acceptable maximum toxicity and minimum efficacy, stopping is governed in ET by the parameters $p_E$ and $p_T$. We adjust these until the requisite stopping probability is achieved in the baseline scenario. Our preference is to have similar values for $p_E$ and $p_T$ because we have no particular motivation to prioritise one source of error. In ET1, a modest amount of trial-and-error lead to $p_E = 0.15$ and $p_T = 0.16$, yielding a design that stops 60%[3] of the time in 10,000 iterations. Likewise, the values $p_E = 0.16$ and $p_T = 0.16$ lead to stopping in 61% of iterations in ET2. We must stress at this juncture that the values we have derived are greater than: i) those used by the original authors; and ii) those used in the Matchpoint trial described in Chapter 2. Thall & Cook use in their example[92] the slightly lower values $p_E = p_T = 0.1$. Compared to their parameterisation and our Matchpoint design, our ET1 and ET2 will less readily find a dose admissible and thus will stop more often.

The stopping mechanism in WATU is similar to ET and requires values for the same two parameters. The pair $p_E = 0.2$ and $p_T = 0.2$ lead to a stopping probability of 61.3% in WATU1 using 10,000 iterations. Despite the similarity in the stopping rules between ET and WATU, the specific values for $p_E$ and $p_T$ are different, hinting at the role played by the methods of modelling the dose-event curves. In WATU2, the same pair $p_E = 0.2$ and $p_T = 0.2$ lead to stopping 60.2% of the time. In WATU3, $p_E = 0.22$ and $p_T = 0.22$ lead to stopping 60.9% of the time.

The probability of WT stopping can be modified via the $\alpha_E$ and $\alpha_T$ parameters in the safety and futility stopping rules. Greater significance values will lead to narrower confidence intervals and a greater chance of stopping. In WT1, $\alpha_E = 0.3$ and $\alpha_T = 0.3$ gives a stopping probability of 61.5% in 10,000 simulations of the baseline scenario. In WT2, the same pairing $\alpha_E = 0.3$ and $\alpha_T = 0.3$ stops 60.8% of the time. In WT3, the pair $\alpha_E = 0.32$ and $\alpha_T = 0.32$ stops 58.8% of the time. Although fractionally less than 60%, this pair was chosen over $\alpha_E = 0.33$ and $\alpha_T = 0.33$, which stopped 63.2% of the time.

Once again, we note that $\alpha_E = 0.3$ and $\alpha_T = 0.3$ in WT are far from the values proposed by Wages & Tait[98]. This reflects that we have asked the designs to be

[3]The EffTox software published by the MD Anderson Centre gives stopping and dose selection probabilities rounded the nearest whole percent

| Pr(Tox) | ET1 | ET2 | WT1 | WT2 | WT3 | WATU1 | WATU2 | WATU3 |
|---|---|---|---|---|---|---|---|---|
| 0.1 | 0.21 | 0.23 | 0.032 | 0.033 | 0.118 | 0.127 | 0.122 | 0.151 |
| 0.2 | 0.23 | 0.24 | 0.051 | 0.059 | 0.131 | 0.136 | 0.123 | 0.163 |
| 0.3 | 0.29 | 0.31 | 0.131 | 0.141 | 0.183 | 0.181 | 0.171 | 0.196 |
| 0.4 | 0.42 | 0.44 | 0.314 | 0.314 | 0.323 | 0.328 | 0.317 | 0.327 |
| **0.5** | **0.60** | **0.61** | **0.615** | **0.608** | **0.588** | **0.613** | **0.602** | **0.609** |
| 0.6 | 0.79 | 0.79 | 0.885 | 0.876 | 0.859 | 0.871 | 0.873 | 0.858 |
| 0.7 | 0.91 | 0.91 | 0.987 | 0.986 | 0.980 | 0.982 | 0.982 | 0.977 |

TABLE 3.5: Stopping probabilities under increasing, flat toxicity rates. In each case, the probability of efficacy is 45% at each dose. The probability of toxicity is 10% at each dose in the first row, increasing in further rows. The baseline stopping scenario, where all design should stop with probability approximately 60%, is shown in bold. The EffTox software reports outcomes to the nearest whole percent.

quite willing to stop when toxicity is a fairly modest 10% greater than the threshold value. We will see how this affects the performance of the designs in non-toxic scenarios in subsequent sections. Despite choosing values different to those proposed by their authors, we do not expect *a priori* that this exercise favours any particular design because the methods have been calibrated to behave similarly in a neutral scenario.

The stopping parameterisations are summarised in Table 3.4.

### 3.5.5 Horizon Stopping Probabilities

Having calibrated the designs to stop with common probability at $(\pi_E, \pi_T) = (0.45, 0.5)$, we examined how the designs varied in their stopping probabilities in similar scenarios with flat efficacy and toxicity curves.

In Table 3.5, we kept the probability of efficacy at dose $i$ equal to 0.45 for $i = 1, .., 4$. We then analysed the stopping probabilities of all designs in scenarios with increasing levels of uniform toxicity. For example, in the first row, we set the probability of toxicity equal to 0.1 at all doses and analysed how reliably the designs advocated stopping. We did this for toxicity probabilities $0.1, ..., 0.7$. It is preferable that the designs do not stop when the toxicity rate is less than the upper threshold of 0.4 and they should show an increasing propensity to stop as the toxicity rate increases above 0.4.

WT is the design least likely to stop when the toxicity rate is 40% and below. At toxicity rates 50% and above, however, WT is the most likely to stop. ET is much

| Pr(Eff) | ET1 | ET2 | WT1 | WT2 | WT3 | WATU1 | WATU2 | WATU3 |
|---|---|---|---|---|---|---|---|---|
| 0.15 | 0.98 | 0.97 | 0.972 | 0.963 | 0.981 | 0.993 | 0.989 | 0.995 |
| 0.25 | 0.90 | 0.91 | 0.790 | 0.797 | 0.852 | 0.888 | 0.867 | 0.888 |
| 0.35 | 0.70 | 0.71 | 0.515 | 0.517 | 0.577 | 0.605 | 0.578 | 0.607 |
| 0.45 | 0.42 | 0.43 | 0.314 | 0.312 | 0.323 | 0.328 | 0.317 | 0.327 |
| 0.55 | 0.20 | 0.22 | 0.227 | 0.218 | 0.190 | 0.198 | 0.187 | 0.182 |
| 0.65 | 0.10 | 0.10 | 0.213 | 0.192 | 0.149 | 0.152 | 0.152 | 0.120 |
| 0.75 | 0.06 | 0.06 | 0.207 | 0.194 | 0.136 | 0.147 | 0.141 | 0.113 |

TABLE 3.6: Stopping probabilities under increasing, flat efficacy rates. In each case, the probability of toxicity is 40% at each dose. The probability of efficacy is 15% at each dose in the first, increasing in further rows. The EffTox software reports outcomes to the nearest whole percent.

more likely to stop than the other designs when toxicity is as low as 10%. Generally, WATU stops with probability between that of ET and WT. Looking within model, the different ET variations do not stop materially differently. In WT and WATU, suppressing the first stage in each model (WT3 and WATU3) increases the chances of stopping when toxicity is low, notably so in WT. The first stage exists to explore doses and learn about the dose-event probabilities. It is intuitive that removing it increases the probability of stopping even when treatment is tolerable and effective.

We perform a similar exercise in Table 3.6 to analyse stopping over a horizon of efficacy probabilities. In this analysis, the probability of toxicity is held constant at 0.4 at dose $i$ for $i = 1, ..., 4$ in every scenario, being the threshold maximum that we would accept. In the first row of Table 3.6, the probability of efficacy is 0.15 at each dose. We look at efficacy probabilities $0.15, 0.25, ..., 0.75$, being increments of 0.1 from the minimum efficacy threshold, 0.45. Once again, the efficacy and toxicity curves in each scenario are flat. Here, the designs should stop when the efficacy rate is less than 45% but they should show a decreasing propensity to stop as the efficacy rate increases above 45%.

All designs stop quite reliably when efficacy is very low, at 15%. ET is the design most likely to stop with modest efficacy at 25% and 35%. WT is the least likely to stop at low efficacy. A pertinent aspect of the WT design is that it only stops for low efficacy in its second stage, the maximisation stage. WT3 has a larger second stage than WT1 and WT2, so it follows that it stops more frequently for low efficacy. Once again, WATU sits between ET and WT, in the main. At high efficacy, ET is least likely to stop. The stopping probability in WT barely changes as efficacy increases

from 55% to 75%, and only falls modestly in WATU.

In summary, the described method has yielded parameterisations that stop with 60% probability in the baseline stopping scenario. Tables 3.5 and 3.6 show that the designs stop with broadly similar probabilities over efficacy and toxicity horizons. However, despite this systematic calibration, some heterogeneity persists.

### 3.5.6 General Simulation Study

The simulations presented above were about calibration. In this section, we investigate by simulation the performance of the designs in scenarios that are more likely to manifest in a trial situation, with efficacy and toxicity rates that vary by dose.

The scenarios under consideration are given in Table 3.7. Scenarios 1 to 5 have monotonically increasing efficacy and toxicity curves. The optimal dose differs in scenarios 1 to 3. In Scenario 4, efficacy escalates rapidly up to the highest dose. In Scenario 5, toxicity escalates rapidly up to the highest dose. Scenarios 6 and 7 show plateau efficacy curves. Scenarios 8 and 9 show unimodal efficacy curves. In scenario 10, all doses are inefficacious. Finally, in scenario 11, all doses are excessively toxic.

The selection probabilities of each of the designs are also given in Table 3.7. An admissible dose is one that has associated probability of efficacy greater than 45% and probability of toxicity less than 40%. Each scenario has exactly one optimal decision, be it to select the best admissible dose or to stop the trial where no dose is admissible. Where there are several admissible doses, the optimal dose is the one with the highest utility score, as determined by 2.5. Where there is no admissible dose, the optimal and admissible decisions are to stop the trial. In each row, the optimal decision is in bold text and the admissible decisions are underlined.

#### 3.5.6.1 Results

Scenario 1 is evidently one where it is easy to select an admissible dose but relatively difficult to select the optimal dose. Testament to this is that seven of the eight designs pick the true optimum dose less than half of the time. The efficacy curve is monotonically increasing and performance in WT and WATU is convincingly improved by tilting the models towards the monotonic efficacy skeleton. In contrast, ET2 loses

TABLE 3.7: Simulated selection probabilities. The probabilities of efficacy and toxicity are given at each dose, and utilities determined by (2.5). The optimal dose is shown in bold and acceptable doses are underlined. The EffTox software gives selection probabilities to the nearest whole percent.

| Scenario | | Dose 1 | Dose 2 | Dose 3 | Dose 4 | Stop |
|---|---|---|---|---|---|---|
| 1 | Pr(Eff) | 0.21 | 0.46 | 0.58 | 0.69 | |
| | Pr(Tox) | 0.11 | 0.16 | 0.24 | 0.30 | |
| | Utility | -0.32 | 0.07 | 0.23 | **0.34** | |
| | ET1 | 0.00 | 0.02 | 0.50 | **0.44** | 0.04 |
| | ET2 | 0.00 | 0.02 | 0.52 | **0.41** | 0.05 |
| | WT1 | 0.029 | 0.212 | 0.545 | **0.174** | 0.040 |
| | WT2 | 0.013 | 0.122 | 0.445 | **0.392** | 0.028 |
| | WT3 | 0.003 | 0.060 | 0.336 | **0.565** | 0.037 |
| | WATU1 | 0.005 | 0.078 | 0.579 | **0.249** | 0.089 |
| | WATU2 | 0.003 | 0.031 | 0.500 | **0.383** | 0.082 |
| | WATU3 | 0.001 | 0.024 | 0.515 | **0.372** | 0.088 |
| 2 | Pr(Eff) | 0.21 | 0.49 | 0.55 | 0.61 | |
| | Pr(Tox) | 0.05 | 0.32 | 0.53 | 0.69 | |
| | Utility | -0.32 | **0.04** | -0.05 | -0.17 | |
| | ET1 | 0.02 | **0.36** | 0.23 | 0.00 | 0.39 |
| | ET2 | 0.02 | **0.36** | 0.22 | 0.00 | 0.40 |
| | WT1 | 0.041 | **0.521** | 0.294 | 0.002 | 0.142 |
| | WT2 | 0.029 | **0.488** | 0.344 | 0.003 | 0.136 |
| | WT3 | 0.046 | **0.544** | 0.216 | 0.000 | 0.194 |
| | WATU1 | 0.026 | **0.336** | 0.348 | 0.001 | 0.289 |
| | WATU2 | 0.028 | **0.248** | 0.392 | 0.002 | 0.332 |
| | WATU3 | 0.033 | **0.223** | 0.338 | 0.001 | 0.404 |
| 3 | Pr(Eff) | 0.21 | 0.52 | 0.64 | 0.77 | |
| | Pr(Tox) | 0.08 | 0.17 | 0.32 | 0.53 | |
| | Utility | -0.32 | 0.17 | **0.25** | 0.16 | |
| | ET1 | 0.00 | 0.08 | **0.81** | 0.05 | 0.06 |
| | ET2 | 0.00 | 0.08 | **0.81** | 0.06 | 0.06 |
| | WT1 | 0.017 | 0.225 | **0.675** | 0.059 | 0.025 |
| | WT2 | 0.009 | 0.136 | **0.691** | 0.144 | 0.021 |
| | WT3 | 0.007 | 0.122 | **0.729** | 0.106 | 0.037 |
| | WATU1 | 0.006 | 0.127 | **0.792** | 0.018 | 0.058 |
| | WATU2 | 0.005 | 0.047 | **0.865** | 0.024 | 0.059 |
| | WATU3 | 0.002 | 0.031 | **0.870** | 0.026 | 0.070 |
| 4 | Pr(Eff) | 0.04 | 0.15 | 0.32 | 0.63 | |
| | Pr(Tox) | 0.07 | 0.12 | 0.19 | 0.31 | |
| | Utility | -0.60 | -0.43 | -0.16 | **0.25** | |
| | ET1 | 0.00 | 0.00 | 0.08 | **0.73** | 0.19 |
| | ET2 | 0.00 | 0.00 | 0.08 | **0.71** | 0.20 |
| | WT1 | 0.001 | 0.013 | 0.162 | **0.442** | 0.382 |
| | WT2 | 0.001 | 0.007 | 0.118 | **0.581** | 0.293 |
| | WT3 | 0.000 | 0.003 | 0.092 | **0.648** | 0.257 |
| | WATU1 | 0.003 | 0.005 | 0.082 | **0.433** | 0.476 |
| | WATU2 | 0.002 | 0.002 | 0.069 | **0.533** | 0.395 |
| | WATU3 | 0.003 | 0.001 | 0.070 | **0.534** | 0.391 |
| 5 | Pr(Eff) | 0.21 | 0.37 | 0.51 | 0.58 | |
| | Pr(Tox) | 0.07 | 0.16 | 0.28 | 0.52 | |
| | Utility | -0.32 | -0.07 | **0.10** | -0.01 | |
| | ET1 | 0.00 | 0.02 | **0.66** | 0.12 | 0.20 |
| | ET2 | 0.00 | 0.02 | **0.65** | 0.12 | 0.21 |
| | WT1 | 0.045 | 0.194 | **0.566** | 0.083 | 0.112 |
| | WT2 | 0.026 | 0.121 | **0.590** | 0.174 | 0.088 |
| | WT3 | 0.009 | 0.085 | **0.638** | 0.140 | 0.129 |
| | WATU1 | 0.017 | 0.084 | **0.586** | 0.112 | 0.201 |
| | WATU2 | 0.016 | 0.045 | **0.583** | 0.179 | 0.177 |
| | WATU3 | 0.014 | 0.039 | **0.571** | 0.174 | 0.201 |
| 6 | Pr(Eff) | 0.37 | 0.51 | 0.51 | 0.51 | |
| | Pr(Tox) | 0.04 | 0.11 | 0.28 | 0.38 | |
| | Utility | -0.05 | **0.17** | 0.10 | 0.03 | |
| | ET1 | 0.02 | **0.04** | 0.51 | 0.26 | 0.17 |
| | ET2 | 0.00 | **0.06** | 0.53 | 0.23 | 0.17 |
| | WT1 | 0.175 | **0.408** | 0.339 | 0.053 | 0.025 |
| | WT2 | 0.118 | **0.304** | 0.375 | 0.177 | 0.027 |
| | WT3 | 0.046 | **0.192** | 0.442 | 0.250 | 0.070 |
| | WATU1 | 0.070 | **0.247** | 0.443 | 0.137 | 0.103 |
| | WATU2 | 0.063 | **0.156** | 0.385 | 0.297 | 0.099 |
| | WATU3 | 0.056 | **0.119** | 0.389 | 0.298 | 0.138 |
| 7 | Pr(Eff) | 0.22 | 0.37 | 0.51 | 0.51 | |
| | Pr(Tox) | 0.04 | 0.11 | 0.28 | 0.38 | |
| | Utility | -0.30 | -0.06 | **0.10** | 0.03 | |
| | ET1 | 0.01 | 0.01 | **0.42** | 0.42 | 0.15 |
| | ET2 | 0.00 | 0.02 | **0.46** | 0.37 | 0.15 |
| | WT1 | 0.052 | 0.194 | **0.534** | 0.118 | 0.103 |
| | WT2 | 0.030 | 0.131 | **0.509** | 0.261 | 0.070 |
| | WT3 | 0.011 | 0.085 | **0.518** | 0.275 | 0.111 |
| | WATU1 | 0.018 | 0.097 | **0.514** | 0.164 | 0.206 |
| | WATU2 | 0.015 | 0.062 | **0.421** | 0.319 | 0.183 |
| | WATU3 | 0.014 | 0.052 | **0.404** | 0.309 | 0.222 |
| 8 | Pr(Eff) | 0.37 | 0.51 | 0.4 | 0.27 | |
| | Pr(Tox) | 0.04 | 0.11 | 0.28 | 0.38 | |
| | Utility | -0.05 | **0.17** | -0.07 | -0.32 | |
| | ET1 | 0.08 | **0.04** | 0.25 | 0.07 | 0.55 |
| | ET2 | 0.02 | **0.09** | 0.26 | 0.06 | 0.57 |
| | WT1 | 0.220 | **0.521** | 0.192 | 0.013 | 0.055 |
| | WT2 | 0.181 | **0.467** | 0.233 | 0.043 | 0.077 |
| | WT3 | 0.101 | **0.350** | 0.236 | 0.054 | 0.259 |
| | WATU1 | 0.150 | **0.396** | 0.190 | 0.022 | 0.242 |
| | WATU2 | 0.152 | **0.351** | 0.169 | 0.066 | 0.261 |
| | WATU3 | 0.133 | **0.323** | 0.140 | 0.066 | 0.337 |
| 9 | Pr(Eff) | 0.22 | 0.51 | 0.59 | 0.33 | |
| | Pr(Tox) | 0.04 | 0.11 | 0.20 | 0.35 | |
| | Utility | -0.30 | 0.17 | **0.26** | -0.21 | |
| | ET1 | 0.04 | 0.05 | **0.53** | 0.14 | 0.25 |
| | ET2 | 0.00 | 0.09 | **0.62** | 0.08 | 0.21 |
| | WT1 | 0.028 | 0.293 | **0.621** | 0.036 | 0.023 |
| | WT2 | 0.022 | 0.275 | **0.564** | 0.117 | 0.022 |
| | WT3 | 0.010 | 0.315 | **0.460** | 0.120 | 0.095 |
| | WATU1 | 0.014 | 0.332 | **0.523** | 0.033 | 0.098 |
| | WATU2 | 0.018 | 0.319 | **0.403** | 0.126 | 0.134 |
| | WATU3 | 0.015 | 0.286 | **0.384** | 0.125 | 0.190 |
| 10 | Pr(Eff) | 0.04 | 0.15 | 0.24 | 0.32 | |
| | Pr(Tox) | 0.04 | 0.11 | 0.20 | 0.35 | |
| | Utility | -0.60 | -0.42 | -0.29 | -0.23 | |
| | ET1 | 0.00 | 0.01 | 0.02 | 0.19 | **0.77** |
| | ET2 | 0.00 | 0.02 | 0.03 | 0.17 | **0.79** |
| | WT1 | 0.003 | 0.021 | 0.125 | 0.154 | **0.697** |
| | WT2 | 0.003 | 0.024 | 0.098 | 0.222 | **0.654** |
| | WT3 | 0.002 | 0.008 | 0.071 | 0.173 | **0.746** |
| | WATU1 | 0.021 | 0.006 | 0.037 | 0.070 | **0.866** |
| | WATU2 | 0.024 | 0.006 | 0.020 | 0.120 | **0.829** |
| | WATU3 | 0.017 | 0.006 | 0.012 | 0.108 | **0.857** |
| 11 | Pr(Eff) | 0.21 | 0.37 | 0.51 | 0.58 | |
| | Pr(Tox) | 0.47 | 0.55 | 0.62 | 0.69 | |
| | Utility | -0.47 | -0.30 | -0.19 | -0.20 | |
| | ET1 | 0.02 | 0.12 | 0.06 | 0.00 | **0.80** |
| | ET2 | 0.03 | 0.11 | 0.06 | 0.00 | **0.80** |
| | WT1 | 0.045 | 0.028 | 0.003 | 0.000 | **0.924** |
| | WT2 | 0.047 | 0.026 | 0.005 | 0.000 | **0.921** |
| | WT3 | 0.052 | 0.030 | 0.007 | 0.000 | **0.912** |
| | WATU1 | 0.020 | 0.060 | 0.033 | 0.000 | **0.886** |
| | WATU2 | 0.022 | 0.052 | 0.038 | 0.000 | **0.888** |
| | WATU3 | 0.027 | 0.044 | 0.036 | 0.000 | **0.894** |

relatively little using a much vaguer prior on $\beta_{E,2}$ to facilitate an efficacy curve that initially increases and then decreases as dose is increased. WT1 and WATU1 perform quite poorly compared to ET1 and ET2, however, WT3 performs the best in this scenario. The chances of choosing the optimum dose with WT and WATU in this scenario seem to be sensitive to choosing the right parameterisation, a feat that is very difficult prior to experimentation. Notably, no design is likely to do anything highly undesirable like stopping or selecting the lowest dose. Each dose selects an admissible dose with probability at least 90%.

The challenge with scenario 2 is that dose 2 is the only admissible dose. Generally, WT performs much better than ET and WATU. All designs show a predilection for dose 3, allowing a little too much toxicity when looking for superior efficacy. Scenario 2 demonstrates the increased risk of stopping when only one dose is admissible, especially in ET and WATU here. Surprisingly, the performances of WT2 and WATU2 have been hindered by biasing the models towards the monotonic (i.e. correct) efficacy skeleton.

In scenario 3, toxicity ramps up at the highest dose. All designs perform well, correctly picking dose 3 as optimal with high probability. The performance of WT slightly lags that of ET and WATU. All designs avoid doses 1 (for inactivity) and 4 (for excess toxicity) very well.

In scenario 4, efficacy escalates rapidly at the highest dose, with the first three doses being inefficacious. ET most reliably identifies dose 4 as the optimum. WT and WATU only identify the correct dose in the majority of cases once they have been inclined towards the monotonic efficacy scenario. All designs stop quite frequently.

In scenario 5, efficacy is high at the top two doses but toxicity is also high at dose 4. Dose 3 is the optimum dose and the only admissible dose. All designs are 56%-66% likely to correctly recommend dose 3. Once again, ET outperforms WT and WATU.

Scenario 6 is the first of two scenarios where the efficacy curve plateaus, in this instance at dose 2. Doses 2, 3 and 4 are all admissible but dose 2 is optimal. ET identifies the true optimal dose less than 10% of the time, opting for the admissible but inferior doses 3 & 4 more frequently. ET2 performs similarly poorly, suggesting that the obstacle in this scenario is not surmounted by tweaking the prior for $\beta_{E,2}$

alone. WT1 performs relatively well and is the only method more likely to choose dose 2 than 3. As might be imagined, WATU has performance between that of ET and WT. Performance in WT and WATU diminishes understandably as the models are biased towards the monotonic efficacy prior. It diminishes further still as the first stage is suppressed, suggesting that both methods benefit from their exploratory stages when biased towards an incorrect efficacy skeleton.

Efficacy plateaus again in Scenario 7, with dose 3 being the optimal and doses 3 & 4 admissible. WT is slightly superior to ET and WATU here. As with the previous scenario, WATU is harmed by inclining towards the monotonic efficacy skeleton. Interestingly, WT scarcely is.

Scenario 8 is the first of two unimodal efficacy curves. Dose 2 is the single admissible dose, and thus the optimal dose. Once again, WT is the superior design and WT1 shows the best performance. ET is very unlikely to select the best dose. The performance of ET, and to a lesser extent, WATU, is compromised by a propensity to stop too often. This suggests that idiosyncrasies in stopping behaviour persist, despite the calibration exercise.

Scenario 9 is another unimodal example. The optimal dose is dose 3 but dose 2 is also admissible. Once again, WT1 is the best performing design. ET performs much better in this scenario than in scenario 8, however, it still retains a tendency to stop too often. This is the first example where the vague prior for $\beta_{E,2}$ in ET2 has materially improved model performance. Consistent with this, WT2 and WATU2 both suffer from the inclination towards the monotonic efficacy skeleton.

Scenarios 10 and 11 both call for stopping and selecting no dose. Overall, all designs do this quite reliably, with overall better performance from WATU. Despite the exercise to calibrate stopping probabilities, there is reasonable heterogeneity in scenario 10, with the lowest probability (WT2) more than 20% less than the greatest probability (WATU1). WT designs are the least likely to stop here. This is slightly surprising given the extent to which we promoted stopping by choosing parameters such as $\alpha_E = \alpha_T = 0.3$ rather than the $\alpha_E = \alpha_T = 0.05$ proposed by Wages & Tait[98]. In contrast, WT designs are most likely to stop in scenario 11. This questions the success of our stopping calibration exercise and highlights the difficulty in parameterising stopping mechanisms.

TABLE 3.8: Probabilities of each design making optimal and admissible selections in the scenarios listed in Table 3.7, plus summary statistics. Information ratio (IR) is calculated as Mean / StDev. In each column, the best score is bolded.

| | Pr(Optimal) | | | | | | | | | | | All Scenarios | | | Scenarios 1-9 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | Mean | StDev | IR | Mean | StDev | IR |
| ET1 | 0.44 | 0.36 | 0.81 | **0.73** | **0.66** | 0.04 | 0.42 | 0.04 | 0.53 | 0.77 | 0.80 | 0.509 | 0.281 | 1.8 | 0.448 | 0.275 | 1.6 |
| ET2 | 0.41 | 0.36 | 0.81 | 0.71 | 0.65 | 0.06 | 0.46 | 0.09 | 0.62 | 0.79 | 0.80 | 0.524 | 0.271 | 1.9 | 0.463 | 0.264 | 1.8 |
| WT1 | 0.174 | 0.521 | 0.674 | 0.442 | 0.566 | **0.408** | **0.534** | **0.521** | **0.621** | 0.697 | **0.924** | 0.553 | 0.189 | 2.9 | 0.496 | 0.146 | 3.4 |
| WT2 | 0.392 | 0.488 | 0.691 | 0.581 | 0.590 | 0.304 | 0.509 | 0.467 | 0.564 | 0.654 | 0.921 | 0.560 | **0.164** | **3.4** | 0.509 | **0.115** | **4.4** |
| WT3 | **0.565** | **0.544** | 0.729 | 0.648 | 0.638 | 0.192 | 0.518 | 0.350 | 0.460 | 0.746 | 0.912 | **0.573** | 0.198 | 2.9 | **0.516** | 0.164 | 3.1 |
| WATU1 | 0.249 | 0.336 | 0.792 | 0.433 | 0.586 | 0.247 | 0.514 | 0.396 | 0.523 | **0.866** | 0.887 | 0.530 | 0.232 | 2.3 | 0.453 | 0.174 | 2.6 |
| WATU2 | 0.383 | 0.248 | 0.865 | 0.533 | 0.583 | 0.156 | 0.421 | 0.351 | 0.403 | 0.829 | 0.888 | 0.514 | 0.251 | 2.0 | 0.438 | 0.206 | 2.1 |
| WATU3 | 0.372 | 0.223 | **0.870** | 0.534 | 0.571 | 0.119 | 0.404 | 0.323 | 0.384 | 0.857 | 0.894 | 0.505 | 0.268 | 1.9 | 0.422 | 0.218 | 1.9 |
| | Pr(Admissible) | | | | | | | | | | | All Scenarios | | | Scenarios 1-9 | | |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | Mean | StDev | IR | Mean | StDev | IR |
| ET1 | **0.96** | 0.36 | 0.89 | **0.73** | **0.66** | 0.81 | **0.84** | 0.04 | 0.58 | 0.77 | 0.80 | 0.676 | 0.267 | 2.5 | 0.652 | 0.292 | 2.2 |
| ET2 | 0.95 | 0.36 | 0.89 | 0.71 | 0.65 | 0.82 | 0.83 | 0.09 | 0.71 | 0.79 | 0.80 | 0.691 | 0.252 | 2.7 | 0.668 | 0.276 | 2.4 |
| WT1 | 0.931 | 0.521 | 0.899 | 0.442 | 0.566 | 0.800 | 0.652 | **0.521** | **0.913** | 0.697 | **0.924** | 0.715 | 0.186 | 3.8 | 0.694 | 0.193 | 3.6 |
| WT2 | 0.958 | 0.488 | 0.827 | 0.581 | 0.590 | 0.856 | 0.770 | 0.467 | 0.839 | 0.654 | 0.921 | 0.723 | **0.174** | **4.2** | 0.708 | **0.179** | **4.0** |
| WT3 | 0.961 | **0.544** | 0.851 | 0.648 | 0.638 | **0.885** | 0.793 | 0.350 | 0.775 | 0.746 | 0.912 | **0.737** | 0.180 | 4.1 | **0.716** | 0.190 | 3.8 |
| WATU1 | 0.906 | 0.336 | **0.919** | 0.433 | 0.586 | 0.827 | 0.679 | 0.396 | 0.855 | **0.866** | 0.887 | 0.699 | 0.224 | 3.1 | 0.660 | 0.231 | 2.9 |
| WATU2 | 0.915 | 0.248 | 0.912 | 0.533 | 0.583 | 0.838 | 0.740 | 0.351 | 0.722 | 0.829 | 0.888 | 0.687 | 0.230 | 3.0 | 0.649 | 0.239 | 2.7 |
| WATU3 | 0.911 | 0.223 | 0.901 | 0.534 | 0.571 | 0.806 | 0.712 | 0.323 | 0.670 | 0.857 | 0.894 | 0.673 | 0.238 | 2.8 | 0.628 | 0.241 | 2.6 |

In a real trial situation, we pick one design that we trust will perform well in all (or most) scenarios. That trust is motivated by analysing performance over an indicative range of scenarios, as we have done here. The probabilities of each design selecting the optimal and admissible doses are shown in Table 3.8 for each scenario.

We can crudely estimate broad model performance across the range of scenarios by taking the mean probability of each design selecting the optimal or admissible dose. Likewise, we can estimate the variability of performance by calculating the standard deviation. A smaller standard deviation reflects greater homogeneity in performance. Means and standard deviations are included in Table 3.8. We have also calculated the *information ratio* (IR) for each design. In finance, IR is a metric used to appraise the risk-adjusted performance of a security or fund, defined as the expected value of a set of returns (potentially less some risk-free or benchmark rate) divided by the standard deviation of those returns, defined in equation 3.18.

$$\text{IR}\left(x_1, ..., x_n\right) = \frac{\text{Mean}\left(x_1, ..., x_n\right)}{\text{SD}\left(x_1, ..., x_n\right)} \tag{3.18}$$

A high IR is better, signifying high average value for relatively low variability. We have used the measure here to compare dose selection strategies across the scenarios, defined as the mean probability of selection divided by standard deviation of those probabilities. We seek a design that performs well on average in many scenarios with relatively little variability, and IR scores allow us to measure this. For

instance, a design that selects the best dose 50% of the time in scenarios A and B is preferable to one that selects the best dose with probability 100% in A and 0% in B. In this contrived example, the two strategies have the same selection probability but the first strategy has better risk-adjusted performance and greater IR.

The designs in the WT family offer the highest average selection percentages, the lowest standard deviations and naturally, the greatest IRs. Table 3.8 shows that WT does this when all scenarios are grouped, and when scenarios 1-9, the non-stopping scenarios, are considered. The IRs are smallest in this particular study for the ET designs. The WATU designs sit in between WT and ET. This is perhaps to be expected, given the fact that WATU combines elements from WT and ET.

Table 3.9 shows the average number of patients allocated by each design to each dose.

WT uses more patients in scenarios 10 & 11 where the correct decision is to stop. WT1 and WT2 use considerably more in scenario 10 where the doses are ineffective but tolerable. This is partly because WT designs do not stop for futility in the randomisation stage. WT3 sets the randomisation stage size to 0 and is able to stop for toxicity or futility at each dose decision. As such, it uses fewer patients than WT1 and WT2. In non-stopping scenarios, there is considerable heterogeneity in allocations.

Table 3.10 summarises the allocations to optimal and admissible doses. In scenarios 10 and 11, the desired action is to stop without allocating many patients so in these two scenarios we have instead provided the number of patients left unallocated, i.e. 30 minus the numbers allocated to doses 1, ...., 4. The IR statistics are again provided as a measure of risk-adjusted performance.

ET treats the greatest number of patients at optimal and admissible doses, on average. However, ET also has the greatest variability in allocation. Notably, in scenario 8 it allocates only a single patient on average to the single admissible dose. WT generally provides the lowest average allocation to attractive doses. This is understandable given its use of randomised allocation in the first half of the trial. Note for example that WT3, the variant with no randomisation, has allocation figures that are more comparable to the other designs. As we might expect by now, WATU provides performance between the two.

TABLE 3.9: Mean numbers of patients allocated to each dose. Figures for the optimal dose is shown in bold and acceptable doses are underlined. For stopping scenarios, the total is bolded and underlined.

| Scenario | | Dose 1 | Dose 2 | Dose 3 | Dose 4 | Total |
|---|---|---|---|---|---|---|
| 1 | Pr(Eff) | 0.21 | 0.46 | 0.58 | 0.69 | |
| | Pr(Tox) | 0.11 | 0.16 | 0.24 | 0.30 | |
| | Utility | -0.32 | 0.07 | 0.23 | **0.34** | |
| | ET1 | 0.1 | 0.6 | 16.5 | **11.9** | 29.1 |
| | ET2 | 0.0 | 0.6 | 16.9 | **11.3** | 28.8 |
| | WT1 | 4.9 | 8.6 | 13.2 | **2.9** | 29.6 |
| | WT2 | 4.1 | 7.0 | 12.9 | **5.8** | 29.8 |
| | WT3 | 0.9 | 2.4 | 11.8 | **14.4** | 29.5 |
| | WATU1 | 0.9 | 2.6 | 15.1 | **9.7** | 28.3 |
| | WATU2 | 0.7 | 1.5 | 13.7 | **12.5** | 28.4 |
| | WATU3 | 0.6 | 0.9 | 16.0 | **10.8** | 28.2 |
| 2 | Pr(Eff) | 0.21 | 0.49 | 0.55 | 0.61 | |
| | Pr(Tox) | 0.05 | 0.32 | 0.53 | 0.69 | |
| | Utility | -0.32 | **0.04** | -0.05 | -0.17 | |
| | ET1 | 0.4 | **7.2** | 13.9 | 0.9 | 22.4 |
| | ET2 | 0.4 | **7.1** | 13.7 | 0.9 | 22.1 |
| | WT1 | 7.0 | **12.2** | 9.6 | 0.2 | 29.0 |
| | WT2 | 6.6 | **11.6** | 10.5 | 0.3 | 29.0 |
| | WT3 | 4.4 | **12.1** | 10.9 | 0.8 | 28.2 |
| | WATU1 | 3.4 | **8.2** | 13.5 | 0.9 | 25.9 |
| | WATU2 | 3.1 | **6.1** | 15.1 | 1.0 | 25.3 |
| | WATU3 | 2.7 | **4.8** | 15.5 | 1.2 | 24.1 |
| 3 | Pr(Eff) | 0.21 | 0.52 | 0.64 | 0.77 | |
| | Pr(Tox) | 0.08 | 0.17 | 0.32 | 0.53 | |
| | Utility | -0.32 | 0.17 | **0.25** | 0.16 | |
| | ET1 | 0.1 | 2.1 | **23.0** | 3.7 | 28.9 |
| | ET2 | 0.0 | 1.9 | **23.1** | 3.6 | 28.6 |
| | WT1 | 5.1 | 9.0 | **14.2** | 1.5 | 29.8 |
| | WT2 | 4.6 | 7.7 | **14.5** | 3.0 | 29.8 |
| | WT3 | 1.5 | 4.2 | **18.1** | 5.7 | 29.5 |
| | WATU1 | 1.1 | 4.4 | **19.0** | 4.4 | 28.9 |
| | WATU2 | 0.9 | 2.6 | **20.5** | 4.9 | 28.9 |
| | WATU3 | 0.6 | 1.2 | **23.6** | 3.2 | 28.7 |
| 4 | Pr(Eff) | 0.04 | 0.15 | 0.32 | 0.63 | |
| | Pr(Tox) | 0.07 | 0.12 | 0.19 | 0.31 | |
| | Utility | -0.60 | -0.43 | -0.16 | **0.25** | |
| | ET1 | 0.0 | 0.1 | 7.4 | **18.8** | 26.3 |
| | ET2 | 0.0 | 0.1 | 7.4 | **18.8** | 26.3 |
| | WT1 | 3.7 | 6.2 | 10.6 | **6.4** | 26.9 |
| | WT2 | 2.6 | 4.9 | 10.9 | **9.1** | 27.5 |
| | WT3 | 0.5 | 0.9 | 8.0 | **18.0** | 26.4 |
| | WATU1 | 1.7 | 1.4 | 7.6 | **10.0** | 20.6 |
| | WATU2 | 1.4 | 0.4 | 6.5 | **13.2** | 21.5 |
| | WATU3 | 1.3 | 0.3 | 6.5 | **13.4** | 21.5 |
| 5 | Pr(Eff) | 0.21 | 0.37 | 0.51 | 0.58 | |
| | Pr(Tox) | 0.07 | 0.16 | 0.28 | 0.52 | |
| | Utility | -0.32 | -0.07 | **0.10** | -0.01 | |
| | ET1 | 0.1 | 0.7 | **19.3** | 6.1 | 26.2 |
| | ET2 | 0.0 | 0.7 | **19.4** | 6.0 | 26.1 |
| | WT1 | 5.4 | 8.5 | **13.3** | 2.0 | 29.2 |
| | WT2 | 4.3 | 7.2 | **14.1** | 3.8 | 29.4 |
| | WT3 | 1.3 | 3.5 | **16.8** | 6.8 | 28.4 |
| | WATU1 | 1.7 | 3.5 | **16.0** | 5.6 | 26.8 |
| | WATU2 | 1.4 | 2.1 | **15.9** | 7.6 | 26.9 |
| | WATU3 | 1.3 | 1.3 | **16.9** | 7.0 | 26.6 |
| 6 | Pr(Eff) | 0.37 | 0.51 | 0.51 | 0.51 | |
| | Pr(Tox) | 0.04 | 0.11 | 0.28 | 0.38 | |
| | Utility | -0.05 | **0.17** | 0.10 | 0.03 | |
| | ET1 | 0.4 | **1.0** | 16.4 | 8.8 | 26.6 |
| | ET2 | 0.1 | **1.2** | 17.1 | 8.0 | 26.4 |
| | WT1 | 7.2 | **10.4** | 10.7 | 1.5 | 29.8 |
| | WT2 | 5.6 | **8.6** | 11.8 | 3.8 | 29.8 |
| | WT3 | 2.0 | **4.9** | 13.6 | 8.7 | 29.2 |
| | WATU1 | 2.7 | **5.7** | 13.8 | 6.2 | 28.4 |
| | WATU2 | 2.2 | **3.7** | 12.8 | 9.6 | 28.4 |
| | WATU3 | 2.4 | **2.6** | 13.8 | 9.0 | 27.8 |
| 7 | Pr(Eff) | 0.22 | 0.37 | 0.51 | 0.51 | |
| | Pr(Tox) | 0.04 | 0.11 | 0.28 | 0.38 | |
| | Utility | -0.30 | -0.06 | **0.10** | 0.03 | |
| | ET1 | 0.3 | 0.4 | **14.1** | 12.4 | 27.2 |
| | ET2 | 0.0 | 0.5 | **15.1** | 11.3 | 26.9 |
| | WT1 | 5.3 | 8.4 | **13.0** | 2.5 | 29.2 |
| | WT2 | 4.2 | 7.1 | **13.4** | 4.8 | 29.5 |
| | WT3 | 1.2 | 3.6 | **14.8** | 9.0 | 28.6 |
| | WATU1 | 1.7 | 3.9 | **14.6** | 6.6 | 26.8 |
| | WATU2 | 1.4 | 2.6 | **13.1** | 9.8 | 26.9 |
| | WATU3 | 1.5 | 1.8 | **13.8** | 9.3 | 26.4 |
| 8 | Pr(Eff) | 0.37 | 0.51 | 0.4 | 0.27 | |
| | Pr(Tox) | 0.04 | 0.11 | 0.28 | 0.38 | |
| | Utility | -0.05 | **0.17** | -0.07 | -0.32 | |
| | ET1 | 1.4 | **1.0** | 11.5 | 5.6 | 19.5 |
| | ET2 | 0.4 | **1.5** | 12.1 | 5.0 | 19.0 |
| | WT1 | 8.3 | **11.5** | 8.9 | 0.9 | 29.6 |
| | WT2 | 6.7 | **10.1** | 10.3 | 2.3 | 29.4 |
| | WT3 | 3.4 | **7.7** | 10.6 | 5.0 | 26.7 |
| | WATU1 | 5.0 | **7.9** | 10.5 | 3.2 | 26.7 |
| | WATU2 | 4.8 | **6.5** | 9.7 | 5.4 | 26.4 |
| | WATU3 | 5.4 | **6.3** | 8.2 | 5.4 | 25.3 |
| 9 | Pr(Eff) | 0.22 | 0.51 | 0.59 | 0.33 | |
| | Pr(Tox) | 0.04 | 0.11 | 0.20 | 0.35 | |
| | Utility | -0.30 | 0.17 | **0.26** | -0.21 | |
| | ET1 | 1.2 | 1.3 | **15.9** | 7.4 | 25.8 |
| | ET2 | 0.1 | 1.8 | **18.3** | 5.8 | 26.0 |
| | WT1 | 4.6 | 9.4 | **13.9** | 2.0 | 29.9 |
| | WT2 | 3.8 | 8.5 | **13.3** | 4.3 | 29.9 |
| | WT3 | 0.8 | 8.0 | **12.6** | 7.5 | 28.9 |
| | WATU1 | 1.0 | 7.6 | **13.4** | 6.4 | 28.4 |
| | WATU2 | 1.1 | 6.8 | **11.4** | 8.9 | 28.1 |
| | WATU3 | 1.4 | 6.6 | **12.3** | 7.2 | 27.4 |
| 10 | Pr(Eff) | 0.04 | 0.15 | 0.24 | 0.32 | |
| | Pr(Tox) | 0.04 | 0.11 | 0.20 | 0.35 | |
| | Utility | -0.60 | -0.42 | -0.29 | -0.23 | |
| | ET1 | 0.1 | 0.3 | 5.4 | 9.8 | **15.6** |
| | ET2 | 0.0 | 0.3 | 5.7 | 9.1 | **15.1** |
| | WT1 | 4.0 | 6.8 | 10.0 | 4.0 | **24.8** |
| | WT2 | 2.8 | 5.4 | 10.5 | 6.5 | **25.2** |
| | WT3 | 0.9 | 2.4 | 7.9 | 9.6 | **20.8** |
| | WATU1 | 2.9 | 2.0 | 6.1 | 4.0 | **15.1** |
| | WATU2 | 2.7 | 1.2 | 5.1 | 6.2 | **15.3** |
| | WATU3 | 2.8 | 1.2 | 4.2 | 6.4 | **14.6** |
| 11 | Pr(Eff) | 0.21 | 0.37 | 0.51 | 0.58 | |
| | Pr(Tox) | 0.47 | 0.55 | 0.62 | 0.69 | |
| | Utility | -0.47 | -0.30 | -0.19 | -0.20 | |
| | ET1 | 0.7 | 4.6 | 9.2 | 0.6 | **15.1** |
| | ET2 | 0.8 | 4.4 | 9.2 | 0.5 | **14.9** |
| | WT1 | 11.7 | 2.7 | 3.7 | 0.1 | **18.2** |
| | WT2 | 11.7 | 2.7 | 4.0 | 0.1 | **18.5** |
| | WT3 | 8.5 | 2.5 | 4.9 | 0.4 | **16.3** |
| | WATU1 | 4.6 | 4.4 | 5.9 | 0.4 | **15.3** |
| | WATU2 | 4.2 | 3.5 | 6.9 | 0.4 | **15.1** |
| | WATU3 | 3.6 | 2.5 | 8.4 | 0.7 | **15.3** |

We further discuss model performance and mitigating factors in the next section.

TABLE 3.10: Mean number of patients that each design allocates to optimal and admissible doses, plus summary statistics. Information ratio (IR) is calculated as Mean / StDev. Scenarios 10 & 11 show mean patients left unallocated. The best score in each column is bolded. Full data is listed in Table 3.9.

| | Mean treated at optimal dose | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | Mean | StDev | IR |
| ET1 | 11.9 | 7.2 | 23.0 | **18.8** | 19.3 | 1.0 | 14.1 | 1.0 | 15.9 | 14.4 | 14.9 | 12.9 | 7.1 | 1.8 |
| ET2 | 11.3 | 7.1 | 23.1 | **18.8** | **19.4** | 1.2 | **15.1** | 1.5 | **18.3** | 14.9 | **15.1** | **13.3** | 7.2 | 1.8 |
| WT1 | 2.9 | **12.2** | 14.2 | 6.4 | 13.3 | **10.4** | 13.0 | **11.5** | 13.9 | 5.2 | 11.8 | 10.4 | 3.8 | 2.7 |
| WT2 | 5.8 | 11.6 | 14.5 | 9.1 | 14.1 | 8.6 | 13.4 | 10.1 | 13.3 | 4.8 | 11.5 | 10.6 | **3.3** | **3.2** |
| WT3 | **14.4** | 12.1 | 18.1 | 17.0 | 16.8 | 4.9 | 14.8 | 7.7 | 12.6 | 9.2 | 13.7 | 12.8 | 4.1 | 3.1 |
| WATU1 | 9.7 | 8.2 | 19.0 | 10.0 | 16.0 | 5.7 | 14.6 | 7.9 | 13.4 | 15.0 | 14.7 | 12.2 | 4.1 | 3.0 |
| WATU2 | 12.5 | 6.1 | 20.5 | 13.2 | 15.9 | 3.7 | 13.1 | 6.5 | 11.4 | 14.8 | 15.0 | 12.1 | 4.9 | 2.5 |
| WATU3 | 10.8 | 4.8 | **23.6** | 13.4 | 16.9 | 2.6 | 13.8 | 6.3 | 12.3 | **15.4** | 14.8 | 12.2 | 6.0 | 2.1 |
| | Mean treated at admissible dose | | | | | | | | | | | | | |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | Mean | StDev | IR |
| ET1 | **29.0** | 7.2 | **25.1** | **18.8** | 19.3 | 26.2 | **26.5** | 1.0 | 17.2 | 14.4 | 14.9 | 18.1 | 8.6 | 2.1 |
| ET2 | 28.8 | 7.1 | 25.0 | **18.8** | **19.4** | 26.3 | 26.4 | 1.5 | 20.1 | 14.9 | **15.1** | **18.5** | 8.5 | 2.2 |
| WT1 | 24.7 | **12.2** | 23.3 | 6.4 | 13.3 | 22.6 | 15.5 | **11.5** | **23.3** | 5.2 | 11.8 | 15.4 | 7.0 | 2.2 |
| WT2 | 25.7 | 11.6 | 22.2 | 9.1 | 14.1 | 24.2 | 18.2 | 10.1 | 21.8 | 4.8 | 11.5 | 15.8 | 7.0 | 2.3 |
| WT3 | 28.6 | 12.1 | 22.3 | 17.0 | 16.8 | **27.2** | 23.8 | 7.7 | 20.6 | 9.2 | 13.7 | 18.1 | 7.0 | **2.6** |
| WATU1 | 27.4 | 8.2 | 23.4 | 10.0 | 16.0 | 25.7 | 21.2 | 7.9 | 21.0 | 15.0 | 14.7 | 17.3 | **6.9** | 2.5 |
| WATU2 | 27.7 | 6.1 | 23.1 | 13.2 | 15.9 | 26.1 | 22.9 | 6.5 | 18.2 | 14.8 | 15.0 | 17.2 | 7.2 | 2.4 |
| WATU3 | 27.7 | 4.8 | 24.8 | 13.4 | 16.9 | 25.4 | 23.1 | 6.3 | 18.9 | **15.4** | 14.8 | 17.4 | 7.5 | 2.3 |

## 3.6 Discussion

We proposed a fusion of the EffTox and Wages & Tait methods to create a new seamless phase I/II trial design that we call WATU. Our primary motivation was to remove the need for adaptive randomisation in WT for situations where frequent adjustments to a randomisation algorithm could lead to operational inefficiency. We also examined by way of comparison a WT variant that also abrogates randomisation. We described a systematic method of calibrating designs to control for disparate stopping probabilities and used this method to calibrate eight phase I/II designs with a common stopping probability in a baseline case. We then conducted a broad simulation study, inspired by a real trial that used an EffTox design, to analyse performance of these designs. We used a novel approach to compare designs, borrowing a measure from finance to identify the design that provided the best risk-adjusted performance. In summary, we generally found that WT has superior performance in this setting in terms of determining the optimal doses, but not in terms of optimal allocation to doses. We found that on average, WATU performs similarly to ET.

Despite our exercise to calibrate stopping across the designs, we noted material

heterogeneity. For instance, Table 3.7 shows that the WT designs are on average much less likely to stop than ET and WATU in scenarios 1-10. This naturally makes us question the value of calibration exercise and ponder whether we are truly comparing like-for-like. Calibration could have been conducted differently. Instead of setting the prevailing efficacy equal to the threshold rate 45% at all doses, we could have chosen 100% to completely remove the estimation of efficacy as a source of variability in stopping.

Of our 11 simulation scenarios, two required stopping, five had an optimal dose in a monotonic efficacy curve, two used a plateau efficacy curve, and two an unimodal curve. Instead of taking the uniformly-weighted mean of selection probabilities, as we have in Table 3.8, we might have weighted the performance numbers by the scenario importance or prior likelihood. However, our counts of stopping, monotonic, plateau and unimodal scenarios broadly match our prior beliefs on the shape of the dose-efficacy curve. Thus, the scenarios have already been implicitly weighted in our situation.

Our simulation study investigates only a small number of the practically infinite possible scenarios. Different scenarios might have provided different conclusions. The scenarios we have chosen are motivated by a genuine clinical trial situation. They do not cover all eventualities but reflect those that are plausible and pertinent in this setting.

When randomly sampling efficacy and toxicity outcomes in simulations, we have assumed that the two events are independent. In a real trial, it is natural to consider that efficacy and toxicity might be dependent. For instance, a patient that ceases treatment early because of toxicity has less opportunity to receive the therapeutic benefit of treatment and is, presumably, less likely to achieve an efficacy event. The EffTox software offers the ability to sample dependent efficacy and toxicity outcomes, as does Wages' implementation of the WT design. It remains an exercise for further study to verify whether the conclusions we have made from this work persist in scenarios where efficacy and toxicity occurrences are associated.

It should be stressed that WT does not maximise utility so it may (legitimately) favour a different dose to ET and WATU. Trialists will appraise designs on their ability to select a dose with attractive qualities. In trial settings that explicitly quantify

utility using a metric, this can sensibly be interpreted as the dose with the highest utility or any dose with positive utility. In a setting that does not quantify utility, however, any dose satisfying efficacy and toxicity criteria may be attractive. As such, comparing the probability of selecting the optimal dose is not necessarily fair because 'optimal' is not uniquely defined. It is arguably fairer to compare the probability of designs selecting an admissible dose. Nevertheless, WT outperformed the two utility-maximising designs so the putative hindrance has not comparatively impaired the design in this study.

Our motivation for analysing different parameterisations of the three designs was to assess the extent to which performance might vary in our trial setting. For instance, ET2 has a much vaguer prior than ET1 on the coefficient of the squared-term in the efficacy logit model. In the monotonic efficacy scenarios (1-5), this vague prior reduces performance by 1-3%. In the plateau scenarios (6-7), performance improves marginally, and in the unimodal scenarios (8-9), performance improves 5-9%. IRs improve from 1.8 to 1.9 and 2.5 to 2.7, suggesting a slight model improvement overall in this setting. Naturally, if there was strong prior evidence to suspect a non-monotonic dose-efficacy curve, different efficacy priors would be used. We reiterate that the prior for $\beta_{E,2}$ is fixed by default to be $N(0, 0.2)$ in the MD Anderson EffTox implementation[45] but can be changed to suit.

The extent of variation within the WT family of designs is more noteworthy. In the monotonic scenarios, there are material improvements in the chances of picking the optimal dose in scenarios 1 and 4 for WT2 compared to WT1. With WT3, there is further benefit, between 3 and 18 percentage points, to suppressing the adaptive randomisation stage. On average, WT3 is 14.9% better than WT1 at selecting the optimal dose in the monotonic scenarios. In the non-monotonic scenarios, comparing WT2 to WT1 and WT3 to WT2, the probability of selecting the optimum falls with each model change. In scenarios 6-9, WT3 is 14.1% on average less likely to pick the optimum dose than WT1. Although we have only considered a modest number of scenarios, it appears that WT2 and WT3 are better than WT1 in monotonic scenarios and worse in non-monotonic scenarios, as expected. Naturally, the model used will be parameterised to match the investigators' prior beliefs on efficacy and toxicity.

This, however, suggests that it is difficult to improve on WT1 without prior information on the prevailing efficacy and toxicity scenarios, which is unlikely to be known with confidence in a situation where a dose-finding clinical trial was deemed necessary. Similar, albeit less pronounced, effects can be observed in the WATU family of designs.

We sought to remove randomisation from WT in pursuit of operational efficiency. Over the scenarios presented, the mean performance penalty to using WATU1 over WT1 in selecting the optimal dose is approximately 2.3%. This can be interpreted as the expected cost of using WATU1 over WT1. Notional operational efficiency might have been achieved but statistical efficiency, in this situation the probability of making the correct decision with a given set of resources, has marginally diminished. Whether the trade-off is acceptable depends on the prevailing trial situation.

Part of the performance difference will stem from removing randomisation. Randomisation in WT provides a facility to assess outcomes at different doses. This information is useful in estimating the dose-event curves. If randomisation is to be removed, as may or may not be desirable, the challenge is to remove it in the way with least loss. There will be trial scenarios when a modest performance penalty is an acceptable price to pay. However, in this study, it would have been superior to simply implement WT3, another design that avoids randomisation. WT3 has higher average performance and lower variability than each of the WATU designs we studied. Comparisons between WT3 and WATU2 or WATU3 are natural because each is biased towards the monotonic efficacy skeleton and avoids randomisation. WT3 outperforms WATU2 in making the optimal choice by 5.8 percentage points on average and performs better in eight of eleven scenarios. We did not consider a non-randomising variant of WT with uniform values for $\tau(k)$.

The conclusion from our simulation study is that WT3 abrogates the need for randomisation, thus achieving our operational efficiency objective, whilst offering superior performance, thus maintaining statistical efficiency. Our hybrid design WATU achieves the same operational objective but offers slightly inferior mean statistical performance, has greater heterogeneity in performance, and allocates marginally fewer patients at attractive doses.

In completing this research, we have gained a lot of experience working with

ET and WT. If designing a phase I/II trial from fresh, speaking purely in a personal capacity, the author would start with the WT design. The underlying probability models are simpler so it is easier to select parameters and to calculate the next dose. Our simulation study has suggested that WT has superior performance in non-monotonic efficacy scenarios. This would presumably be important where a phase I/II trial design was deemed preferable. If adaptive randomisation is operationally tolerable, we would use half of the patients in the first stage, as recommended by Wages & Tait.

These phase I/II trial designs are so-called because they perform tasks typical of phase I and II clinical trials. They are considered efficient because they potentially reduce the number of trials required to approve a treatment. However, they do not repeal the potential need for randomised phase II studies, where an experimental treatment is compared to a control to assess whether a likely large and expensive phase III trial is warranted. Retaining randomisation in WT's design actually offers an opportunity to further increase efficiency in the clinical trial pathway by additionally achieving the objective of these randomised phase II trials. In situations where a treatment can be ethically compared to a placebo, or can be offered adjunctly with a standard of care, it is possible to include the comparator in WT as a zero-dose control arm. This effectively prepends $d_0$ onto the list of ordered doses $d_1, ..., d_n$ under investigation. Patients can be randomised to $d_0$ or one of the admissible non-zero-doses and the randomisation probabilities would need to be adjusted to incorporate the new arm and provide a reasonable allocation of control patients throughout the trial. Depending on the clinical scenario, the stopping criteria would scrutinise toxicity at $d_1$ rather than $d_0$, the 'lowest' dose. As before, if there are no admissible non-zero doses, the trial would end with no dose being selected. At the culmination of the trial, the efficacy outcomes yielded by the optimal dose level could be compared to those yielded by $d_0$, achieving the goal of a randomised phase II trial. This complex design has the alluring potential to provide a genuine single-trial solution before phase III, achieving toxicity- and efficacy-oriented dose-finding and a randomised comparison with a legitimate control arm. A trial designed by the author is currently in set-up at the Cancer Research UK Clinical Trials Unit that seeks to implement this design.

## Chapter 4

# Design of a practice-changing clinical trial in an ultra-rare condition

**Background:** Wolfram syndrome is an ultra-rare neurological condition in children and young adults. One of the symptoms is progressive loss of visual acuity. Many clinical trials that analyse visual acuity as a primary outcome have taken place but none in Wolfram syndrome.

**Notable methods in this chapter:** We use mixed effects models, simulation, and consider several patterns of data-missingness to prospectively estimate the power of a clinical trial of sodium valproate. Our motivation to consider this level of detail is the severely constrained feasible sample size in this rare disease. Our literature review shows that this approach is novel in clinical trials of visual acuity.

**The implications on efficiency:** We demonstrate that clinical trials in ultra-rare diseases that achieve conventional error rates are feasible. Key to achieving this in our situation was using a repeated measures analysis to make use of all outcome information. Furthermore, using simulation allowed us to verify that randomisation favouring the experimental treatment could be used whilst maintaining a defensible design.

## 4.1 Introduction

The previous chapters were concerned with dose-finding clinical trials. In this chapter, we focus on a randomised efficacy study in a rare disease.

### 4.1.1 Wolfram Syndrome

Wolfram Syndrome (WS) is an ultra-rare, neurodegenerative disorder of children and young adults. It was first described in 1938 by Wolfram & Wagener[107], who reported on a family of nine siblings. Four of the siblings were affected with childhood onset diabetes mellitus, progressive optic atrophy leading to blindness, sensorineural deafness, and diabetes insipidus.

Wolfram Syndrome is caused by homozygous or compound heterozygous mutation of the WFS1 gene that encodes wolframin. It is registered in the Online Mendelian Inheritance in Man (OMIM) database with identifier 222300. The syndrome is also known as DIDMOAD, for Diabetes Insipidus, Diabetes Mellitus, Optic Atrophy and Deafness. The natural history typically involves diabetes mellitus in the first decade of life together with progressive optic atrophy. Deafness, neuropathic bladder and cranial diabetes insipidus appear in the second decade. The minimum criteria for diagnosis are diabetes mellitus and optic atrophy in an individual under 16 years of age.

The diagnosis of Wolfram syndrome is devastating for the affected person and their family or carers, as it virtually guarantees progressive sensory, motor, autonomic and mental faculty loss, and reduced life expectancy. The median age of death in patients is around 30 years and usually arises from respiratory failure as a result of brain stem atrophy[7]. Thankfully, Wolfram Syndrome is very rare, having a prevalence of approximately 1 in 770,000[8]. Using a population size of 64 million, being the estimated UK population in 2013 by the World Bank, this suggests there are approximately 83 affected patients in the UK.

There is no pharmaceutical treatment for Wolfram Syndrome. Instead, current therapies focus on the clinical management of symptoms. Being a multisystemic syndrome, different treatments exist to manage the different elements.

Treatments for diabetes mellitus aim to control metabolism by maintaining glycaemic targets. Interventions may include insulin injections, blood glucose and ketone testing, exercise, nutrition and smoking avoidance, and management of diabetic ketoacidosis (DKA) and hypoglycaemia. After the onset of diabetes insipidus, patients may receive desmopressin to treat bed-wetting. Vision loss may lead to cataract surgery or correction of refractive error, as appropriate. Hearing loss may lead to the use of hearing aids or cochlear implants. Neuropathic bladder may require clean intermittent self-catheterisation.

Reduction in the amount or activity of wolframin is associated with the death of neurons. At present, there are no known methods to prevent the neurodegeneration observed in Wolfram syndrome. Nagy *et al.*[67] showed that the re-entry of neurons into the cell cycle may be a step on the pathway to apoptosis in neurodegeneration. They also showed that $p21^{cip1}$ acts as an anti-apoptotic molecule. Significant down-regulation of $p21^{cip1}$ was seen in wolframin-depleted cells compared with controls. Those cells that retained $p21^{cip1}$ expression had much lower levels of apoptosis compared to those cells without. This led to the hypothesis that increased $p21^{cip1}$ expression may prevent neuronal death even in Wolframin-depleted cells.

### 4.1.2 Sodium Valproate

Nagy and colleagues at the University of Birmingham conducted a screen of 1,040 US Food and Drug Administration-approved drugs and short-listed 22 drugs that:

- are known to increase expression of $p21^{cip1}$;
- would likely be tolerable in children for chronic administration.

Five drugs were identified that showed clear evidence for protecting nerve cells from death in a Wolfram syndrome disease model. Sodium valproate, the sodium salt of valproic acid, was one of these drugs. It is classed as an anticonvulsant and currently approved in the treatment of epilepsy and bipolar disorder. It is known to cross the blood-brain barrier. Sodium valproate was selected for further study in Wolfram syndrome because it has been used for decades in children as an approved medicinal product and thus has an established safety profile. In patients with Wolfram syndrome, we expect sodium valproate to increase $p21^{cip1}$ expression

levels, increase wolframin expression, and ultimately to diminish neurodegeneration. There have been no clinical trials of sodium valproate in Wolfram syndrome to date.

### 4.1.3 The TreatWolfram Trial

We propose a randomised clinical trial to test the hypothesis that sodium valproate reduces the rate of neurodegeneration in patients with Wolfram syndrome. An early version of this trial was proposed by Lucinda Billingham and the lead clinical investigator, Timothy Barrett (TB). After feedback from the regulator advising against the use of Bayesian statistics, the trial underwent a complete redesign by Kristian Brock (KB) and TB. From this point, all aspects of a statistical nature were led by KB.

The treatment will be considered successful with respect to an outcome if it is associated with a significant, clinically relevant reduction in the rate of degradation. Naturally, improvements in symptoms would be very welcome but we do not necessarily expect this. There are many symptoms mentioned in the previous section that generally degrade over time. Knowing that the sample size of our trial would be severely constrained by the rarity of the syndrome, it would be important to identify the outcomes that are most conducive to study. Notwithstanding the fundamental requirement that outcomes are relevant and important to patients, we specifically seek outcomes with maximal information content. That is, we would like to identify outcomes associated with disease progression that typically see large changes over time with relatively low variability. To these ends, we were incredibly appreciative that Professor Tamara Hershey of the Wolfram Syndrome Research Clinic, Washington University in St Louis, USA, provided longitudinal data on 26 patients with Wolfram syndrome from her clinical cohort, collected under grant NIH HD070855, “Tracking Neurodegeneration in Early Wolfram Syndrome".

The remainder of this chapter proceeds as follows. In Section 4.2, we elaborate in detail on the St Louis dataset. We systematically search for a primary outcome measure that will promote an efficient analysis, conduct some preliminary regression modelling of the candidate outcome variable, and calculate sample sizes for hypothesis-testing trials, appraising these in light of the severely constrained feasible size of accrual. In Section 4.3, we apply the inferences from the St Louis dataset to

describe an experimental design for the TreatWolfram trial and a proposed method of analysis. We investigate the efficiency of our efforts through a simulation study, including considerations for missing data. In Section 4.4, we demonstrate by detailed literature review the novelty of our simulation-based approach incorporating different schemes for missing data to study a visual acuity outcome. In Section 4.5 we provide some discussion and in Appendix B, we appraise our efforts using the recently-published framework on randomised trials in small populations by Parmar *et al.*[72]. Finally, we conclude in Section 4.6.

## 4.2 The St Louis Cohort

The St Louis dataset contains observations over a period of six years on 26 patients with Wolfram syndrome of the Washington University Wolfram Syndrome Research Clinic. These patients would be candidates for the TreatWolfram trial because many meet the inclusion criteria. Data were assessed approximately annually. Figure 4.1 shows the frequency of the assessment times, relative to the first visit for each patient. We see that there is less long-term data than short-term data. The study recruited 11 patients in its first year and added between three and seven new patients each year in its second to fifth years. There has been very little drop-out. The decreasing number of observations in Figure 4.1 is consistent with the staggered recruitment times and patients passing through the follow-up schedule. Most observations times fall close to the anniversary of the initial visit, but there are a small number of assessment times that occur part way through the year.

The dataset contains up to six measurements per patient for each of the variables listed in Table 4.1. We will give sodium valproate to Wolfram patients under the expectation that it will diminish the rate of progression, rather than reverse their symptoms. The patients on trial will be children and young adults, with potentially very different baseline values. Furthermore, WS is a lifelong condition characterised by chronic deterioration of performance. Our primary interest in this dataset will be to characterise the typical *rate of change* in the clinical symptoms associated with WS. In Table 4.1 under $n$, we have listed the number of year-on-year change values. Much more data is provided for visual acuity than balance, for instance. We have

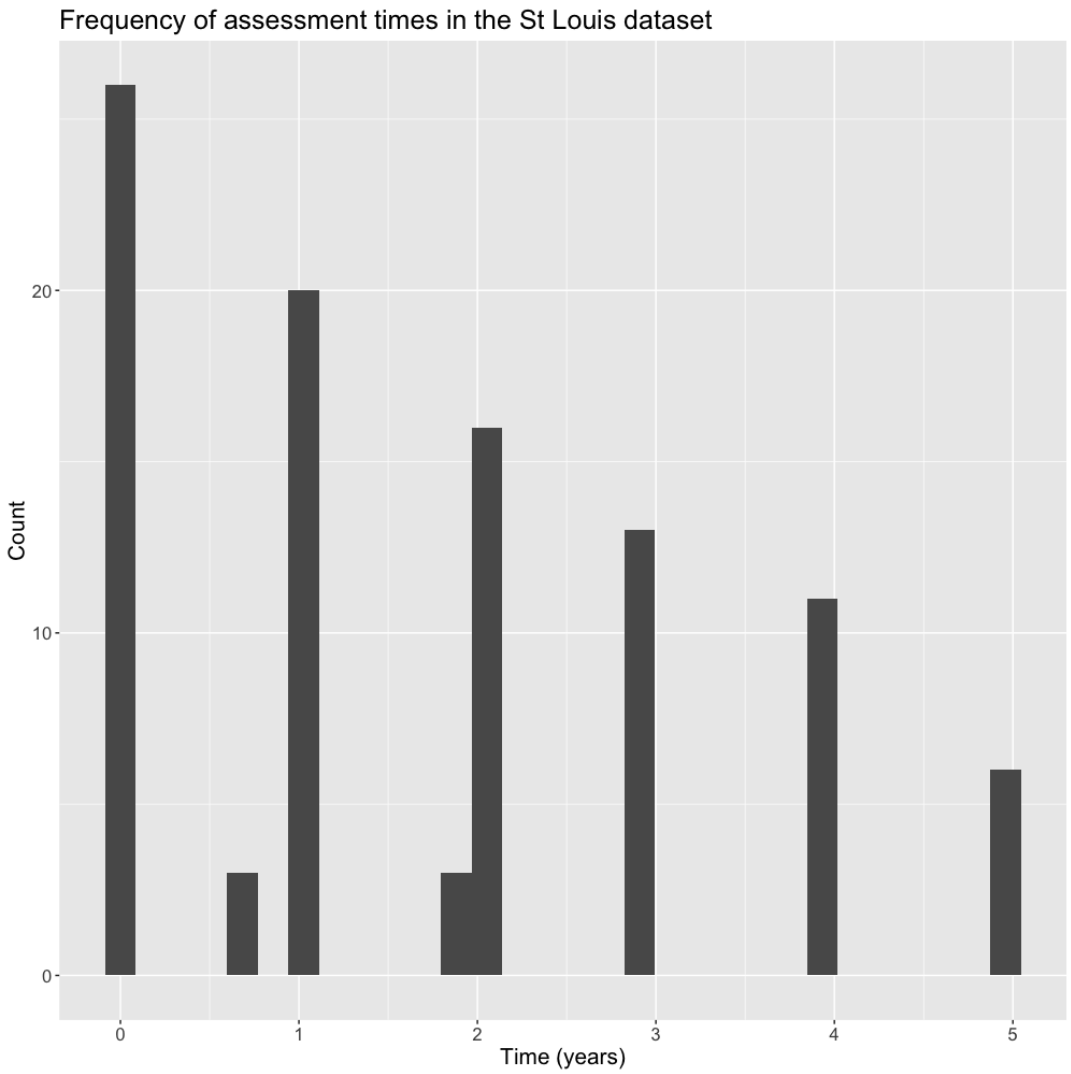



FIGURE 4.1: Assessment times of patients in the St Louis cohort. There are a small number of observations not near an anniversary of the initial visit.

| Variable | NumObs | NumPats | StdEff |
|---|---|---|---|
| Visual acuity | 68 | 24 | 0.6 |
| Colour vision | 59 | 23 | 0.5 |
| Ventral pons volume | 61 | 21 | 0.4 |
| Brainstem volume | 61 | 21 | 0.4 |
| Balance | 22 | 13 | 0.3 |
| Upsit (smell) | 66 | 24 | 0.2 |
| RNFL | 62 | 22 | 0.2 |
| Humphrey visual field, mean defect | 41 | 17 | $< 0.1$ |
| Humphrey visual field, pattern standard deviation | 41 | 17 | $< 0.1$ |

TABLE 4.1: Volume of information for variables in the St Louis dataset. NumObs is the number of observed year-on-year differences. NumPats is the number of patients that contributed at least one year-on-year difference, i.e. two consecutive values. StdEff is the absolute value of the mean of the year-on-year differences, divided by their standard deviation.

also presented the absolute value of the standardised effect size, being the mean year-on-year difference divided by the standard deviation of the differences. We use the absolute value because the direction of change is not pertinent to quantifying the volume of information.

Variables with higher standardised effect sizes are more conducive to study because the variability of the change is small relative to the mean. This makes it easier to observe a trend amidst the noise. By this measure, the variables most conducive to study are visual acuity, colour vision, ventral pons volume, and brainstem volume. These variables broadly cluster as measures of vision and brain size. Primary outcomes in clinical trials should be important to patients and conducive to study. Visual acuity is the most important of these variables to patients and their carers so it is extremely fortuitous that it ranks highest by our information measure. Any treatment that ameliorates the loss of vision will be welcome, for obvious reasons. Colour vision is understandably regarded as less important. Based on this, we investigate visual acuity as the potential primary outcome of our clinical trial. In the next section, we provide a detailed examination of this variable.

### 4.2.1 Visual Acuity

Visual acuity (VA) is measured on the LogMAR scale in clinics using *Early Treatment Diabetic Retinopathy Study* (ETDRS) charts. Patients read letters from a set distance and the scores reflect the number of letters correctly identified. In best corrected VA, patients wear glasses to correct for refraction disorders. Values are taken for each eye and generally range from 0, which represents perfect vision, to +2.0, which represents near blindness. Thus, increases in LogMAR represent deterioration. A LogMAR score of 0 is also referred to as "20/20", reflecting that a person can at 20 feet read letters that most humans will also be able to read at 20 feet. On this scale, LogMar 2.0 is expressed as 20/2000, to reflect that the person can read at 20 feet what most others could read at 2000 feet, making quite tangible the paucity of visual acuity in a patient with LogMAR 2.0. Values less than 0 are also possible, reflecting that the patient can read at distances greater than 20 feet what most others could read at 20 feet.

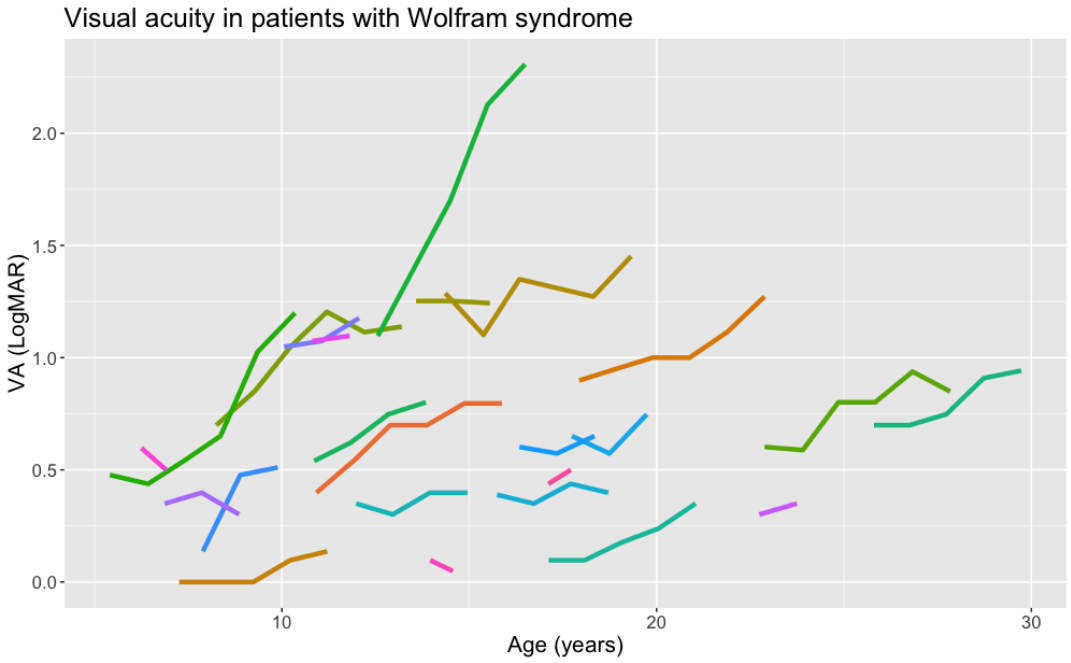



FIGURE 4.2: Visual acuity in 26 patients with Wolfram syndrome in the St Louis cohort.

Figure 4.2 shows VA of the 26 patients in the St Louis dataset. We took VA to be the average of the LogMAR scores in the left eye and right eye in 95 complete pairs of data. We use the simple mean because it is preferable to maintain vision in each eye and neither takes precedence over the other. For analysis, the mean of scores is an equivalent statistic to their sum in binocular patients and can be interpreted as a measure of the overall quality of vision. In the discussion, we consider a model that analyses eyes separately.

Figure 4.3 shows a scatter plot and locally estimated scatterplot smoothing (loess) line of the concurrent left-eye and right-eye assessments. Loess is an example of local regression, fitting simple regression models to small localised subsets of the data. The fits from these local models are combined to produce a smooth non-linear overall model.

Differences of up to 0.4 are observed between eyes but the average relationship showed by the blue line follows the line $y = x$ for $x \in (0, 1.5)$. There are two observations with very large VA values where this relationship breaks down. These relate to a single individual that experienced rapid progression in symptoms.

At one assessment, this same patient yielded a VA measurement in one eye only. Figure 4.4 shows two methods we considered for imputing an average value on this

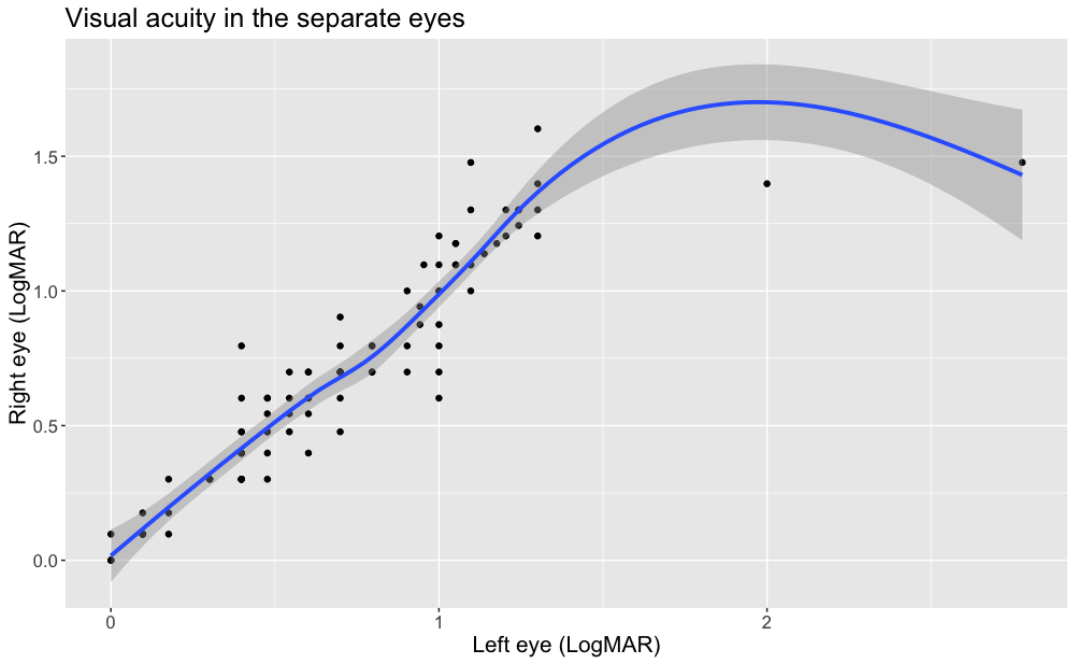
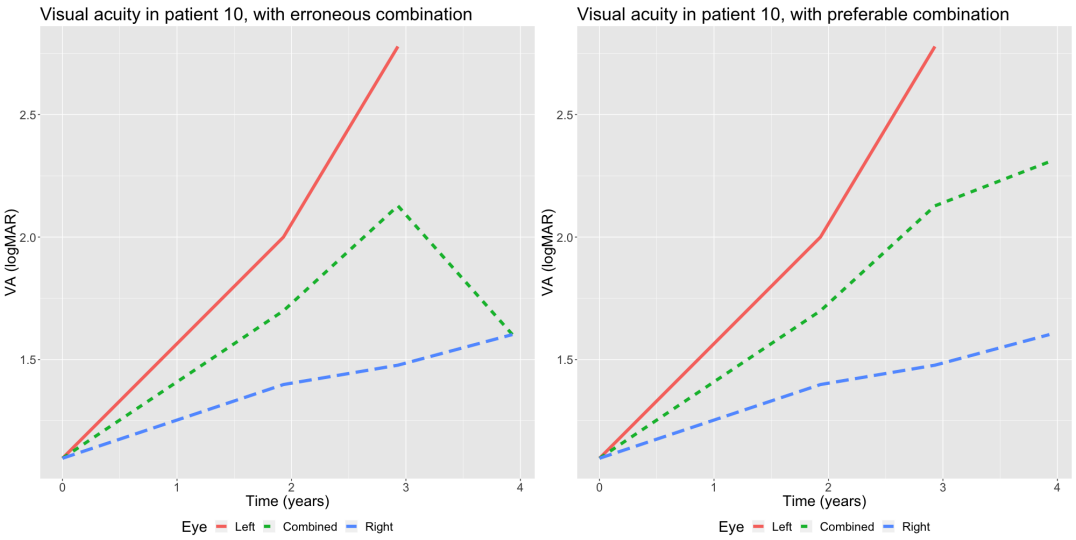



FIGURE 4.3: Left-eye vs right-eye visual acuity in the St Louis cohort.




(A) Assume same score (B) Assume same year-on-year change

FIGURE 4.4: Two methods for dealing with a missing VA value in one eye to create combined VA score.

occasion. There are values for the right eye at all periods but no observation for the left eye at Time = 4. In both methods, the combined values at Time = 0-3 are the simple arithmetic means. In Figure 4.4a, the combined value at Time = 4 is taken to be the single value provided in the right eye, effectively assuming the same score in each eye. We see that this is inappropriate because it artificially suggests that the patient experienced an overall improvement. In truth, the patient's vision had deteriorated over each period in each eye. In Figure 4.4b, the combined value at Time = 4 is created by imputing that the left eye deteriorated at the same rate as the right eye, and then taking the arithmetic mean. We used this method to impute the single missing value because it conveys deterioration at all periods and pragmatically makes use of all available data.

Including the imputed value, we have 96 observations for VA in total, an average of 3.7 observations per patient. The ages of patients at time 0 ranged from 5.4 to 25.8 years. The mean VA score at time 0 was 0.59 LogMAR units (range, 0.0 to 1.3).

Figure 4.2 demonstrates many noteworthy characteristics. We see that VA generally increases over time but is subject to a reasonable amount of natural variation. Patients appear to progress at a similar rate, irrespective of age and VA level. There is a stark outlier series that we have already identified. The patient with a LogMAR score of approximately 1.1 at age 13 progresses more rapidly than the rest of the patients, albeit from a high starting value. This demonstrates the types of progression that can occur, perhaps in a relative minority of cases. We will address the implications of this series with respect to modelling and hypothesis testing in later sections.

#### 4.2.1.1 Classical sample size calculations

Let us briefly consider conducting a standard parallel-groups randomised controlled trial (RCT), where patients are assigned to receive either sodium valproate or placebo. Let our primary outcome be change in VA. In a so-called *analysis of change scores*, VA would be assessed in all patients at baseline, again after a period of treatment, and changes calculated as the latter minus the former. The mean changes in each group would be compared using a two-sample $t$-test (for approximately normally-distributed data) or Mann-Whitney-U test (for non-normal data) to assess whether the rate of progression significantly differed. A persistent therapeutic benefit is

sought. Let us assume for elucidation that a 0.04 LogMAR units reduction in the rate of progression per annum is a meaningful treatment effect. Below we demonstrate using hierarchical regression models that the annual average progression in VA is approximately 0.08 units LogMAR. Thus, an annual difference between groups of 0.04 units would represent a treatment effect that halved natural progression. Here, we power tests to detect a difference of 0.04 in mean annual change in VA. We revisit effect sizes later in this chapter.

To estimate required sample sizes, we require estimates of the mean and variability of changes in VA. There are 68 one-year changes in VA in the St Louis dataset, with mean 0.067 and standard deviation 0.110 units LogMAR. We assume that the standard deviation of one-year changes is 0.110 in each arm. To achieve 80% power at a 5% significance level using a one-tailed $t$-test to detect a difference of 0.04 requires 95 patients per arm, thus a total sample size of 190. We perform this calculation using the software provided with the book by Machin & Campbell[61]. The required sample size vastly exceeds the number of patients in the UK. The sample size for a two-tailed test would be larger still. Greater efficiency is needed.

VA scores are not perfectly correlated so the standard deviation of two-year changes is less than twice the standard deviation of one-year changes. Perhaps increasing the assessment period will lead to a feasible sample size. There are 46 two-year changes in VA, with mean 0.157 and standard deviation 0.164 units LogMAR. To detect a 0.08 difference in the average two-year change with 80% power at a 5% significance level using a one-tailed $t$-test requires 53 patients per arm, or a total of 106. This too, is infeasible.

There are 28 three-year changes in VA, with mean 0.240 and standard deviation 0.224 units LogMAR. To detect a 0.12 difference in the mean three-year change with 80% power at a 5% significance level using a one-tailed $t$-test requires 44 patients per arm, or 88 in total. For this sample size to be feasible, we would have to recruit every patient in the UK. This is unrealistic.

The above power calculations are based on two-sample $t$-tests. This requires the changes to be approximately normally distributed. Non-parametric tests would require further increases in sample size to achieve the same power. We have established that an RCT using an analysis of change scores requires an infeasibly high

sample size. To achieve defensible power with a feasible sample size, it will require a more efficient approach to suit this particular situation. The series in Figure 4.2 seem amenable to repeated measures analysis. To these ends, we investigate modelling VA using mixed-effects regression models.

#### 4.2.1.2 Characterising VA through time

We note that the majority of the series in Figure 4.2 appear to progress at a similar rate over multi-year periods. There are instances where VA increases and decreases in consecutive years. This may be a manifestation of measurement error or reversion to the mean. This supports the use of longitudinal analysis to distinguish the multi-year trend from the short-term noise.

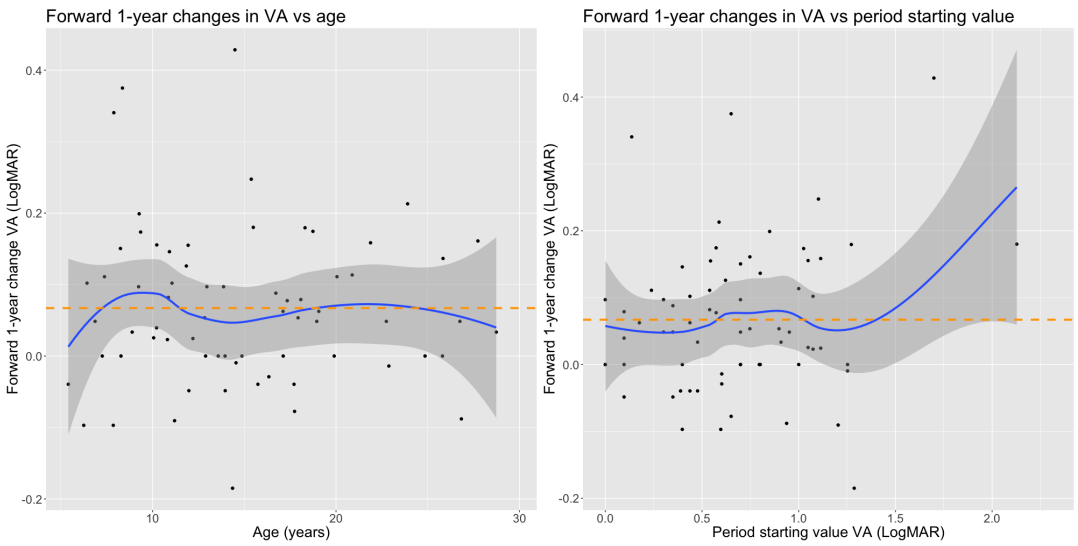
We also note the presence of outliers. For the purpose of assessing the treatment in a controlled experiment, we seek to estimate the average rate of change and outliers present a challenge. A single period analysis that compares VA scores before and after treatment in different groups would generally be more at risk of being affected by outliers than a longitudinal analysis, which has the opportunity to smooth out outliers if regression to the mean is subsequently observed.

We seek to characterise the dynamics of VA in Wolfram patients using linear mixed effects models. This hierarchical approach lets us reflect that repeated measures are nested within individuals through time.

There is evidently a population-level effect in time, because VA deteriorates as practically all patients age. We can immediately see from Figure 4.2 that patient-level intercepts are warranted because the series start at different values, irrespective of age.

Figure 4.5 shows the relationship of one-year forward changes in VA with A) age; and B) VA at the start of the period. For example, VA = 0.4 in an 11-year old at $t = 0$, increasing to VA = 0.6 one year later in the same individual would appear in plot 4.5a as the point (11, 0.2) and in plot 4.5b as (0.4, 0.2). Figure 4.5a provides no reason to believe that changes in VA systematically vary by age.

Figure 4.5b suggests that changes in VA can reasonably be assumed to be independent of the level of VA for VA < 1.5. However, there is a suggestion that outcomes increase more rapidly at higher values. This remains just a suspicion,




(A) Age (B) VA

FIGURE 4.5: Forward one-year change in VA vs age and VA at the start of the period. The blue lines shows the loess mean and the grey shaded regions the 95% uncertainty interval of the mean. The dashed orange lines show the mean one-year change in VA.

however, because both VA values greater than 1.5 in this dataset are yielded by the single aforementioned individual. Symptoms could deteriorate more rapidly when disease and symptoms are already well-developed. The World Health Organisation defines blindness to be best-corrected VA worse (i.e. scores greater) than 1.3 LogMAR[96]. It is plausible that the accuracy of visual acuity measurements decreases as blindness becomes more comprehensive. Alternatively, rapid progression might be a characteristic of this particular individual. The St Louis dataset does not allow us to distinguish amongst these scenarios because only one patient is seen at such levels. Patients with baseline LogMAR less than or equal to 1.6 are eligible for the TreatWolfram trial, so we may encounter patients with high values and thus strongly prefer a model that will handle outcome heterogeneity. As much as possible, we resist the temptation to remove this patient from the modelling.

Let $\tau_{ij}$ be the age in years of patient $i$ at VA observation $j$, for $i = 1, ..., 26$ and $j = 0, ..., 5$. Let $t_{ij} = \tau_{ij} - \tau_{i0}$ so that $t_{ij}$ is the time after baseline of observation $j$ for patient $i$. The $t_{ij}$ are continuous values, not integers or factors. This is desirable because, as Figure 4.1 demonstrates, assessments are not always conducted on anniversaries of the first visit. We see that $t_{i0} = 0 \ \forall \ i$. Let $y_{ij}$ be the VA observation for patient $i$ at time $t_{ij}$.

We consider the hierarchical model

$$y_{ij} = \alpha + a_i + \beta t_{ij} + e_{ij} \quad e_{ij} \sim N(0, \sigma^2), \quad a_i \sim N(0, \sigma_a^2) \tag{4.1}$$

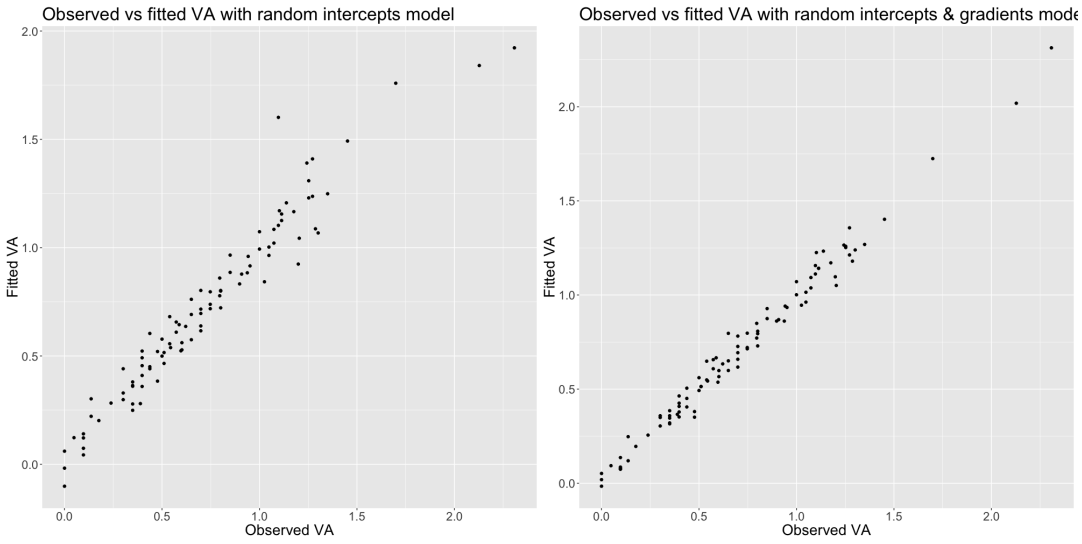
Here, $\alpha$ is the fixed-effect intercept, interpretable as the average baseline VA score; $a_i$ is the random intercept adjustment for patient $i$, assumed normally distributed with mean 0; and $\beta$ is the fixed effect for mean change in VA per annum, assumed constant in time and uniform across patients. We call this the random intercepts model.

Using the `nlme`[74] package in R[76], the estimated parameters are $\alpha = 0.567$ (s.e. 0.080) and $\beta = 0.082$ (s.e. 0.010), both with $p < 0.001$. In this cohort, this model estimates the mean annual increase in VA to be 0.082 LogMAR units per annum. This estimate differs from the simple mean period-on-period change of 0.067 given in Section 4.2.1.1 for two reasons: (i) the regression model has adjusted for some sources of variability to produce a better estimate of the change in VA attributable to the passage of time; and (ii) some of the period-on-period changes did not strictly cover periods of one-year, as demonstrated by Figure 4.1. The estimates of the standard deviations are $\sigma = 0.125$ (95% CI, 0.106 - 0.147) and $\sigma_a = 0.382$ (95% CI, 0.287 - 0.508). We have reported standard errors for coefficients but confidence intervals for standard deviations to reflect the summary statistics provided by the `nlme` package. The model was fit using general positive-definite structure for the variance-covariance matrix.

This model yields the fitted values shown in Figure 4.6a. Overall, we see that the fitted values are close to those observed for VA scores up to approximately 1.75. There are two values observed greater than 2 that are not fit particularly well. These values relate to the same individual noted above with rapid progression at high VA scores, suggesting the benefit in accounting for heterogeneity in gradients. First, however, we consider further population-level terms.

Seeking to improve the model fit, particularly at high VA values, we consider non-linear functions of time. We investigate alternative models that use the square of $t_{ij}$

$$y_{ij} = \alpha + a_i + \beta t_{ij} + \gamma t_{ij}^2 + e_{ij} \quad e_{ij} \sim N(0, \sigma^2), \quad a_i \sim N(0, \sigma_a^2) \tag{4.2}$$




(A) Random intercepts (B) Random gradients

FIGURE 4.6: Fitted and observed VA values under the two models. Each was fit using the `nlme` package using REML.

and the square-root of $t_{ij}$

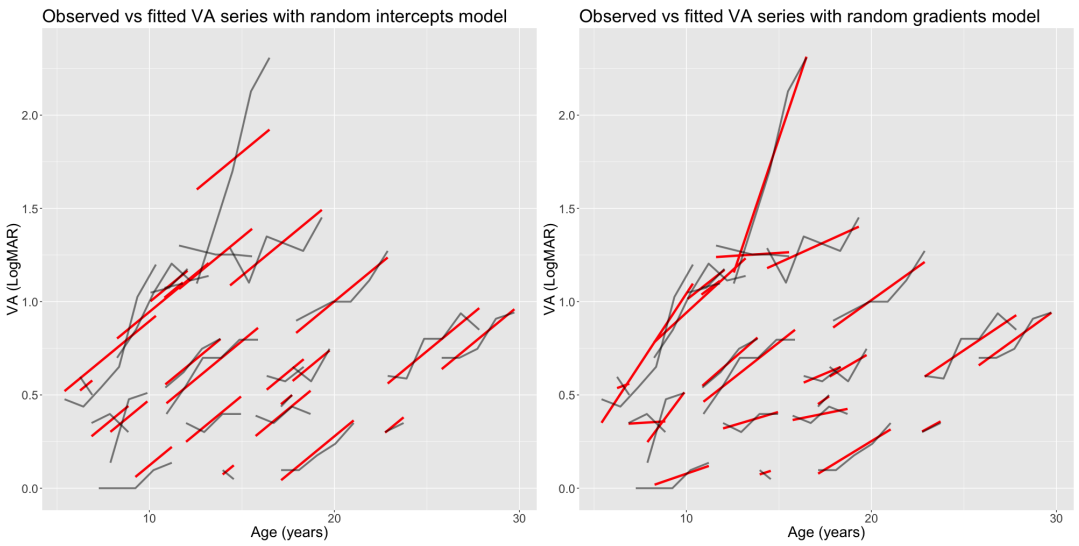
$$y_{ij} = \alpha + a_i + \beta t_{ij} + \zeta\sqrt{t_{ij}} + e_{ij} \quad e_{ij} \sim N(0, \sigma^2), \quad a_i \sim N(0, \sigma_a^2) \tag{4.3}$$

as additional fixed effects. The incremental benefit of each of these models is assessed by testing the nested models via likelihood ratio test using the `anova.lme` function in `nlme`[74]. As recommended by Pinheiro & Bates[73], these models and the comparator (4.1) were fit using maximum likelihood (ML) because testing nested models with different fixed effects structures is invalid under *restricted maximum likelihood* (REML). The $p$ values are 0.45 and 0.25 respectively. The case for including the extra variables is not particularly strong. More importantly, neither rectifies the poor model fit at high VA values. We seek improvements elsewhere.

Our difficulties with the outlier patient have been largely driven by their heterogeneous rapid rate of progression. This suggests we extend (4.1) by considering the following model with patient-specific gradients with respect to time:

$$\begin{gathered} y_{ij} = \alpha + a_i + (\beta + b_i)t_{ij} + e_{ij}, \\ e_{ij} \sim N(0, \sigma^2), \;\; a_i \sim N(0, \sigma_a^2), \;\; b_i \sim N(0, \sigma_b^2) \end{gathered} \tag{4.4}$$

We call this the random gradients model. It reflects that individuals will commence




(A) Random intercepts (B) Random gradients

FIGURE 4.7: Observed (dark grey) VA series and those estimated by the two mixed effects models (red).

our study under different levels of visual acuity and that through repeated measures in time, each will experience their own rate of progression. It yields the fitted values shown in Figure 4.6b. We see that allowing heterogeneity in gradients improves model fit at high VA values. A likelihood ratio test of the nested random intercepts and random gradients models yields $p < 0.001$, confirming that this model is very likely superior for modelling the St Louis data. For this test, models (4.4) and (4.1) were fit by REML. As described in Pinheiro & Bates[73], tests of nested models fit by REML that differ only in random effects are valid. Furthermore, the ML method of fitting mixed models has an undesirable tendency to underestimate the size of the variance components, a flaw rectified by fitting by REML.

The plots in Figure 4.7 clearly demonstrate how including random gradients improves model fit. They show the fitted (red) series superimposed on the observed (dark grey) VA series for each patients using the random intercepts (left) and random gradients (right) models. The random intercepts model is surprisingly good for the majority of patients. However, it does not model at all well those that progress quickly. These patients still progress approximately linearly, albeit at a much faster rate. The random gradients model facilitates this.

The random gradients model yields estimates $\alpha = 0.571$ (s.e. 0.072), $\beta = 0.070$ (s.e. 0.017), $\sigma = 0.074$ (95% CI, 0.061 - 0.090), $\sigma_a = 0.356$ (95% CI, 0.266 - 0.477) and

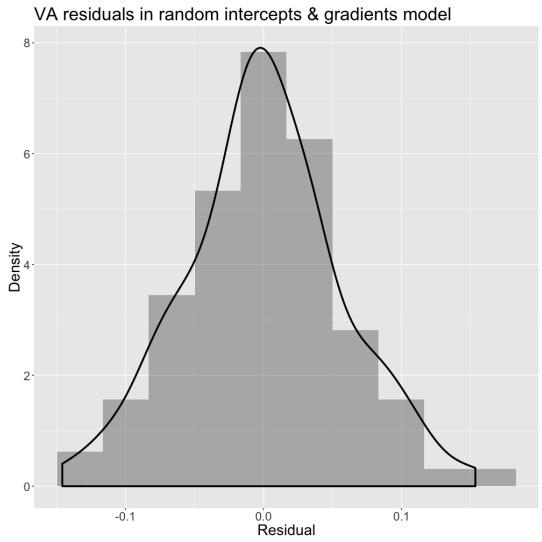


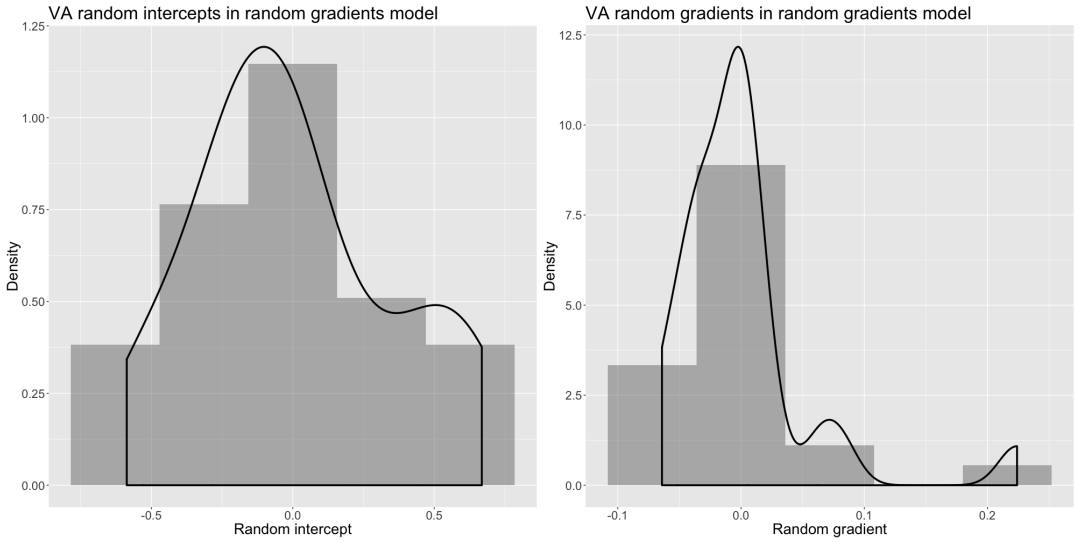

FIGURE 4.8: Raw residuals of the random gradients model. The solid black line shows a Gaussian kernel smoother to estimate the distribution.

$\sigma_b = 0.071$ (95% CI, 0.049 - 0.103). The estimated mean annual progression is 0.012 LogMAR units lower when we allow heterogeneity in gradients. This reflects the reduced influence of the patient with rapid progression.

Figure 4.8 shows the distribution of the "raw" residuals, i.e. the observed values less the fitted values, of the random gradients model. We see that they are approximately normal, as required.

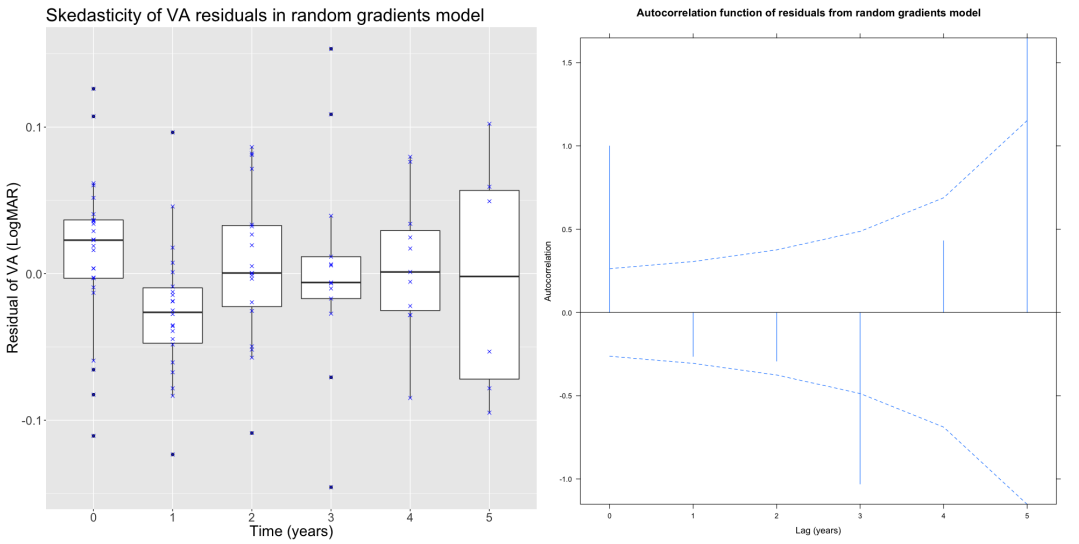
Figure 4.9 shows the distributions of the random parameters, both also assumed normal in (4.4). Figure 4.9a shows that the random intercepts are indeed approximately normal. In contrast, Figure 4.9b also shows clear central tendency, but also shows a large positive outlier in the random gradients. Perhaps unsurprisingly, the outlier is the patient with very large VA values. The result is that the estimate above for $\sigma_b^2$ is possibly inflated by the data for this single patient. All else being equal, this would overestimate the variability of outcomes and dictate an inflated sample size for a given power. We return to this in following sections.

Figure 4.10 shows two further diagnostic plots. Figure 4.10a shows that residuals are centred at zero for each assessment point, and that their variance is approximately constant through time. Finally, we consider the autocorrelation of residuals. Figure 4.10b shows the auto-correlation function for the residuals, produced using




(A) intercepts (B) gradients

FIGURE 4.9: Distributions of the patient-specific parameters from the random gradients model, assumed normal. The solid black lines show Gaussian kernel smoothers.




(A) Skedasticity of residuals (blue crosses). Loess smoother demonstrates the mean. (B) Auto-correlation function, with 1% significance bounds.

FIGURE 4.10: Further diagnostic plots of residuals from the random gradients model.

the `ACF` command in `nlme`[74]. Mirroring Pinheiro & Bates[73, p. 241], we add significance bounds (blue lines) at the 1% level. Assuming the observations are taken at points equally-spaced through time, which Figure 4.1 demonstrates to be overwhelmingly the case, we see from Figure 4.10b that the residuals at lags 1 and 2 years do not show significant autocorrelation. In contrast, the residuals are apparently significantly autocorrelated at lags 3 and 5 years. Two factors motivate us to cautiously interpret this finding as chance. The first is the lack of material autocorrelation at lower lags. If the residual process retained information from previous observations, for instance via an auto-regressive moving average (ARMA) process, we would expect to see significant autocorrelation at short lags too. Secondly, there are very few pairs of observations at these long lags.

Of our two candidate models for fitting VA in the St Louis data, the random-gradients model is clearly superior. We consider further embellishments to this model by testing two other population-level effects. We have demonstrated that VA is related to age. Time is already included as a population-level effect and is perfectly correlated with age, so age at each assessment point is not a sensible covariate to add. However, age at baseline could add marginal information. Testing the addition of this covariate via nested models estimated by ML yields $p = 0.77$. The data are largely consistent with the additional effect associated with this variable being zero. This is perhaps not surprising because the information contained in baseline VA is already reflected in the model by the patient-level intercepts, $a_i$.

Lastly, we consider a population-level effect with respect to sex. Again, a test via nested models yields $p = 0.46$ and no strong case for inclusion.

We have demonstrated in figures above that the fit of our random gradients model is good and that errors are reasonably independent. The model has validity because it maps to the research question we seek to answer and incorporates effects that are intuitive and biologically plausible. The assumption of additive effects is reasonable given the small number of terms that combine to essentially yield patient-specific straight lines. The one modelling assumption that is questionable is normality of the random terms. Gelman & Hill[38] identify this as the least important of the modelling assumptions. Nevertheless, it is a topic we repeatedly visit in subsequent sections when we use the random gradients model to simulate and

analyse VA paths.

## 4.3 TreatWolfram Statistical Design

We present our design for an international, double-masked, randomised, placebo-controlled trial of sodium valproate versus control in patients with Wolfram syndrome. This design was conceived and developed by KB. There is currently no pharmaceutical treatment for Wolfram syndrome so the control arm will be a placebo, manufactured to match the appearance of sodium valproate.

A severe constraint in designing a pivotal clinical trial in an ultra-rare disease is the limited sample size. We have already noted above that a conventional experimental design requires infeasibly high accrual. A more efficient analysis is required to achieve conventional error rates with our restricted sample size.

In this chronic disease setting, we are able to measure the outcome variable many times. In fact, the treatment is only likely to materially improve the lives of patients if it demonstrates prolonged efficacy. In the following sections, we calculate the required sample size for the TreatWolfram trial assuming a repeated measures analysis of the candidate primary outcome, VA. The sample size analysis will be informed by the St Louis cohort.

A key component of a repeated measures is the frequency with which the outcomes are measured. All else held constant, more frequent assessments yield more information and a more powerful analysis.

VA is measured in clinics using standardised charts. The assessment is not costly nor invasive and the frequency of assessment is constrained only by how often it is reasonable to expect patients to attend clinic. We propose to measure this outcome at baseline and every six months for three years, giving the seven assessment times $t_{\text{VA}} = (0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0)$.

We are keen to maximise the chances that each patient will experience therapeutic benefit on the trial. Given the absence of a standard treatment, we will investigate the feasibility of randomisation that favours the experimental arm. However, the most efficient allocation is equal-sized groups. We will tolerate modest deterioration in efficiency arising from non-equal randomisation if it achieves the patients'

stated preference to increase the chances of receiving the experimental drug. We investigate this via simulation.

As with any clinical trial, we expect to collect less than complete outcomes as some assessments may not be performed as planned, or some patients may drop out of the study. Less than complete data collection reduces the efficiency of analysis. Furthermore, the pattern in which data is missing is potentially pertinent in longitudinal analyses. We present three methods for simulating missing data and analyse their effects on statistical efficiency.

### 4.3.1 Sample size for longitudinal analysis

We propose a longitudinal analysis because it will use more information and be more efficient than a single post-baseline comparison. We showed in section 4.2.1.2 that VA series are amenable to analysis by mixed effects models. We investigate in this section whether we can achieve conventional clinical trial error rates using our limited sample size.

We follow the example of Diggle *et al.*[31, p. 30] to calculate the required sample size for a test by repeated measures model of a continuous outcome. Using equal-sized arms, for a two-tailed test with power $P$ and significance $\alpha$, Diggle gives the required per-arm sample size to be

$$N = \frac{2(z_\alpha + z_Q)^2\sigma^2(1-\rho)}{ns_t^2d^2} \tag{4.5}$$

where $Q = 1 - P$; $z_\alpha$ and $z_Q$ are quantiles from the unit normal distribution; $\sigma^2$ is residual variance discussed below; $\rho = \text{Corr}(y_{ij}, y_{ik})$ for all $j \neq k$; $d$ is the difference in slope coefficients to be detected; $s_t^2 = \sum_j (t_j - \bar{t})^2/n$ is the within-subject variance of the explanatory variables, $t$; and $n$ is the number of assessments times in $t$. For a two-arm trial, the total sample size is $2N$.

Diggle *et al.* estimate the variability of residuals, $\sigma^2$, using an ordinary least squares regression model

$$y_{ij} = \alpha + \beta t_{ij} + e_{ij} \quad e_{ij} \sim N(0, \sigma^2) \tag{4.6}$$

TABLE 4.2: Serial correlations of VA in the St Louis dataset, at baseline ($VA_0$) and years 1-3.

| | $VA_0$ | $VA_1$ | $VA_2$ | $VA_3$ |
|---|---|---|---|---|
| $VA_0$ | 1 | | | |
| $VA_1$ | 0.949 | 1 | | |
| $VA_2$ | 0.923 | 0.978 | 1 | |
| $VA_3$ | 0.867 | 0.970 | 0.980 | 1 |

for $i = 1, ..., 2N$ and $j = 1, ..., n$. This model is used only to estimate $\sigma^2$. It is not the proposed analysis model. Model (4.6) fit to the St Louis data yields the estimate $\sigma^2 = 0.164$.

We will initially investigate $\alpha = 0.05$ and power $\geq 80\%$ to comply with conventional clinical trial error rates. Where possible, we prefer to increase power. We estimate the required parameters using the St Louis dataset. The correlation parameter in (4.5) is assumed to be the same at all lags. Smaller values for $\rho$ demand larger sample sizes because previous response values contain less information about future values. Thus, to avoid the risk of under-powering the study, we seek to estimate the lower bound of $\rho$.

The serial correlations in VA at baseline and years 1-3 years, chosen to match the time-frame over which we will analyse this outcome in TreatWolfram, are shown in Table 4.2. We see that the serial correlation values are at least 0.867. For conservative sample size estimation and the reasons explained above, we assume $\rho = 0.867$.

With assessment times $t_{\text{VA}}$ (defined above), we have $s^2_{t_{VA}} = 1.0$ and $n = 7$. As before, we power to detect a reduction of 0.04 LogMAR units in the rate of increase in VA per annum. Using these values, by (4.5) we require $N = 25$ patients per arm to achieve 80% power to detect the specified difference at a 5% significance level. Similarly, we require $N = 29$ to achieve 85% power and $N = 34$ to achieve 90% power. These represent marked improvements over the sample sizes in Section 4.2.1.1 and ably demonstrate the boost to efficiency that comes from using a repeated measures analysis.

Paradoxically, one of the distinct benefits of designing a clinical trial in a condition as rare as Wolfram syndrome is that many feasible patients can be identified before the trial has commenced. Many countries, including the United Kingdom,

have registries of patients and/or patient and carer support groups. Patients attend routine clinics for monitoring and management of symptoms. Patient groups and clinical leads have been consulted in several European countries to estimate the likely number of patients that can be recruited. Thus, in a disease where trial recruitment will be highly constrained, we can say with reasonable confidence the exact number of patients that are feasible to recruit. In the UK, the lead investigator predicts they can recruit 48 patients. Sites in France, Spain and Poland indicate that they will be able to recruit 11, 6 and 5 patients respectively, leading to a maximum feasible sample size of 70 patients. Further recruitment would require many centres with low potential recruitment and this would contribute materially to the trial cost.

In the proceeding sections, we treat 70 patients as the maximum feasible accrual. We have shown above that power up to 90% can be expected to detect an annual difference in progression of 0.04 LogMAR units *if all data is collected*. We investigate the sensitivity of efficiency to this last assumption in coming sections.

### 4.3.2 Measuring statistical performance using simulation

The sample sizes in the previous section are feasible given the expected number of patients in the UK and our European neighbour countries. Those sample sizes assume 1:1 randomisation. Given the dearth of pharmaceutical treatments for Wolfram syndrome, the fact that it afflicts children, and the fact that symptoms generally progress continuously, we and our funders are highly motivated to use randomisation that allocates more patients to the experimental treatment. Equation (4.5) above assumes equal-sized arms. Diggle *et al.*[31] do not give a version for general randomisation ratios, and we could not find one in the literature. We use simulation to investigate the feasibility of unequal arms by estimating statistical power at various allocation ratios in favour of sodium valproate. Furthermore, it is inevitable in a longitudinal analysis that some data will be missing. Simulation allows us to easily incorporate various patterns for data loss and assess their impact on statistical efficiency.

We describe in the next section our methods for sampling patient outcomes. In Section 4.3.2.2, we describe three schemes for simulating missing data. In Sections

4.3.3 and 4.3.4, we investigate the power of two proposed models, including scenarios with missing data. Finally, we summarise the benefits of using simulation in Section 4.3.5.

#### 4.3.2.1 Methods for simulating VA paths

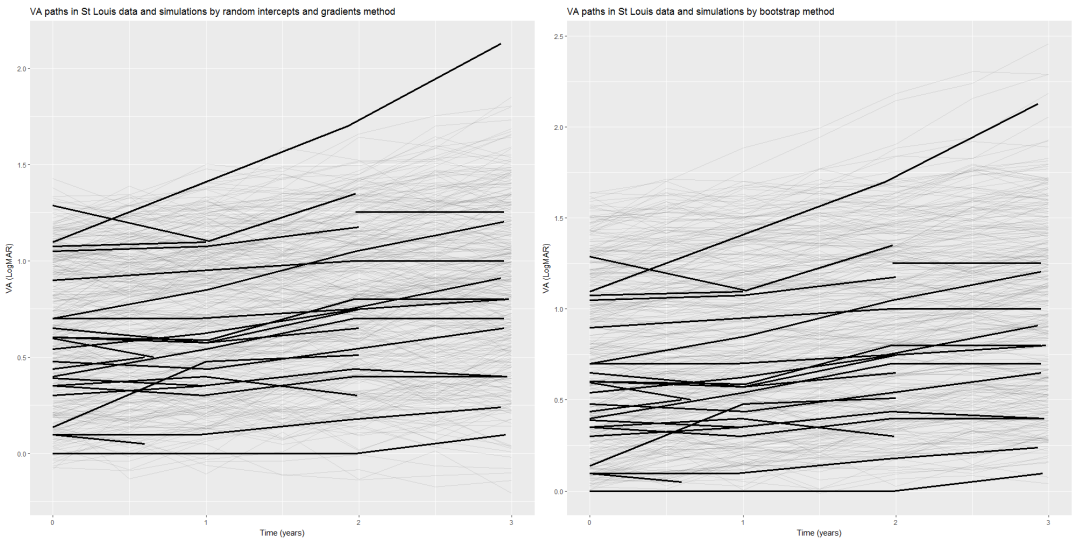
To make inference by simulation, we require a method of randomly sampling patient VA paths. We desire that these paths mirror the statistical characteristics of those observed in the St Louis dataset under the belief that the patients we recruit will be similar. We achieve this using two methods of path generation, each inspired by the random gradients model described in Section 4.2.1.2. We will refer to these as the *parametric method* and the *parametric bootstrap method*.

In both methods, we sample VA path starting values to be uniformly distributed on (0.0, 1.6), to reflect the patients we will recruit on trial. Patients with VA greater than 1.6 are ineligible.

The rate of increase in VA per annum is assumed to have a fixed component, common to all patients, and a patient-specific component. In each method, the fixed component for patients allocated to the control treatment is 0.07 LogMAR units, to match $\beta$, the estimated fixed effect with respect to time in (4.4). In the *parametric method*, the patient-specific gradient components, $b_i$, are randomly drawn from a $N(0, \sigma_b^2)$ distribution, with $\sigma_b = 0.071$ to match the estimates yielded by the random gradients model.

Figure 4.10a shows the skedasticity of the residuals from the REML random gradients model derived in Section 4.2.1.2. The raw residuals are shown as blue crosses and a loess smoother with confidence intervals is overlaid. The residuals are distributed about zero with some modest outliers. They could reasonably be described as homoskedastic for $t \in (0, 4)$ where the width of the confidence interval is approximately constant. The variability in year 5 appears to be slightly larger, but there are very few observations at this point. As such, random errors $\epsilon_{ij}$ for each VA assessment are assumed time-invariant and sampled from a $N(0, 0.074^2)$ distribution. Once again, these fixed values are chosen to match the result of fitting model (4.4) to the St Louis data.

With each of these components, the paths are calculated using $t_{VA}$ and (4.4).




(A) Parametric method (B) Bootstrap method

FIGURE 4.11: Three-hundred simulated VA series (grey lines) produced using the methods described. The overplotted black lines show the actual 26 St Louis patients.

TABLE 4.3: Correlation matrix of 300 3-year VA paths simulated using the parametric method.

| | $VA_0$ | $VA_{0.5}$ | $VA_1$ | $VA_{1.5}$ | $VA_2$ | $VA_{2.5}$ | $VA_3$ |
|---|---|---|---|---|---|---|---|
| $VA_0$ | 1 | | | | | | |
| $VA_{0.5}$ | 0.96 | 1 | | | | | |
| $VA_1$ | 0.95 | 0.96 | 1 | | | | |
| $VA_{1.5}$ | 0.93 | 0.94 | 0.96 | 1 | | | |
| $VA_2$ | 0.91 | 0.92 | 0.95 | 0.96 | 1 | | |
| $VA_{2.5}$ | 0.87 | 0.89 | 0.93 | 0.95 | 0.96 | 1 | |
| $VA_3$ | 0.86 | 0.87 | 0.91 | 0.94 | 0.96 | 0.97 | 1 |

Figure 4.11a shows 300 paths simulated using the parametric method. Overlaid in black are the paths observed in the St Louis cohort. The rapidly progressing individual is simple to identify. We see that the *parametric method* yields paths that are generally less extreme in high values than the St Louis paths. For instance, none of them progresses as rapidly as the patient in the St Louis cohort. Table 4.3 shows the serial correlation matrix for these paths. These broadly match the serial correlations observed in the St Louis cohort in Table 4.2.

A reasonable theoretical flaw of the parametric method is that the random gradients are not strictly normally distributed, as shown in Figure 4.9b. They are positively skewed by the single, rapidly-progressing individual. To investigate performance under non-normal random progression, in the *parametric bootstrap method* of

TABLE 4.4: Correlation matrix of 300 3-year VA paths simulated using the parametric bootstrap method.

| | $VA_0$ | $VA_{0.5}$ | $VA_1$ | $VA_{1.5}$ | $VA_2$ | $VA_{2.5}$ | $VA_3$ |
|---|---|---|---|---|---|---|---|
| $VA_0$ | 1 | | | | | | |
| $VA_{0.5}$ | 0.98 | 1 | | | | | |
| $VA_1$ | 0.98 | 0.98 | 1 | | | | |
| $VA_{1.5}$ | 0.97 | 0.98 | 0.98 | 1 | | | |
| $VA_2$ | 0.95 | 0.96 | 0.97 | 0.98 | 1 | | |
| $VA_{2.5}$ | 0.94 | 0.95 | 0.97 | 0.98 | 0.99 | 1 | |
| $VA_3$ | 0.92 | 0.94 | 0.95 | 0.97 | 0.98 | 0.99 | 1 |

path generation, we sample with replacement random gradients from the 26 values calculated on the St Louis dataset, depicted in Figure 4.9b. Likewise, we re-sample errors $\epsilon_{ij}$ from the model residuals in Figure 4.8. All other aspects remain the same as the parametric method and the components are combined using (4.4).

Three hundred paths by this method are shown in Figure 4.11b. We see that the simulations now yield paths as extreme as the St Louis cohort. Analysis of paths generated by this method will provide a valuable measure of the sensitivity of our statistical design to modest departures from the Gaussian assumptions. The serial correlations of these paths are shown in Table 4.4. These paths have generally greater serial correlation that those generated by the parametric method.

In both methods, paths for patients allocated to the *experimental* treatment are simulated in a similar way. The single difference is that the fixed effect gradient $\beta$ is assumed to be $0.07 - \lambda$, for some treatment effect $\lambda$. The random gradients are not adjusted, nor are the starting values or the measurement errors.

We expect some data to be missing on trial. Having described our way of simulating repeated measures data, in the next section we describe ways to obscure some data to estimate the power of our statistical test when data coverage is less than 100%.

#### 4.3.2.2 Missing data

Some data loss is practically unavoidable in TreatWolfram so it is conservative to factor this into the power estimation, especially when sample size is so severely constrained and the efficiency of the design is so critical. For $n$ assessments of $m$ patients, the full dataset should contain $mn$ data points.

Define $R_{ij}$ to be a data presence indicator variable that takes the value 1 if the outcome $y_{ij}$ is observed, else 0 if it is missing. Data are said to be Missing Completely At Random (MCAR) if $P(R_{ij} = 1)$ is constant over patients and time-horizons. This is not likely in a multi-year longitudinal analysis where later observations are naturally more likely to be missing for a variety of reasons, e.g. people move home. Data are said to be Missing At Random (MAR) if $P(R_{ij} = 1)$ is a function of the value or presence of previous observations or contemporaneous covariates. Critically, data-missingness is independent of the current outcome, $y_{ij}$ under MAR. If $P(R_{ij} = 1)$ is a function of $y_{ij}$, the data are said to be Missing Not At Random (MNAR).

Mixed effects models assume that data is MAR. Analysing MCAR or MAR data using mixed effects models does not result in bias but does lead to a loss of precision compared to an analysis of the complete dataset. Notably, analysing MNAR data using mixed effects models *results in bias* and a loss of precision. A distinct complication is that "distinguishing between MAR and MNAR is not trivial and relies on fundamentally untestable assumptions"[30]. We revisit this in the Discussion.

We investigate three methods for simulating missing data, illustrated in Figure 4.12. In the method depicted in 4.12a (that we will refer to as *missingness 1*), a number of series are assumed to be completely missing. All other series are fully observed. This naturally leads to the interpretation that it is the patients, rather than the observations, that go missing. The percentage of missing patients is equal to the percentage of missing data points. This method maximises the number of completely observed series.

Under missingness 2 in 4.12b, all data points are missing with equal probability, unaffected by whether other data are available for that patient.

Under missingness 3 in 4.12c, patient discontinuation points are randomly sampled iteratively until a threshold amount of missing information has been achieved. Once discontinued, a patient yields no further data. Some patients provide full data series, some provide no data, and some provide truncated series. To simulate this method, we used the following algorithm:

`While target level of data loss is not yet reached:`

- `Select a patient at random;`

| Patient | t0 | t1 | t2 | t3 | t4 |
|---|---|---|---|---|---|
| 1 | | | | | |
| 2 | | | | | |
| 3 | | | | | |
| 4 | | | | | |
| 5 | | | | | |
| 6 | | | | | |

(A) Whole series are missing

| Patient | t0 | t1 | t2 | t3 | t4 |
|---|---|---|---|---|---|
| 1 | | | | | |
| 2 | | | | | |
| 3 | | | | | |
| 4 | | | | | |
| 5 | | | | | |
| 6 | | | | | |

(B) Points are missing completely at random

| Patient | t0 | t1 | t2 | t3 | t4 |
|---|---|---|---|---|---|
| 1 | | | | | |
| 2 | | | | | |
| 3 | | | | | |
| 4 | | | | | |
| 5 | | | | | |
| 6 | | | | | |

(C) Discontinuation points are randomly sampled

FIGURE 4.12: Three methods of simulating missing outcome data. The orange cells represent missing observations.

- `Select a time point` $t^*$`, at random;`
- `Remove all observations for that patient with` $t \geq t^*$`;`
- `Loop.`

Under missingness 2, responses are MCAR. Under missingness 1 and 3, responses are MAR because missingness is dependent on the presence of the trailing observation: once an observation in a series is missing, no subsequent observations are made.

### 4.3.3 Power of the random intercepts model

In the TreatWolfram setting with experimental and control treatment arms, the random intercepts model generalises to

$$y_{ij} = \alpha + a_i + (\beta + \gamma z_i)t_{ij} + e_{ij} \quad e_{ij} \sim N(0, \sigma^2), \quad a_i \sim N(0, \sigma_a^2) \tag{4.7}$$

where $z_i = 1$ if patient $i$ is allocated to the experimental treatment, else 0. The parameter $\gamma$ estimates the mean adjustment in annual progression in VA associated with receiving the experimental drug compared to control. The presence of a treatment effect is assessed by testing the null hypothesis $H_0 : \gamma = 0$ against the alternative $H_A : \gamma \neq 0$. The other parameters maintain the roles previously described in Section 4.2.1.2.

Our test of $H_0$ entails a test of a fixed effect. As advised by Pinheiro & Bates[73, p. 87-90], we test the marginal significance of $\gamma$ not by likelihood ratio test, which is "sometimes quite badly anticonservative" but by the conditional $t$-test statistics provided in the standard table of regression output. We use a significance level of 5% so that $p$-values $< 0.05$ lead to rejection of the null hypothesis. This test is two-sided although only values of $\gamma < 0$ indicate efficacy, i.e. a reduction in progression. In a randomised controlled trial, a difference would be interpreted as having been caused by the difference in treatments.

The power of this analysis at various sample sizes with patients allocated equally between the arms is shown in Figure 4.13. We see that 70 patients provides approximately 88% power when no data is lost, and at least 80% power when up to 15%

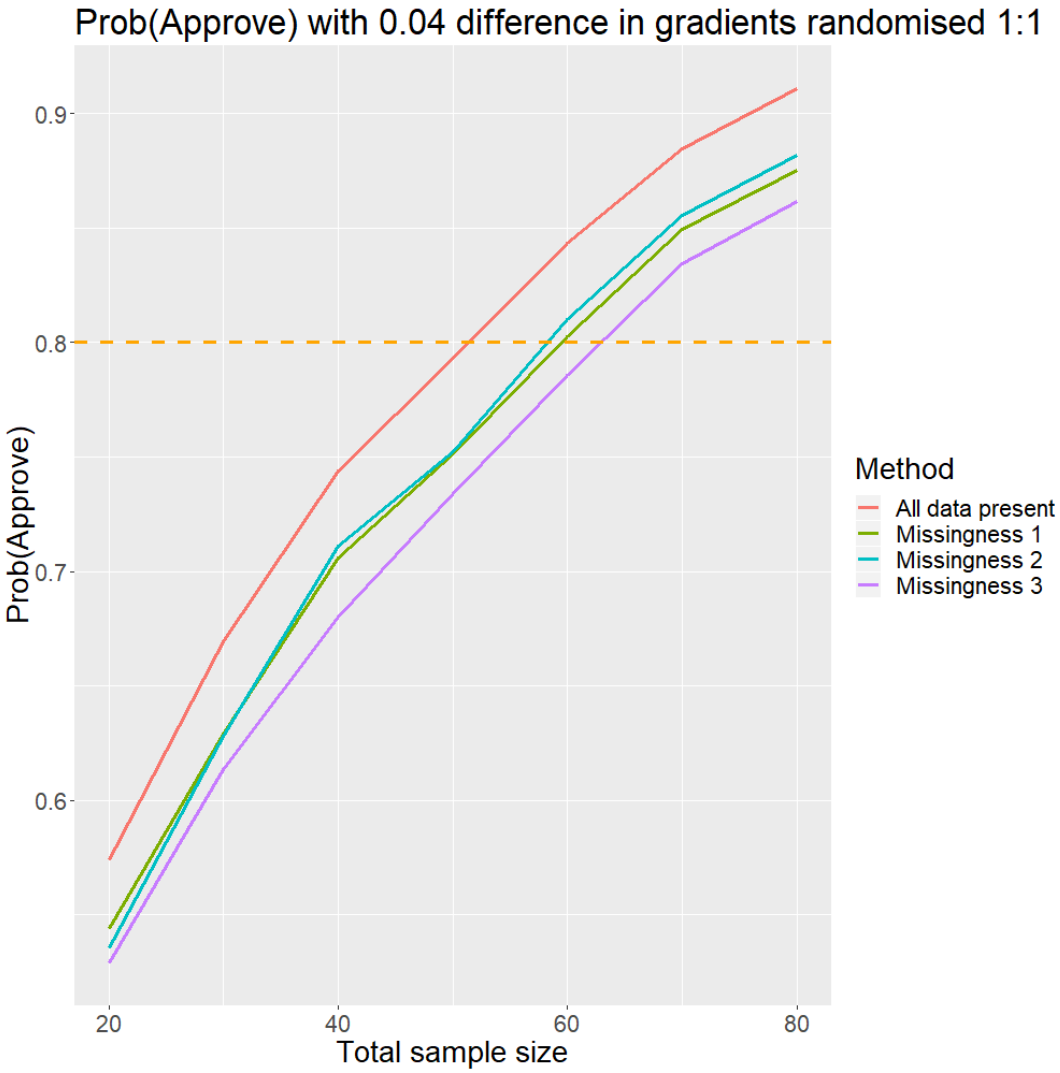
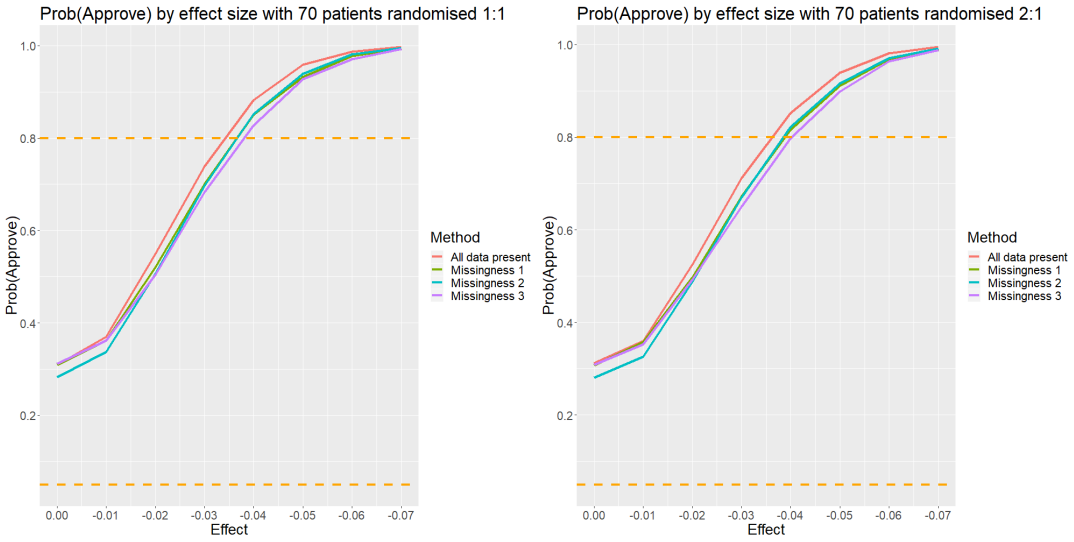



FIGURE 4.13: Power using the random intercepts model to detect a 0.04 unit decrease in the annual rate of increase in VA, with all data observed and under the three methods for removing 15% of outcomes described in Figure 4.12. 10,000 trial iterations were simulated in each scenario. VA outcomes were sampled using the parametric method.

| Power | Diggle *et al.* | Random intercepts model |
|---|---|---|
| 80% | 50 | 51 |
| 85% | 58 | 61 |
| 90% | 68 | 76 |

TABLE 4.5: Total sample size required with equal sized arms to detect a difference of 0.04 LogMAR units per annum. The Diggle *et al.* estimates are derived in the text. The estimates for the random intercepts model are interpolated from Figure 4.13.




(A) 1:1 randomisation

(B) 2:1 randomisation

FIGURE 4.14: Power at various effect sizes using the random intercepts model, with all data observed and under the three methods for removing 15% of outcomes. 10,000 trial iterations were simulated in each scenario. VA outcomes were sampled using the parametric method.

of data is lost, irrespective the pattern of missingness. Table 4.5 shows that when no data is lost, power estimated by the simulation method is close to that implied by the calculations using Diggle *et al.*'s method in Section 4.3.1.

The overwhelming problem with this model is that the type I error is vastly inflated, as demonstrated in Figure 4.14. Even when the true treatment effect is zero, there is approximately 33% probability of incorrectly approving the treatment. Approval probabilities are a few percent lower under 2:1 randomisation. This is because the model misinterprets *any* chance imbalance in patient-specific gradients as treatment effect. This is further demonstration that random gradients are necessary in our analysis model.

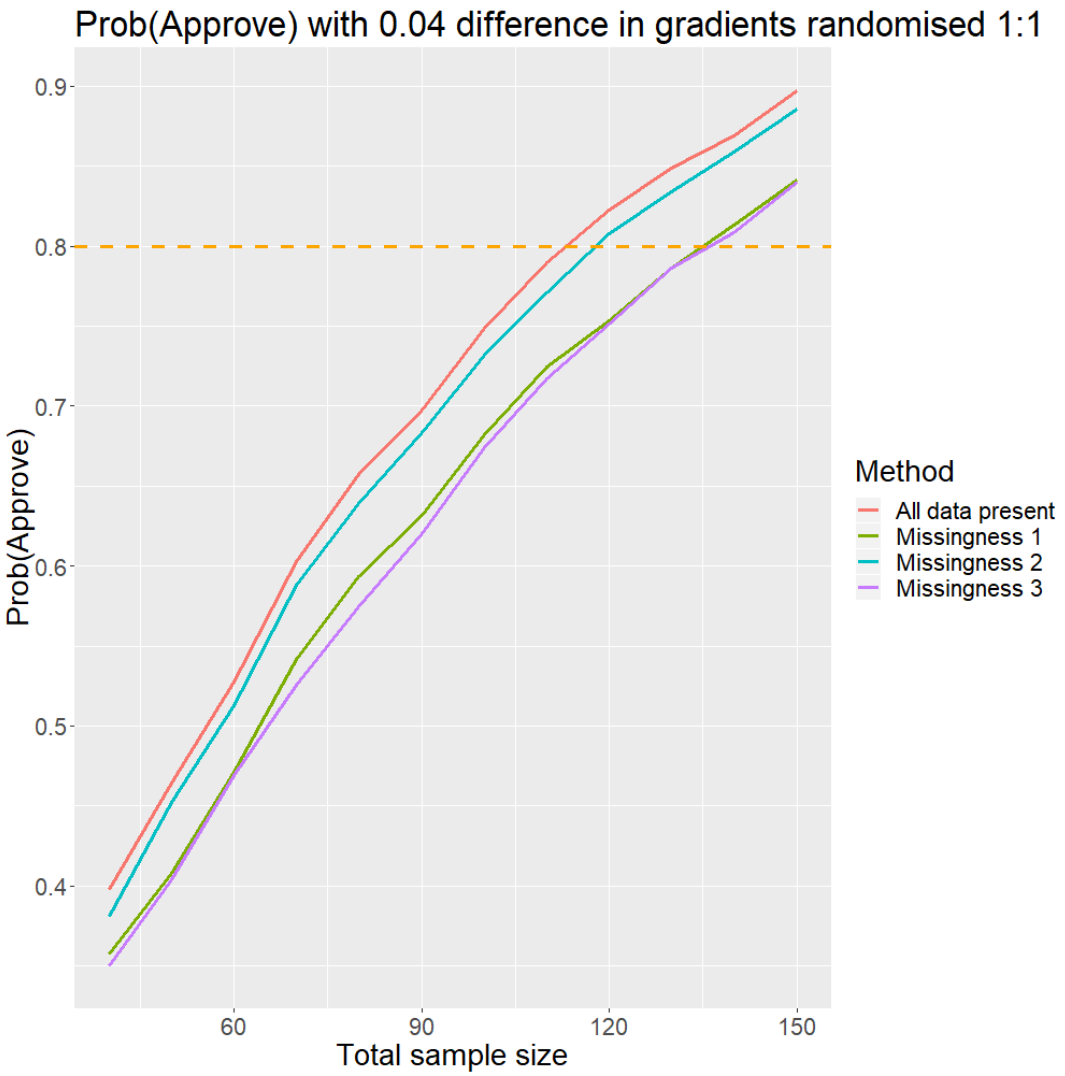


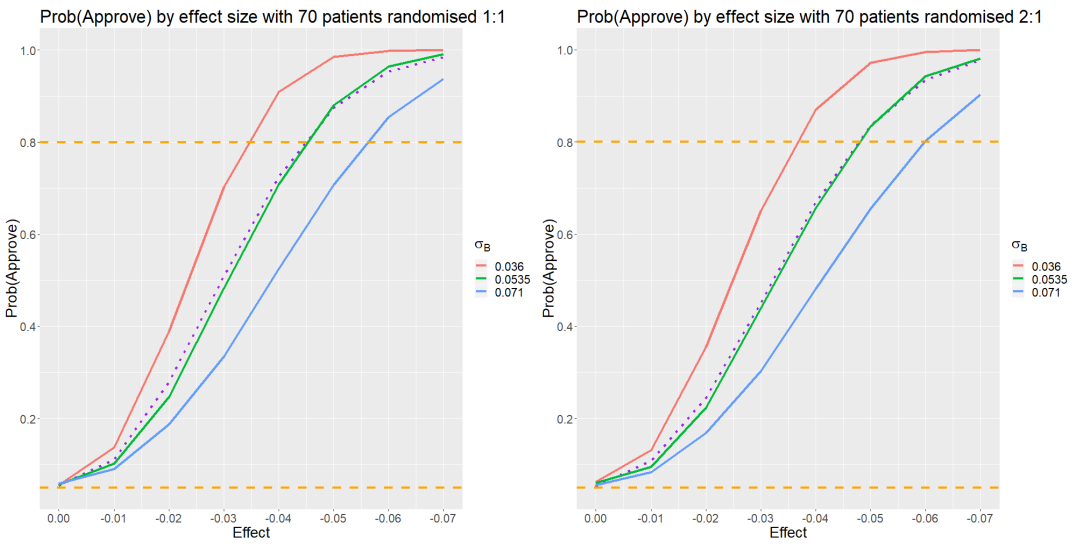

FIGURE 4.15: Power to detect with the random gradients model a 0.04 unit decrease in the annual rate of increase in VA, with all data observed and under the three methods for removing 15% of outcomes. 10,000 trial iterations were simulated in each scenario. VA outcomes were sampled using the parametric method.

### 4.3.4 Power of the random gradients model

We demonstrated that the random gradients mixed effects model fits the St Louis data better than the random intercepts model. As before, with two treatment arms, the model generalises to

$$y_{ij} = \alpha + a_i + (\beta + b_i + \gamma z_i)t_{ij} + e_{ij} \tag{4.8}$$

where $b_i \sim N(0, \sigma_b^2)$ are the random gradients with respect to time and all other parameters retain their previous definitions.




(A) 1:1 randomisation (B) 2:1 randomisation

FIGURE 4.16: Power at various effect sizes using the random gradients model, and different values for $\sigma_B^2$. We assume that 15% of data is lost under missingness method 3. 10,000 trial iterations were simulated in each scenario. VA outcomes were sampled using the parametric method (solid lines) and parametric bootstrap method (dotted purple line).

The power of this proposed analysis using 1:1 randomisation is shown in Figure 4.15. The random gradients model requires much larger sample sizes to achieve the same level of power as the random intercepts model. To achieve 80% power when no data is lost, we would expect to require about 110 patients, a material increase on the random intercepts model.

It is noteworthy that under the random gradients model, power is similar under missingness methods 1 and 3. It is intuitive to think that a regression model would be adept at dealing with data missing completely at random as in missingness method 2 because the model simply interpolates using the points on either side. Figure 4.15 shows that losing sequences of points is more detrimental to power in this scenario.

Figure 4.16 shows that the type I error rate is under control, as required. It also shows that 70 patients randomised 2:1 in favour of valproate yields roughly 80% power to detect a 0.06 LogMAR unit treatment effect when 15% of data is missing by method 3 and $\sigma_b = 0.071$. Under 1:1 randomisation, the same power would be available to detect an effect size of approximately 0.056, showing a modest performance penalty to using non-equal arms. Figure 4.16 also makes plain the sensitivity

of power to $\sigma_b$. The red line shows estimated power when $\sigma_b = 0.036$. This is the value estimated by the random gradients model fit to the St Louis data when the entire series for the rapidly-progressing patient is removed, i.e. the standard deviation approximately halves. We see that 80% power is now estimated to be achieved at a treatment effect between 0.035 and 0.040 LogMAR units with 2:1 randomisation.

The green series in Figure 4.16 shows power when $\sigma_b = 0.0535$, being the value halfway between the two extreme values. Notably (and unexpectedly), this series is very close to the line depicting power estimated using the random gradients model on outcomes simulated by the parametric bootstrap method (dotted purple line). This makes plain the uncertainty surrounding power and the role played by the variability of the patient-specific gradients. If rapid patient-specific progression is relatively common, we could expect our true, notional power curve to resemble the blue line with $\sigma_b = 0.071$. We highlighted when developing the random gradients model the large positive outlier in the random gradients, and how this value overstates $\sigma_b$. We have seen that the distribution of patient gradients is not normally distributed with mean 0 and standard deviation of 0.071. Instead, it is more like a combination of two distributions: a normal distribution with mean 0 and standard deviation of 0.036; and a single heterogeneous case. Thus, we expect the simulations using bootstrapped outcomes to be more indicative of what we may see on trial. The best outcome we could reasonably expect is represented by the red line where there is no rapid progression.

There is one additional qualitative explanatory factor that is exhibited by our single rapidly progressing patient, and this variable will be used on trial to maintain statistical efficiency. We expand on this in the Discussion.

### 4.3.5 Benefits of simulation

In previous sections, we described a method for simulating VA paths that statistically resemble those observed in the St Louis cohort. We also defined three methods for simulating data missingness. We then used these methods with mixed effects model analysis, and used computer simulation to infer statistical performance of our clinical trial design. The resulting power estimates reflect the implications of reasonably punitive data loss. It was important for us to understand the effect of data loss

and the various patterns of missingness that may manifest in our scenario where the size of feasible recruitment cohort is so severely constrained. Similarly, it was important for us to be able to measure the effect of non-equal randomisation. Simulation afforded us the flexibility to measure the simultaneous effect of these different complications. Furthermore, we learned from simulation that, whilst power under the random intercepts model was close to that suggested by the Diggle method, power under the random gradients model is lower because of the extra source of variability. We conducted this with the eventual goal of testing the effect of a treatment on a visual acuity outcome. A literature review follows in the next section of methods for estimating required sample size in trials that assess visual acuity.

## 4.4 Literature Review

We seek to demonstrate the novelty of our method for prospectively calculating the required sample size of a longitudinal analysis of visual acuity by simulation, including the effect of various patterns of data missingness. For brevity, we will refer to this as "our method" in the remainder of this section.

Appendix B.1 describes how a literature search of "visual acuity sample size trial" yielded ninety manuscripts for review, summarised in Table 4.6.

Primarily, manuscripts concerned trials, but there were many pertaining to reviews and cohort studies as well. Despite the search terms, it was relatively common that studies did not identify visual acuity as an outcome.

Despite being a continuous outcome measure, it was quite common that sample size estimation was conducted using a dichotomisation, e.g. defining response to be change from baseline of at least $x$. Testing differences by $t$-test was relatively common, as were ANOVA and ANCOVA (or their repeated measures analogues). Reporting of sample size methodology was frequently quite vague, making it impossible to identify exactly what method was used. In seven instances, it was possible to determine only that the researchers had not used simulation or repeated measures. These have been enveloped under the category *Other* in Table 4.6.

Methods using simulation or repeated measures were relatively rare, with only five manuscripts identifying either and only one manuscript identifying both.

| | | Manuscripts |
|---|---|---|
| **Manuscript type** | | |
| Clinical trial | | 57 |
| | Randomised groups | 47 |
| | Non-randomised groups | 6 |
| | Single arm | 2 |
| | Design not clear | 2 |
| Review | | 16 |
| Cohort study | | 11 |
| Statistical methodology | | 4 |
| Survey | | 2 |
| **Identified VA as outcome** | | |
| Yes | | 68 |
| No | | 16 |
| Not applicable | | 6 |
| **Sample size method** | | |
| Given | | 39 |
| | Dichotomised response | 10 |
| | $t$-test | 9 |
| | Simulation or repeated measures | 5 |
| | Other | 15 |
| Not given | | 51 |
| **Adjust sample size for missingness** | | |
| Yes | | 10 |
| | Linear inflation | 10 |
| No, or not applicable | | 80 |
| Overall | | 90 |

TABLE 4.6: Summary of manuscripts examined in literature review of the approach used in TreatWolfram to estimate statistical performance by simulation of repeated measures VA whilst incorporating missing data.

Data missingness was also mentioned very infrequently. Only ten manuscripts mentioned adjustment of the sample size to account for missing data. Each of these appeared to use a basic method of linearly inflating the calculated sample size. None gave evidence of having considered further how the data might be missing.

Of particular note were four manuscripts that described using mixed effects models for analysis, or simulation for calculating a sample size. Lambertus *et al.*[56] performed a retrospective cohort study in patients with Stargardt disease, a form of macular degeneration. Instead of modelling the level of visual acuity as we seek to do, they analysed the time to degeneration to a given threshold. Linear mixed models were used to analyse other variables. They used simulation to estimate sample size for an outcome other than visual acuity. They did not include loss-to-follow-up as a factor when estimating sample size.

Wiley *et al.*[106] present a randomised crossover trial of bevacizumab vs ranibizumab in patients with diabetic macular oedema. They analyse mean changes in visual acuity using linear mixed effects models. However, they did not model the manner in which longitudinal observations may be missing.

Lam *et al.*[55] presented the natural history of patients with a type of hereditary optic neuropathy to design a trial of gene therapy. They used mixed effects models to analyse outcomes but calculated sample size for a non-controlled study. They do not consider the effect of missing data.

Finally, Yeh *et al.*[109] present a study of the effects of acupressure and multimedia on the visual health of school children in Taiwan. They used a longitudinal method to calculate sample size using the "G Power" software, but no further details of the method are given. They apparently adjusted their sample size by 20% in anticipation of drop-out, but further details are not given.

In summary, we found evidence that other researchers have used mixed effects models to analyse visual outcomes, and have used simulation to calculate sample size. We found little evidence of complexity in simulating data missingness. Overall, we found no evidence that any researchers have previously used simulation to prospectively calculate a sample size for a longitudinal analysis of visual acuity whilst incorporating assumed patterns of data missingness. The manuscripts that came closest to our proposal are those of Wiley *et al.*[106] and Yeh *et al.*[109].

The reasons for focusing on visual acuity are clear but we believe that this serves as an example of methodologies typically used in studies with repeated numerical measures as outcomes. The results of the review for visual acuity are likely to be generalisable to other types of numerical outcome measured over time.

## 4.5 Discussion

Throughout this chapter, we have demonstrated statistical power to detect a treatment effect using the threshold 0.04 units p.a. LogMAR. The misidentification of a line of five letters on an ETDRS chart adds 0.1 to the LogMAR score, so that a score of 0.04 equates to two letters. Under a treatment that reduces annual average progression by 0.04 units, we would expect a patient to be able to read an extra line and an extra letter (i.e. six letters) after three years, compared to if they had not received the treatment. This makes clear the value of 0.04 units as a treatment effect. Trials that have used visual acuity as a primary outcome have sought larger differences in diseases where greater sample sizes are possible. In this chronic setting where sight progressively diminishes, any positive treatment effect would be of value because it lengthens the time a patient has vision. Clinical acceptability is generally a trade-off between efficacy and toxicity. Sodium valproate is not without side effects so we expect to see some adverse reactions in patients. We would potentially be interested in the treatment if it was demonstrated to be associated with a mean annual effect of 0.035 units that is statistically unlikely to have arisen by chance, and the incidence of adverse reactions was low and events were generally manageable. We anticipate that patients would too. A project supplementary to the trial will analyse patient-reported outcomes with respect to efficacy and toxicity, and seek to clarify a patient-oriented threshold. The difference of 0.04 units LogMAR does not constitute a hard threshold for approving the drug.

The foundation of our simulation study has been the outcomes observed in the St Louis cohort. Wolfram syndrome is a monogenic condition commonly observed in siblings, as in the original description by Wolfram & Wagener[107]. Other unobserved genetic or environmental traits could dictate that the outcomes we observe in the European trial differ from those in the American cohort study. The possibility

that European or British patients progress more slowly on average would be detrimental to our notional power. Nevertheless, the use of randomisation in the trial will promote a fair comparison.

We have made frequent reference to the rapidly progressing patient in the St Louis cohort and the difficulties that this introduced into the analysis. We have resisted the temptation to simply remove the patient because we may observe progression of this ilk in TreatWolfram. There was 1-in-26 in the St Louis cohort so there could be several in our larger cohort. Our model should be able to analyse outcomes from patients like these and provide robust inference. Additional explanatory information pertaining to rapid progressors exists, however.

The highest LogMAR value provided by EDTRS charts is 1.98, where only a single letter is correctly identified. How then, do we have values greater than 2 in Figure 4.2? Ophthalmologists have developed methods to ascribe so-called "off-chart" LogMAR scores to those who fail to read a single letter[57]. If a patient can correctly count fingers held up by the ophthalmologist, they are given a LogMAR score of 2.0. If they can correctly identify the presence of hand waving, they score 2.3. If they can correctly perceive the presence of light, they score 2.6.

These methods were used by the ophthalmologist assessing the St Louis patients and have been used in a published RCT e.g. [52, 53]. They provide pragmatic information on outcomes when patients can no longer be measured by the desired tool. Simply removing these points would understate the average disease progression. However, using off-chart outcomes presents a challenge for analysis because they introduce a discontinuity on an otherwise continuous scale. How do we know that progression assessed by off-chart methods belongs in the same distribution as that assessed on-chart? If we know which VA measures have been recorded using off-chart methods, we will be able to analyse progression under both on-chart and off-chart regimes. For instance, if off-chart measurements become commonplace, a simple method could analyse the on-chart and off-chart subsets separately, estimating the average progression whilst allowing for random patient-specific perturbations in intercepts and gradients in the manner we have demonstrated in this chapter. However, separate models are unlikely to provide the most efficient analysis. We did not do this when modelling the St Louis data because of the very small

number of off-chart measurements. Leaving the patient with rapid progression in the analysis set and allowing shrinkage of regression parameters by attaching probability distributions seemed the most conservative solution.

We investigated age and initial VA as prognostic variables and found that they did not improve our multi-level models. However, in the trial dataset, those same covariates could be predictive of treatment effect. That is, the age or initial VA value of a patient may in part determine the efficacy of the treatment. We will consider this when specifying the statistical analysis plan.

We used computer simulation to gauge the combined effect of non-equal randomisation and missing data on statistical efficiency. Guo *et al.*[42] introduced software to estimate required sample size in parallel groups studies with repeated measures. The method they present applies to complete cases. They acknowledge that missing data is a distinct complication in repeated measures studies. Furthermore, they describe how "validated power and sample size methods exist only for a limited class of mixed models...are based on approximations, and make simple assumptions about the study design". They advocate computer simulations as a general method to obtain reliable sample size estimates when formulae are not available.

More generally, Lu *et al.*[60] introduce a framework for estimating sample sizes in repeated measures analyses with missing data. They assume "monotone" missingness for simplicity, akin to our method in Figure 4.12c. In contrast, our missingness method in Figure 4.12b contravenes their assumption that the number of data-points at any given time never exceeds that at each earlier time. Nevertheless, their method for estimating the inflation factors required to compensate for missing data is valuable for gaining insight into our scenario. Applying their method to the St Louis correlations in Table 4.2 and the expected data presence at each time given 70 patients and 15% missing data under the method shown in Figure 4.12c, yields inflation factor estimates of 1.05 when part-year correlations are linearly interpolated, and 1.07 when only year-end data are used. Thus, even though we expect 15% of data-points to be missing, the sample size need be inflated by less than 15% to compensate and regain power. This is driven by the high expected correlation values.

We have used mixed models but alternative analysis methodologies exist. If assessment times were uniform across all patients, analysis by ANOVA would be

possible. Even though we intend to collect outcome assessments at set times, we would be foolish to expect that they never differ from schedule. The St Louis dataset contains some assessments that were not conducted near an anniversary of the first visit. We prefer a method that allows the time variable to be continuous rather than categorical. This naturally suggests using ANCOVA. However, ANCOVA does not allow the specification of random effects and these have demonstrably improved our modelling of the St Louis outcomes. Furthermore, the multi-level model approach allows the specification of generalised variance-covariance structures. Although we have not needed to use those here, they could well become necessary in the proposal below.

The experimental unit has hitherto been the individual: we have analysed the mean of left and right eye visual acuities for each individual through time. We also described an isolated incident where a measurement was only available in one eye and the care we had to take when imputing the effective "mean" value. An approach to abrogate this complication, and potentially increase statistical efficiency, is to analyse the eyes separately. This is possible in our setting because symptoms affect both eyes. Here, the experimental unit would be eyes rather than individuals, and we would have approximately double the number of series to analyse. However, care would have to be taken to handle the association between eyes. Figure 4.3 shows that contemporaneous left- and right-eye measurements are highly correlated. For this reason, double the number of experimental units would yield less than double the effective sample size. Multilevel models are flexible enough to handle this. Firstly, eyes are nested within individuals. For example, each patient may take their own visual acuity intercept to reflect their general baseline quality of vision, and also eye-specific intercepts to reflect the chance baseline disparity between the two eyes. A similar specification will be possible for random gradients. Multilevel models can specify these types of nested effects: observations are nested within eye through time, and eyes are nested within individuals, all subject to overarching population-level effects. Furthermore, they facilitate covariance structures to model heteroskedasticity and serial correlation in residuals, should they arise.

This potential lift to efficiency and power would be very welcome. It would provide some insurance against a potential decrease in power that would arise if our

repeated measures have a lower serial correlation than those in the St Louis dataset. Analysing series within eye within patient would actually be expected to slightly increase serial correlation as a source of variability, eyes within patient, would have been removed.

A more prosaic option to increase the amount of information is to assess outcomes every three months, for example, rather than every six months, effectively increasing the size of $t_{VA}$ from 7 to 13. Primarily, this would help the model more accurately estimate $\sigma$ and increase power. However, variability across patients is also important. In our random gradients model, more accurate estimation of $\sigma_a$ and $\sigma_b$ requires more patients, not just more assessments of the patients.

In Section 4.3.2.2 we considered missing data. It is customary under some analysis methods to impute missing values. Imputation unavoidably makes assumptions about the distribution of unobserved values. Last Observation Carried Forward (LOCF) is a popular method, where missing values are assumed to take the value that was last observed for a patient. This would be highly inappropriate in TreatWolfram because symptoms demonstrate a tendency to deteriorate through time, as demonstrated by Figure 4.2. LOCF is sometimes justified as a conservative assumption. However, assuming no change here assumes the symptom ceases to deteriorate and this is clearly an optimistic assumption. An analysis using LOCF would show bias in favour of the trial arm with the most drop-out.

Another method is multiple imputation (MI), where likely values for missing observations are calculated from observed outcomes and covariates. However, this requires a model for imputation and if the imputation model is the same as the analysis model, the inference of the analysis using MI will match the inference from fitting a mixed effects model to just the observed data [104]. Thus to improve on our scenario in TreatWolfram, we would need to incorporate auxiliary information like treatment compliance or alternative outcomes. One of the considerable strengths of mixed effects models is that they do not mandate imputation; the model is simply fit to the available data. There is no requirement for us to impute, so we do not.

We cited the impossibility of distinguishing MAR from MNAR. An accepted pragmatic solution is to use sensitivity analyses to distinguish how the inferences of an analysis change if the assumption of MAR is violated [105]. We propose to

do this. Furthermore, we will ask patients why outcomes are not reported. If, for instance, patients stop attending visual acuity clinics because their vision has deteriorated to the extent that they do not feel comfortable travelling, then the assumption of MAR would clearly be violated. In circumstances like this, we would analyse the outcomes using a method that incorporates informative dropout like pattern mixture models.

In Appendix B, we describe our search strategy for identifying papers concerning clinical trials that use a visual acuity outcome and describe a method of calculating sample size. We also compare the methods we used to arrive at a feasible randomised clinical trial design in this rare disease to the framework on designing randomised trials in small populations by Parmar *et al.*[72]. They recommend steps in three sequential categories: *increase what is feasible*; *explore commonly-considered approaches to reducing sample size*; and *explore less common approaches to reducing sample size*. We found high fidelity with the steps we took and the recommendations in their first two categories.

## 4.6 Conclusion

We have succeeded in specifying a defensible clinical trial design with conventional statistical error rates in an ultra-rare disease. Instrumental to this was our ability to select outcomes amenable to repeated measures analysis. We selected parameters for hypothesis testing and simulated frequentist operating performance of our design using the St Louis dataset provided by Prof. Hershey. Using a standard pre- and post-treatment analysis of two groups would have required a sample size exceeding the disease population in the UK. Analysing repeated measures solved this problem. In our ultra-rare disease setting, this boost to efficiency was the critical factor that made the described trial feasible. Simulation allowed us to make informed judgements on preferential allocation to the experimental arm that has proved so important to patients and their carers.

## Chapter 5

# A Phase II Stratified Medicine Trial with Efficacy and Toxicity Outcomes and Predictive Variables

**Background:** PePS2 is a phase II trial of the efficacy and safety of pembrolizumab in performance status 2 non-small-cell lung cancer patients. Previous studies have shown that the chances of clinical response are correlated with baseline covariates, particularly the extent to which PD-L1 is expressed by the cells in a tumour biopsy. There are few clinical trial designs that test co-primary efficacy and toxicity outcomes in phase II, and fewer still that allow the incorporation of stratifying baseline variables.

**Notable methods in this chapter:** The design of Thall, Nguyen and Estey is one such design but it has been scarcely used in actual trials. Furthermore, their model incorporates terms to conduct a dose-finding study. This aspect is not required in PePS2 because an effective and safe dose has already been identified in a closely-related population. We introduce a novel simplification of their design suitable for use in phase II that focuses on testing efficacy and toxicity at a fixed dose whilst adjusting for baseline cohort effects.

**The implications on efficiency:** The method allows sharing of information across cohorts. Using a total of 60 patients to test the treatment in six distinct cohorts, we can expect error rates typical of those used in phase II trials. Our simulations show it is far more efficient than a method that analyses cohorts individually.

## 5.1 Introduction

There is a relative dearth of phase II clinical trial designs that incorporate predictive patient covariates to assess efficacy and toxicity. Thall *et al.*[89] introduced a family of methods that perform dose-finding trials guided by binary efficacy and toxicity outcomes whilst accounting for baseline patient covariates. This enables dose recommendations tailored to individual patients. Our motivation is PePS2, a phase II trial of pembrolizumab in non-small cell lung cancer patients of performance status 2. PePS2 is not a dose-finding trial. Instead, it seeks to estimate the probabilities of efficacy and toxicity at a dose of pembrolizumab previously demonstrated to be safe and effective in performance status 0 and 1 patients. In this chapter we introduce a novel implementation of a simplified version of Thall *et al.*'s method. We remove the dose-finding components but retain aspects to study co-primary efficacy and toxicity outcomes that are associated with baseline covariates. In Section 5.2, we describe the trial setting and review existing clinical trial designs for analysing both efficacy and toxicity. In Section 5.3, we present our proposed alteration to Thall *et al.*'s model. In Section 5.4, we simulate performance in PePS2 and compare it to that of simple Bayesian beta-binomial conjugate models. We discuss some limitations of the model and potential further development in Section 5.5. Finally in Section 5.6 we finish with some conclusions.

## 5.2 Background

### 5.2.1 The PePS2 Trial

PePS2 is a phase II trial of pembrolizumab in non-small cell lung cancer (NSCLC) patients with Eastern Cooperative Oncology Group (ECOG) performance status 2 (PS2). A patient with PS2 is ambulatory and capable of taking care of themselves but typically too ill to work. Critically, it is doubtful that a PS2 patient could tolerate the toxic side effects of chemotherapy.

The joint primary outcomes of the trial are (i) *toxicity*, defined as the occurrence of

a treatment-related dose delay or treatment discontinuation due to adverse event related to pembrolizumab; and (ii) *efficacy*, defined as the occurrence of a complete response (CR), partial response (PR), or stable disease (SD), without prior progressive disease (PD) as measured by RECIST v1.1[35], at or after the second scheduled CT scan that is detailed in the protocol to occur at 18 weeks. For instance, if the second scheduled scan is missed, potentially for reasons of illness, but a subsequent scan confirms absence of progression with respect to baseline, then this will be treated as efficacy. The primary objective of the trial is to learn if the treatment is associated with sufficient efficacy with acceptably low toxicity to approve for further research in performance status 2 patients.

Pembrolizumab inhibits the programmed cell death 1 (PD-1) receptor via the programmed death-ligand 1 (PD-L1) protein. In a phase I study with 495 patients, Garon *et al.*[37] showed pembrolizumab to be active and tolerable in performance status 0 & 1 patients. Overall, 19.4% of patients had an objective response (OR), defined as the occurrence of PR or CR, and 9.5% experienced an adverse event of grade 3 or higher. The rate of toxicity compares favourably to those typically seen in advanced NSCLC patients using chemotherapy [10, 79]. With few treatment options available for PS2 patients, it seemed worthwhile to investigate if similar rates of efficacy and toxicity could be achieved in a PS2 population and thus we hope to show that pembrolizumab is a viable treatment in this specific patient population.

TABLE 5.1: Objective response rate (ORR), where OR = CR or PR, in PD-L1 score cohorts for the 204 patients in the validation sample of Garon, *et al.*[37] with evaluable PD-L1 status.

| Pretreated | PD-L1 Cohort | PD-L1 Criteria | n | ORR%, (95% CI) |
|---|---|---|---|---|
| Yes | Low | PD-L1 $< 1\%$ | 22 | 9.1 (1.1, 29.2) |
| Yes | Medium | $1\% \leq$ PD-L1 $< 50\%$ | 77 | 15.6 (8.3, 25.6) |
| Yes | High | PD-L1 score $\geq 50\%$ | 57 | 43.9 (30.7, 57.6) |
| No | Low | PD-L1 score $< 1\%$ | 6 | 16.7 (0.4, 64.7) |
| No | Medium | $1\% \leq$ PD-L1 score $< 50\%$ | 26 | 19.2 (6.6, 39.4) |
| No | High | PD-L1 score $\geq 50\%$ | 16 | 50.0 (24.7, 75.3) |
| All | Low | PD-L1 score $< 1\%$ | 28 | 10.7 (2.3, 28.2) |
| All | Medium | $1\% \geq$ PD-L1 score $< 50\%$ | 103 | 16.5 (9.9, 25.1) |
| All | High | PD-L1 score $\geq 50\%$ | 73 | 45.2 (33.5, 57.3) |

Garon *et al.* introduce the PD-L1 proportion score biomarker, defined as the percentage of neoplastic cells with staining for membranous PD-L1, hitherto referred

to as *PD-L1 score*. In the nomenclature of Buyse *et al.*[20], they demonstrate PD-L1 to be a *valid* and *predictive* biomarker of pembrolizumab activity. They use the *hold-out method* to identify subgroups based on PD-L1 thresholds, using distinct training and validation subsets of their overall trial population. Efficacy outcomes for the 204 patients in their validation group are shown in Table 5.1. Objective responses are observed in all cohorts and the probability of response generally increases with PD-L1 score.

Based on this information, we expect PD-L1 score to be predictive of response in our PS2 population. Additionally, we expect a mix of patients that have and have not previously received treatment for their cancer. In the Garon trial, 24.8% of treatment-naive (TN) patients achieved a response, whereas only 18.0% did in the pre-treated (PT) patients. A chi-squared test of association between pretreatedness and response yielded a $p$-value of 0.166. A patient with recently diagnosed disease such that no therapy has yet been given could be quite different to a patient that has received previous lines and progressed. Pretreatedness represents a potentially small but important effect that should be considered when testing the treatment.

The PePS2 chief investigator, Gary Middleton (GM), and the lead biostatistician, Lucinda Billingham (LB), proposed a single arm phase II trial that investigates drug in the six cohorts formed by jointly stratifying by: the three Garon PD-L1 classifications; and the PT or TN statuses. Each patient in PePS2 will belong to one of these six cohorts. The trial aims to recruit over one year.

Being a phase II trial, there is strong motivation to deliver findings quickly to inform potential phase III research in a timely manner. It is felt that recruitment in the region of 60 patients within one year would be feasible but that recruitment materially higher would be prohibitive. Given the relative dearth of treatments for PS2 patients and the prior evidence of activity and tolerability in all NSCLC subgroups, GM felt it important to offer a trial aimed at all-comers and not limit the target population by our covariates. Pembrolizumab has not been investigated in PS2 patients so the clinical scenario requires a trial design that tests efficacy and toxicity. Given the evidence that PD-L1 score and previous treatment status are associated with the likelihood of response to this drug in NSCLC patients, it is highly desirable to use a clinical trial design that incorporates these potentially predictive variables to tailor

the treatment approval decision in specific patient subgroups. In the next section, we describe our search for a clinical trial design that achieves these objectives.

### 5.2.2 Review of Competing Trial Designs

The trial statistician, Kristian Brock (KB) sought a clinical trial design that admits explanatory variables to study joint primary outcomes efficacy and toxicity at phase II. The results of our search are summarised in Table 5.2.

| Reference | Design | Co-primary | Covariates | Phase II |
|---|---|---|---|---|
| Braun[12] | BCRM | Yes | No | No |
| Ivanova[46] | | Yes | No | No |
| Zhang *et al.*[112] | TriCRM | Yes | No | No |
| Wang & Day[99] | | Yes | No | No |
| Thall *et al.*[27, 92, 93] | EffTox | Yes | No | No |
| Ghebretinsae *et al.*[40] | | Yes | No | Yes |
| Cook & Farewell [28] | | Yes | No | Yes |
| Brutti *et al.*[18] | | Yes | No | Yes |
| Bouckaert & Mouchart[11] | | Yes | No | Yes |
| Bryant & Day[19] | | Yes | No | Yes |
| Conaway & Petroni[25, 26] | | Yes | No | Yes |
| Thall, Simon & Estey[87, 90] | | Yes | No | Yes |
| Thall & Sung[91] | | Yes | No | Yes |
| Wathen *et al.*[103] | | No | Yes | Yes |
| Thall, Nguyen & Estey [89] | TNE | Yes | Yes | No |

TABLE 5.2: Results of literature review seeking a design for PePS2. Covariates reflects inclusion of baseline data without further adaptation. Phase II reflects original intent.

Using PubMed, KB searched for publications under the MeSH major topic 'clinical trials' that are categorised with the MeSH Terms 'Drug-Related Side Effects and Adverse Reactions' and 'Models, Statistical'. Efficacy was not made explicit in our search because establishing efficacy is such a common motivation for trials. We expected the presence of a toxicity outcome to be a more effective discriminator. On 5-Aug-2015, this query returned 67 documents whose collective focus was primarily statistical clinical trial methodology in scenarios where toxicity is a key outcome.

Forty-eight of the papers were discarded because they focused on a univariate outcome: forty-four focused primarily on toxicity alone and a further four focused on efficacy alone. Four papers were reviews or advisory in nature and did not contain specific model proposals. One paper was discarded because it was in Danish with no English translation.

This left fourteen papers for further consideration. Naturally, given the subject matter, these papers concerned a preponderance of dose-finding and early phase trials. With cytotoxic treatments, dose-finding has typically sought to find the maximum tolerable dose under the assumption that efficacy and toxicity increase in lock step as dose is increased. In so-called cytostatic treatments, disease may be controlled without reducing the overall tumour burden and the probability of efficacy may not be an increasing function of dose. As such, in cytostatic treatments, efficacy and toxicity can be jointly scrutinised to find the optimal dose rather than just the maximal dose. The growth of targeted therapies and immunotherapies is associated with a growing focus on methods that jointly model efficacy and toxicity for dose-finding purposes. These have been already reviewed in Chapters 2 and 3.

Eight of the papers in our search describe dose-finding methods for cytostatic treatments. Although these works detail designs that address a different trial objective (i.e., finding a dose), they are pertinent to our problem because they potentially use probability models that could be redeployed for our purposes. We consider those briefly now.

Braun[12] introduced a bivariate extension of the Continual Reassessment Method (CRM) to two competing outcomes, toxicity and disease progression, where the two events are associated. CRM itself was originally published by O'Quigley *et al.*[69] with the purpose of conducting dose-finding trials under the cytotoxic assumption. Ivanova[46] presented a rule-based up-and-down design that seeks to maximise the number of subjects allocated in the neighbourhood of the optimal dose. Zhang, Mandrekar and Sargent[63, 112] introduced TriCRM, another extension of CRM that considers the ordinal trinary outcome: no response and no serious toxicity; efficacy without serious toxicity; and toxicity so serious that it precludes efficacy. Wang & Day[99] present a method where response and toxicity outcomes occur according to bivariate log-normally distributed patient thresholds. They allocate the next dose to maximise patient-oriented expected utility. Finally, Thall *et al.*[27, 92, 93] present EffTox, the Bayesian adaptive dose-finding design that is the focus of Chapter 2. Generally in dose-finding models, as with EffTox, dose (or transformed dose) is used as the sole explanatory variable that determines outcome probabilities. This provides opportunities to use other explanatory variables in a non-dose-finding

setting.

Five papers present models for efficacy and toxicity in a non-dose-finding setting. Ghebretinsae *et al.*[40] present a method for modelling non-gaussian continuous outcomes from assay data. This is not applicable to our scenario because our outcomes are not continuous. In the single arm setting, Cook & Farewell [28] present a sequential design to analyse correlated bivariate efficacy and toxicity events, accounting for multiple analyses over time. Jin[50] presents a two-stage method accounting for the trade-off between efficacy and toxicity. Brutti *et al.*[18] present a two-stage Bayesian method to compare the overall toxicity rate and the true efficacy-and-safety rate to pre-specified target thresholds. None of these methods explicitly include predictive variables, although that is not to say they could not be adapted to use them.

In the two-arm setting, Bouckaert & Mouchart[11] present a model to analyse a two arm randomised controlled trial from the view that trial outcomes can be attributed to therapeutic effects and toxic effects. They also do not explicitly consider predictive variables but their model uses binary variables to denote arm membership so it is sensible to conclude that this specification could be generalised to include arbitrary explanatory variables.

Finally, our PubMed search returned Bryant & Day[19]. This is perhaps the best known and widely used phase II trial design for studying efficacy and toxicity. Theirs is a two-stage method that offers a chance to reject a treatment for being inactive or excessively toxic at an interim stage. The design takes threshold values for the probabilities of efficacy and toxicity that are acceptable and unacceptable and returns the minimum number of efficacy events and maximum number of toxicity events that should be observed to approve the treatment for further study. For given levels of statistical significance and power, the threshold event counts define the optimal trial of the competing outcomes of efficacy and toxicity. Their method considers different levels of association between efficacy and toxicity events and chooses an optimal design. The design implicitly assumes that the patient population is homogeneous thus it does not use predictive variables.

Using a Bryant & Day optimal design to contrast efficacy rates of 10% and 30% and toxicity rates of 10% and 30%, with efficacy and toxicity significance of 10%

and overall power of 80%, requires the final analysis to use 27 patients. If we were to use this design in each PePS2 cohort, we would require $6 \times 27 = 162$ patients, an infeasibly high number. Even if we were to ignore the potentially important information in the pretreatment variable and analyse three PD-L1 cohorts, we would still require 81 patients using parallel Bryant & Day designs. Analysing the cohorts separately in this way is inefficient. At this juncture, our preference was for a model-based design that could increase power by incorporating predictive information.

Not included in our PubMed search but frequently cited in similar work is Conaway & Petroni[25, 26]. They present sequential designs for phase II trials with bivariate, associated activity and toxicity outcomes. In each case, their emphasis is on the development of stopping rules rather than the incorporation of predictive information.

To further supplement our search, we studied review articles of biomarker-guided clinical trial designs. Table 2 in Buyse *et al.*[20] lists the *targeted* (or selection) design (as used in the ToGA trial[6]) and *Bayesian adaptive* design (as used in the BATTLE trial[51], amongst others) as potential designs for validated, predictive biomarkers of an experimental treatment. These are multi-arm designs, randomly allocating patients to treatments, conditional on biomarker status. Neither of these designs analyse toxicity as a co-primary outcome, although naturally safety would be an important secondary outcome in trials that use either. Freidlin & Korn[36] review randomised designs that can be used to develop or validate biomarkers. Our setting is non-randomised and concerns studying the treatment modification effect of a biomarker that has already been validated in a closely related patient population. More recently, Antoniou *et al.*[5] described in detail the adaptive biomarker-guided clinical trial designs they encountered in a review that covered 171 papers and 14,436 candidate abstracts. None of the eight designs they describe explicitly incorporates a co-primary outcome.

We were also aware of other pertinent publications through knowledge of the field. Thall, Simon & Estey[87, 90] and Thall & Sung's[91] work on monitoring multiple outcomes (commonly, efficacy and toxicity) using Dirichlet-multinomial models and stopping boundaries in single arm phase II trials. These methods do not use predictive information. Wathen *et al.*[103] published a method that uses predictive patient data to study efficacy in patient subgroups, but their method does not study

toxicity.

Finally, Thall, Nguyen & Estey (TNE)[89] introduce an extension of EffTox[92] that adds baseline patient covariates to the analysis of co-primary efficacy and toxicity outcomes at different doses. Theirs is a Bayesian design that uses uninformative priors on dose-effects and informative priors justified by historic data on the covariate effects. The objective achieved by their design is to recommend a personal dose of an experimental agent that is estimated to offer sufficient probability of efficacy and acceptable probability of toxicity, after taking into consideration predictive baseline covariates.

In an example demonstrating their design in AML, TNE use age as a continuous covariate, and a three-level ordinal variable reflecting prognosis with respect to cytogenetic subtype. They conducted a search of previously untreated AML patients aged less than 60 that were then treated with chemotherapy at MD Anderson Cancer Center between January 2000 and December 2004. They found 693 patients treated with three general classes of chemotherapy and tabulated the frequencies of efficacy and toxicity using their covariates, age and cytogenetics. They demonstrated that efficacy decreases and toxicity increases with age. Similarly, they demonstrated that efficacy decreases and toxicity increases as cytogenetic category worsens from *good*, to *intermediate* and ultimately *poor*. In the marginal efficacy and toxicity models they used quadratic terms with respect to dose-level to handle non-linearity, and associated these using a Gaussian copula with probit link.

On 05-Dec-2017, we identified 16 manuscripts listed on PubMed that cite TNE[89]. Of these, 10 were further methodology papers, each concerned with dose-finding. None of these works sought to adapt the design for use in the typical phase II scenario of investigating efficacy and toxicity at a single dose. Three papers were methodological reviews, citing TNE as a potential method. Another was a systematic review of thrombolysis for acute ischaemic stroke that cited TNE but made no explicit reference to it in the main body. A manuscript by Konopleva *et al.*[54] uses TNE in a dose-finding study of PR104 in relapsed or refractory AML and acute lymphoblastic leukaemia (ALL). The final paper was an expert panel recommendation [32] on the diagnosis and management of AML. It referred to TNE simply as a

method that incorporated covariates when dose-finding in contrast to standard designs like 3+3. This literature search suggests that TNE's method has only been used in blood cancer and only for the purposes of dose-finding. We found no suggestion that the method had been adapted for the non-dose-finding context.

The Konopleva study[54] identified above used TNE in a dose-finding study of 17 AML and ALL patients. This study investigated doses of the hypoxia-activated prodrug PR104, ranging from 1.1 to 4 $g/m^2$. A further 8 patients were then treated at selected doses (i.e. not guided by the design), and a further 25 patients were used in an expansion phase. The manuscript mentions that "3 prognostic covariates" were used in the dose-finding study but does not explicitly define them. We sought to identify the covariates. An online supplement is referred to in the manuscript but was not available at the *Haematologica* website on 3-Apr-2018. We contacted the lead author by email but received no response.

PePS2 is not a dose-finding trial. Previous studies of pembrolizumab using collectively over 1,000 PS0/1 NSCLC patients[37, 44] showed that response and adverse event outcomes are not materially affected by dose changes in the range 2 mg/kg to 10 mg/kg. For this reason, subsequent trials of pembrolizumab, including PePS2, used a flat dose of 200mg not adjusted for weight.

We sought a design that: i) studied associated co-primary binary outcomes; ii) and admitted explanatory covariates; iii) at a single common dose. We resolved to remove the dose-finding elements of TNE and retain the model that uses covariates to study correlated co-primary outcomes and tailor the trial decision to each covariate-determined cohort. Of all the candidate designs that could be adapted to achieve these ends, we selected TNE for two reasons. The first was our familiarity with the underlying probability model having used EffTox in the Matchpoint trial, as described in Chapter 2. The second motivation was that TNE offers more than we require, and it is generally easier to simplify something by taking unnecessary elements away than it is to extend something by adding extra complexity. This is the focus of the next section.

## 5.3 A Design for Co-Primary Efficacy and Toxicity Outcomes and Covariates

In this section, we describe novel adaptations to the TNE design to arrive at a model that studies associated, co-primary probabilities of efficacy and toxicity of an experimental agent, adjusted for baseline predictive covariates. We refer to our phase II version of the TNE design as P2TNE. In the following section, we describe the probability model, retaining the elements to incorporate covariates but removing the elements that perform dose-finding tasks.

### 5.3.1 Probability Model in P2TNE

TNE present the marginal probability models of an experimental treatment

$$\text{logit}\, \pi_k(\tau, x, \boldsymbol{\theta}) = f_k(\tau, \boldsymbol{\alpha}_k) + \boldsymbol{\beta}_k x + \tau \boldsymbol{\gamma_k} x \tag{5.1}$$

for $k = E, T$ denoting efficacy and toxicity, respectively. Here, $\tau$ is the given dose; $x$ is a vector of covariates; $\boldsymbol{\theta}$ is the vector of model parameters to be estimated; the $f_k(\tau, \boldsymbol{\alpha}_k)$ characterise the dose effects; $\boldsymbol{\beta}_k$ is the vector of covariate effects; and $\boldsymbol{\gamma}_k$ is a vector of dose-covariate interactions. They also introduce analogous models for the events under historical treatments where covariate effects are present.

As with EffTox[92], let $\boldsymbol{Y} = (Y_E, Y_T)$ be indicators of binary efficacy and toxicity events. Let $\pi_{a,b}(\tau, x, \theta) = \Pr(Y_E = a, Y_T = b | \tau, x, \theta)$ for $a, b \in \{0, 1\}$. The authors associate the marginal probabilities of efficacy and toxicity in a joint model with association parameter $\psi$:

$$\pi_{a,b} = \pi_{a,b}(\pi_E, \pi_T, \psi) \tag{5.2}$$

One possibility for this joint model is that used in EffTox (2.4), sometimes referred to as the Gumbel model.

In P2TNE, we can simplify this if we have no motivation to investigate different doses by removing the terms that pertain to dose-effects.

In our model description for P2TNE, let $x$ denote the baseline covariate information for a given patient. The marginal probabilities of efficacy and toxicity are

estimated using the logit models:

$$\text{logit}\, \pi_E(x, \boldsymbol{\theta}) = g(x, \boldsymbol{\theta}) \tag{5.3}$$

and

$$\text{logit}\, \pi_T(x, \boldsymbol{\theta}) = h(x, \boldsymbol{\theta}) \tag{5.4}$$

where $\boldsymbol{\theta}$ is the vector of all parameters in the model. The exact specifications of $g$ and $h$ are left for the trialists to specify to reflect the perceived relationships of $x$ with the probabilities of efficacy and toxicity. Generally, as with all statistical models, $g$ and $h$ should be both plausible and parsimonious. We present our choices for the PePS2 trial in the next section.

Let patient $i$ have covariate vector $x_i$, and let $a_i = 1$ if they experience efficacy, else 0; and $b_i = 1$ if they experience toxicity, else 0. For trial data

$$\boldsymbol{X} = \{(x_1, a_1, b_1), ..., (x_n, a_n, b_n)\} \tag{5.5}$$

the aggregate likelihood function is

$$\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{X}) = \prod_{i=1}^{n} \pi_{a_i,b_i}(\pi_E(x_i, \boldsymbol{\theta}), \pi_T(x_i, \boldsymbol{\theta}), \psi) \tag{5.6}$$

where $\psi$ is a member of $\boldsymbol{\theta}$. Let $\boldsymbol{\theta}$ have prior distribution function $f(\boldsymbol{\theta})$. For patients with predictive variable vector $x$, the posterior expectation of the probability of efficacy under the treatment is

$$\mathbb{E}(\pi_E(x, \boldsymbol{\theta})|\boldsymbol{X}) = \frac{\int \pi_E(x, \boldsymbol{\theta}) f(\boldsymbol{\theta}) \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{X}) d\boldsymbol{\theta}}{\int f(\boldsymbol{\theta}) \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{X}) d\boldsymbol{\theta}} \tag{5.7}$$

and the posterior probability that the probability of efficacy exceeds $\pi_E^*$ is

$$\Pr(\pi_E(x, \boldsymbol{\theta}) > \pi_E^*|\boldsymbol{X}) = \frac{\int \mathbb{I}(\pi_E(x, \boldsymbol{\theta}) > \pi_E^*) f(\boldsymbol{\theta}) \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{X}) d\boldsymbol{\theta}}{\int f(\boldsymbol{\theta}) \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{X}) d\boldsymbol{\theta}} \tag{5.8}$$

where $\mathbb{I}(A)$ is again the indicator function.

Similarly, the posterior probability of toxicity is

$$\mathbb{E}(\pi_T(x, \boldsymbol{\theta})|\boldsymbol{X}) = \frac{\int \pi_T(x, \boldsymbol{\theta}) f(\boldsymbol{\theta}) \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{X}) d\boldsymbol{\theta}}{\int f(\boldsymbol{\theta}) \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{X}) d\boldsymbol{\theta}} \tag{5.9}$$

and the posterior probability that the probability of toxicity is less than $\pi_T^*$ is

$$\Pr(\pi_T(x, \boldsymbol{\theta}) < \pi_T^*|\boldsymbol{X}) = \frac{\int \mathbb{I}(\pi_T(x, \boldsymbol{\theta}) < \pi_T^*) f(\boldsymbol{\theta}) \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{X}) d\boldsymbol{\theta}}{\int f(\boldsymbol{\theta}) \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{X}) d\boldsymbol{\theta}} \tag{5.10}$$

The posterior expectation of the parameter vector is

$$\hat{\boldsymbol{\theta}} = \mathbb{E}(\boldsymbol{\theta}|\boldsymbol{X}) = \frac{\int \boldsymbol{\theta} f(\boldsymbol{\theta}) \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{X}) d\boldsymbol{\theta}}{\int f(\boldsymbol{\theta}) \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{X}) d\boldsymbol{\theta}} \tag{5.11}$$

The number of dimensions in the integrals (5.7) to (5.11) is equal to the number of elements in $\boldsymbol{\theta}$. The difficulty in solving such integrals increases with dimension, although modern Markov chain Monte Carlo (MCMC) methods like Stan[22] make it relatively simple to sample from the posterior distribution.

Further taking the lead from TNE[89] and Thall & Cook[92], we propose the treatment be approved in patients with predictive variable vector $x_i$ when it is sufficiently likely that the associated efficacy probability exceeds some minimum threshold, $\pi_E^*$, and toxicity probability is less than some maximum threshold, $\pi_T^*$. The acceptance criteria are:

$$\begin{aligned} \Pr(\pi_E(x_i, \theta) > \pi_E^*|\boldsymbol{X}) > p_E \\ \Pr(\pi_T(x_i, \theta) < \pi_T^*|\boldsymbol{X}) > p_T \end{aligned} \tag{5.12}$$

where $p_E$ and $p_T$ are determined using clinician input and simulation. Naturally, $\pi_E^*$, $\pi_T^*$, $p_E$ and $p_T$ can vary by patient cohort. We could, for example, set the efficacy hurdle lower in PT patients if a dearth of feasible alternatives dictates that a lower efficacy hurdle is nevertheless clinically relevant.

The tests in (5.12) can be invoked at any time during the trial with different values for $\pi_E^*, \pi_T^*, p_E$ and $p_T$, making it simple to incorporate interim analyses in a clinical trial, exploiting the flexibility offered by Bayesian cumulative learning. We revisit this in the Discussion.

#### 5.3.1.1 Practical Steps for Implementation

Trialists should assess the operating performance of a design like P2TNE in theoretical scenarios using computer simulation. At the very least, we conduct simulations to estimate the probability that a design will incorrectly approve a poor treatment (similar to the notion of significance in frequentist trial designs) and correctly approve a good treatment (essentially, statistical power). Simulated trials are conducted by randomly sampling outcomes for notional patients and invoking the acceptance decision determined by (5.12) at the final (and potentially also interim) stages. In general, prior to simulating performance, we:

1. Specify forms for the marginal efficacy and toxicity models (5.3) and (5.4).
2. Specify a form for the joint model.
3. Specify $f(\boldsymbol{\theta})$, the prior distribution for $\boldsymbol{\theta}$.
4. Specify efficacy and toxicity thresholds, $\pi_E^*$, $\pi_T^*$ based on clinical rationale. These may vary by cohort or they may be common, as the clinical scenario dictates.

With this information, we simulate trial data sets, $\boldsymbol{X}_j$ for $j = 1, ..., J$ and infer the decision of the design on each. Values for $p_E$ and $p_T$ need not be specified before simulations are run. Instead, it is more flexible to record the value of (5.8) and (5.10) in simulated iterations for each distinct value of $x_i$. Then, we adjust the performance of the design by considering different values for $p_E$ and $p_T$, inferring the operating characteristics of the pair by invoking (5.12) on the simulated output.

In summary, the values for $\pi_E^*$, $\pi_T^*$ are based on clinical rationale and set at runtime. In contrast, the values of $p_E$ and $p_T$ need not be, so it is easier to tweak model operating characteristics by varying $p_E$ and $p_T$. We invoke this algorithm below.

### 5.3.2 P2TNE in PePS2

GM selected $\pi_E^* = 0.1$ and $\pi_T^* = 0.3$ for all cohorts because these represent the thresholds beyond which the treatment would be considered not sufficiently active or too toxic.

We define the predictive variables used in PePS2. Let patient $i$ have $x_{1i} = 1$ if they have been pretreated, else $x_{1i} = 0$. For the primary analysis, we will allocate patients to exactly one of the three PD-L1 groups presented in Table 5.1. Let patient $i$ have $x_{2i} = 1$ and $x_{3i} = 0$ if they belong to the Low PD-L1 cohort; $x_{2i} = 0$ and $x_{3i} = 1$ if they belong to the Medium PD-L1 cohort; and $x_{2i} = 0$ and $x_{3i} = 0$ if they belong to the High PD-L1 cohort. Thus, $x_{2i}$ and $x_{3i}$ are dummy variables that wholly determine membership to the three groups Low, Medium and High PD-L1 [1]. The cohorts and values for $x_i = (x_{1i}, x_{2i}, x_{3i})$ are shown in Table 5.3.

| Cohort | Treatment status | PD-L1 category | $x_i$ |
|---|---|---|---|
| 1 | Treatment naive | Low | (0,1,0) |
| 2 | Treatment naive | Medium | (0,0,1) |
| 3 | Treatment naive | High | (0,0,0) |
| 4 | Pretreated | Low | (1,1,0) |
| 5 | Pretreated | Medium | (1,0,1) |
| 6 | Pretreated | High | (1,0,0) |

TABLE 5.3: Cohorts used in the PePS2 trial. $x_i$ shows the covariate vector for each patient in that cohort.

Using these variables, we propose that the marginal probabilities of efficacy and toxicity be described by logit-models so that, for a patient with predictive data $x_i$:

$$\begin{aligned} \text{logit}\, \pi_E(x_i, \boldsymbol{\theta}) &= \alpha + \beta x_{1i} + \gamma x_{2i} + \zeta x_{3i} \\ \text{logit}\, \pi_T(x_i, \boldsymbol{\theta}) &= \lambda \end{aligned} \tag{5.13}$$

and associate $\pi_E$ and $\pi_T$ using the Gumbel model (2.4 ). In the PePS2 protocol, the toxicity outcome includes occurrence of adverse events that lead to treatment cessation. If patients discontinue treatment, it naturally hinders their ability to gain therapeutic benefit from the treatment and makes response less likely. In contrast, those patients that stay on treatment give themselves the best opportunity for response if the treatment does have a therapeutic effect. As such, it is sensible to facilitate that efficacy and toxicity would be associated. Including $\psi$ to model the association, our parameter vector $\boldsymbol{\theta} = (\alpha, \beta, \gamma, \zeta, \lambda, \psi)$ has six elements.

Under (5.13), the expected probability of efficacy is different for each distinct arrangement of $x_i$. Furthermore, the log-odds of efficacy for TN patients in the three PD-L1 categories are assumed to be a common linear shift of those for PT patients

[1] Using three dummy variables *and* an intercept would yield a singular design matrix

in the same PD-L1 cohorts, determined by $\beta$. Figure 5.1 shows the log-odds of objective response with uncertainty bars by cohort for the validation subgroup of the Garon study. The model we propose effectively assumes that the equivalent lines in our study using our definition of efficacy will be piecewise parallel. We see that this assumption is broadly supported by the small amount of data reported by Garon. The assumption is perhaps modestly violated in the Low cohort, but the estimates are fairly uncertain, particularly in the low PD-L1 groups. A more complicated alternative specification could remove the parallelism assumption by incorporating interaction terms for PD-L1 cohort and pretreatment status. This alternative model would require two extra parameters to handle interactions between $x_{1i}$ and $x_{2i}$, and $x_{1i}$ and $x_{3i}$, respectively, a topic we develop in the next chapter.

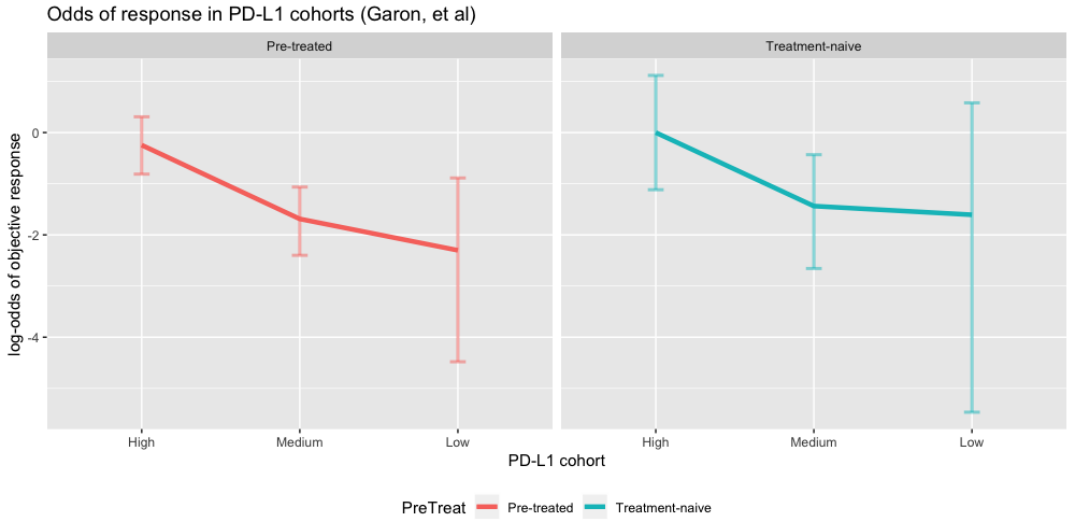



FIGURE 5.1: Log-odds of objective response and 95% uncertainty interval, by cohort of the validation sample ($n = 204$) of the Garon *et al.* study.

Figure 5.1 suggests that parallel lines is not an implausible working model. With accrual anticipated to reach 60 patients, the prospect of using six parameters instead of eight is attractive, so we proceed with the six-parameter model. Nevertheless, with only three data-points per line, we remain mindful to not reach conclusions unmerited by the limited data. In Section 5.4 we simulate performance in a scenario where there is a modest interaction between PD-L1 and pretreatment status, similar to that depicted in Figure 5.1.

In contrast, the probability of toxicity is assumed constant across all cohorts.

Garon, *et al.*[37] do not report in the main paper or supplementary appendix any difference in toxicity in the different PD-L1 or pretreatment groups. Furthermore, no heterogeneity with respect to adverse events is reported by Herbst *et al*[44] in the phase III study of pembrolizumab in NSCLC. However, another topic of the next chapter is an embellishment of our P2TNE model that allows toxicity to vary by PD-L1 and pre-treatedness.

### 5.3.3 Priors in PePS2

We specify normal priors for the elements of $\boldsymbol{\theta}$. In their AML example, TNE use informative priors on parameters that represent covariate effects on outcomes, reflecting historic published data. In contrast, they use uninformative priors on the dose-effects. Their objective is to identify the optimal dose of an experimental agent whilst controlling for baseline heterogeneity. They have deployed uninformative priors on the parameters that are the primary subject of investigation, and informative priors on those that they concede to be "nuisance parameters...for the purposes of dose-finding"[89]. They describe an algorithm for establishing hyperparameters of normal priors. To establish prior means, they elicit expected event rates for at least two dose-levels and solve for the expected values of the dose coefficients $\boldsymbol{\alpha}$ in (5.1) after assuming that dose-covariate parameters $\boldsymbol{\gamma}$ have expected value zero. They also describe a potential algorithm for establishing prior variances. This method controls the effective sample size (ESS) by equating the first two moments of the $\pi_k(\tau, \boldsymbol{Z}, \boldsymbol{\theta})$ to beta distributions and exploiting the fact that the ESS of a beta$(a, b)$ distribution is $a + b$. They advocate that each of the $a + b$ should be small to reflect the limited prior knowledge about dose-effects.

Our primary objective is to estimate the efficacy and toxicity of a treatment *in each distinct cohort of patients*. Thus, the covariates in our setting are rather more than nuisances because they determine these groups. Having observed data in the Garon[37] and Herbst[44] trials, we naturally anticipate that those with higher PD-L1 scores will more likely achieve our efficacy outcome. Likewise, we anticipate that PT patients will be modestly less likely to have the efficacy event.

#### 5.3.3.1 Informative priors

We could develop an analogue of TNE's method described above to establish priors in our setting. Instead, however, we consider different priors based on the event rates they generate. Gelman *et al.*[39] encourage us to think "generatively" in our selection of priors, explaining that "a prior is *generative* if the prior predictive distribution generates only data deemed consistent with our understanding of the problem." In this spirit, we are motivated to select informative priors on model parameters so that the expected efficacy rate in high PD-L1 patients exceeds that of medium patients, which in-turn exceeds that of low patients; and that TN patients are slightly more likely to experience efficacy than PT patients. Furthermore, we may reflect in our priors, information not reported in the Garon and Herbst studies, like the logical expectation that previously treated patients who are further down the disease pathway, may be more likely to experience toxicity because they are more vulnerable than TN patients. We discriminate the priors not by their notional ESS but by the event rates they generate and the associated uncertainty intervals they provide.

Our efficacy outcome in PePS2, repeated here for convenience, is *the occurrence of CR, PR or SD, without prior PD, assessed by RECIST v1.1[35], at or after the second scheduled CT scan expected to occur at 18 weeks*. This outcome is essentially a dichotomisation of progression-free survival (PFS), an outcome used in many cancer trials. We can inform our expectations of our efficacy outcome by analysing PFS reported by PD-L1 and pretreatment status by Garon *et al.*[37]. They do not explicitly report PFS rates at 18 weeks, but in their Figure 3, they show Kaplan-Meier curves that allow us to interpolate values. Their plot of PFS in PT patients includes approximately 300 patients at risk at time = 0, so we expect that the estimates will be relatively precise. In the PT subset, PFS at 18 weeks is approximately 36% in low, 37% in medium and 55% in high PD-L1 patients, as shown in Table 5.4. We expect outcomes in the PS2 population to be similar to PS0/1 but the lower overall level of health suggests considering a modest penalty. To identify prior mean efficacy probabilities in PT patients that reflect a modest penalty, we subtract 15% from the PD-L1-matched PS0/1 groups, as demonstrated in column B of Table 5.4.

| Column | A | B | C | D | E |
|---|---|---|---|---|---|
| Derivation | Garon *et al.*[37] | A - 15 | B + 5 | B $\pm$2 | C $\pm$2 |
| Interpretation | PT (PS 0/1) | PT (PS2) | TN (PS2) | PT target | TN target |
| Low PD-L1 | 36 | 21 | 26 | 19-23 | 24-28 |
| Medium PD-L1 | 37 | 22 | 27 | 20-24 | 25-29 |
| High PD-L1 | 55 | 40 | 45 | 38-42 | 43-47 |

TABLE 5.4: Derivation of prior mean efficacy rates to motivate informative parameter priors. We start in column A with the PS0/1 efficacy rate observed by Garon *et al.* in PT patients. We subtract 15% from this in column B to reflect that we expect PS2 PT patients to be weaker and have worse chances of efficacy than PS0/1 PT patients. In column C we add 5% to B to reflect that we expect PS2 TN patients to do slightly better than PS2 PT patients. In columns D and E, we create target ranges for the expected efficacy rates by adding and subtracting 2% to B and C, knowing that we will not obtain parameter priors that yield efficacy means that exactly match B and C. Parameters in Table 5.5 were chosen so that the expected efficacy rates fell in the ranges in D and E. Numbers are %.

Garon *et al.*'s subset of TN patients is much smaller, however, with only 62 subjects at time = 0 split between the three subgroups. The Kaplan-Meier plot for TN patients shows large decreases in PFS for each single event, with large changes in the survival curve being associated with small changes in time. This prohibits reading off accurate values. We can see from the summary statistics they report that TN patients generally do slightly better than PD-L1-matched PT patients. To identify suitable prior efficacy estimates, we instead estimate the TN efficacies to be a modest improvement on the PT efficacies. We increase the PD-L1-matched PT estimates by 5%, as shown in column C. These coarse adjustments are intended only to identify plausible expected values. Neighbouring efficacy rates will be facilitated by the uncertainty parameters in our normal parameter priors.

TABLE 5.5: Informative normal prior distributions on $\boldsymbol{\theta}$.

| | $\mu$ | $\sigma^2$ |
|---|---|---|
| $\alpha$ | -0.3 | 4 |
| $\beta$ | -0.7 | 4 |
| $\gamma$ | -2.0 | 4 |
| $\zeta$ | -2.0 | 4 |
| $\lambda$ | -2.2 | 2.9 |
| $\psi$ | 0 | 1 |

Having identified candidate prior mean efficacy rates in each cohort, we add and subtract 2% from each to generate a target range, as shown in columns D and E. We then select by trial-and-error hyperparameters that achieve means in the target

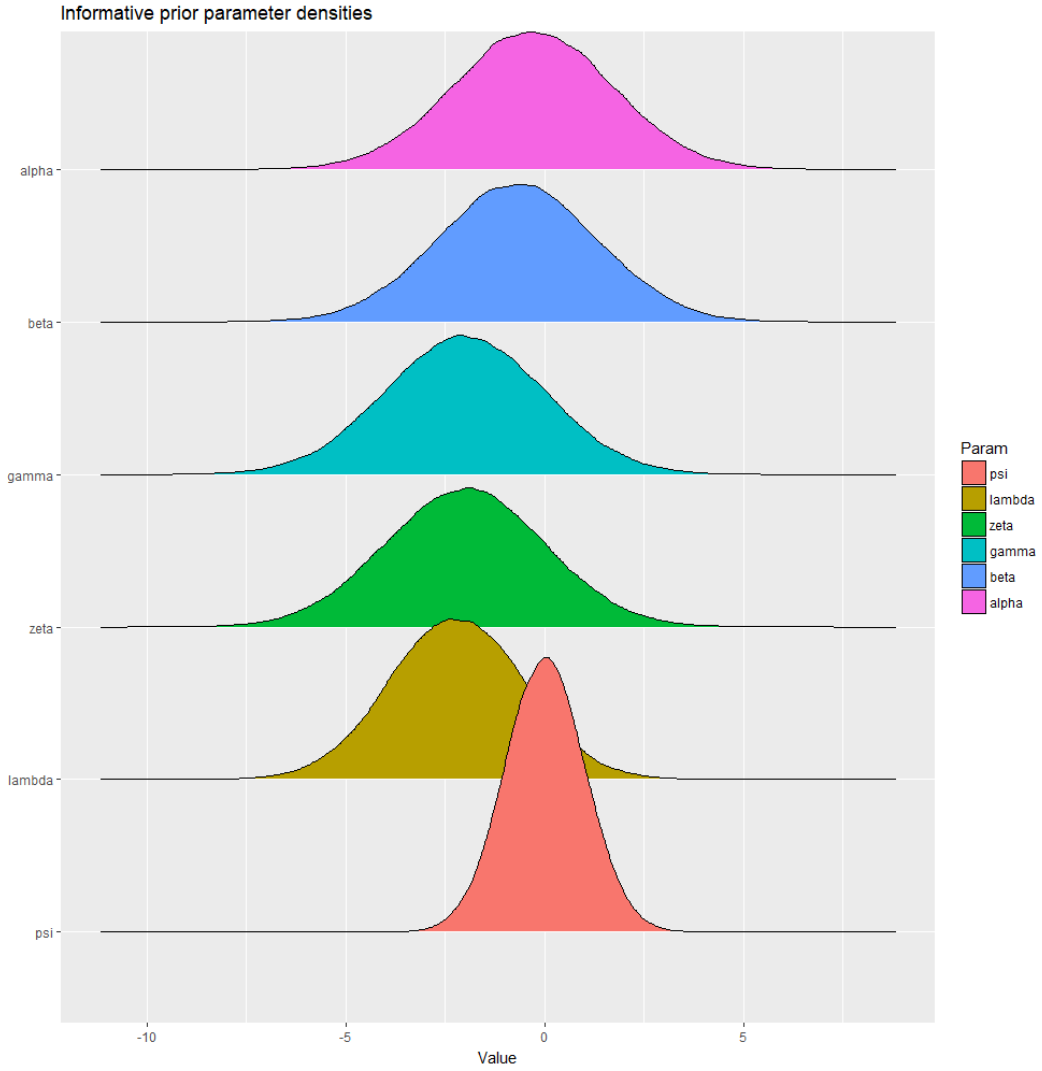



FIGURE 5.2: Informative prior distributions on the parameters.

range. We have described this logic to give transparent justification to the prior mean efficacy probabilities generated by our informative priors. We feel transparency is important here given how contentious informative priors are.

Our chosen hyperparameters for our informative priors are shown in Table 5.5 and the prior parameter densities are shown in Figure 5.2. The event rates they generate with credible intervals (CI) are shown in Table 5.6. The upper case L and U adorn the 90% CI and lower case letters adorn the 50% CI. The variability parameters were selected to yield 50% and 90% prior predictive CIs that felt appropriate. For instance, efficacy probabilities over 40% are possible in TN low and medium PD-L1 patients, but not particularly likely. Likewise, we would not want to rule out an efficacy probability in high PD-L1 patients that exceeds 70%, but it is much more

likely to be lower.

| Treatment status | PD-L1 | EffL | Effl | Eff | Effu | EffU |
|---|---|---|---|---|---|---|
| TN | Low | 0.00 | 0.01 | 0.24 | 0.40 | 0.91 |
| TN | Med | 0.00 | 0.01 | 0.25 | 0.40 | 0.91 |
| TN | High | 0.03 | 0.16 | 0.46 | 0.74 | 0.95 |
| PT | Low | 0.00 | 0.00 | 0.22 | 0.34 | 0.94 |
| PT | Med | 0.00 | 0.00 | 0.22 | 0.34 | 0.94 |
| PT | High | 0.00 | 0.05 | 0.38 | 0.72 | 0.97 |
| Treatment status | PD-L1 | ToxL | Toxl | Tox | Toxu | ToxU |
| TN/PT | Low-High | 0.01 | 0.03 | 0.18 | 0.26 | 0.64 |

TABLE 5.6: Credible intervals for events rates drawn from the prior predictive distribution of the informative priors in Table 5.5. Eff and Tox show the probability of efficacy and toxicity, respectively. Lower-case l and u show the central 50% credible interval and upper-case L and U show the central 90% credible interval.

We now consider priors on our toxicity outcome, again repeated for convenience: *treatment delay or discontinuation caused by an adverse event related to pembrolizumab*. Garon *et al.*[37] refer to only a solitary incident of treatment discontinuation after an infusion reaction. Although they do not report treatment delays arising from pembrolizumab-emergent adverse events, it is likely that they occurred. In stark contrast, Herbst *et al*[44] report that 34 / 345 (9.9%) of patients allocated to pembrolizumab 2 mg/kg and 32 / 346 (9.2%) of patients allocated to pembrolizumab 10 mg/kg discontinued due to an adverse event. These events may have manifested primarily because of treatment or disease. We are not told which but, for the purposes of forming prior beliefs on our toxicity outcome, it is sensible to assume that some are down to disease and some down to treatment. Once again, treatment delays are not explicitly described or quantified but will almost certainly have occurred. With three-weekly administrations in sick patients, treatment-related delays could be very common. For example, it is highly likely that treatment delays will occur in patients that do not eventually discontinue. Our priors should reflect this level of ignorance. We expect a toxicity rate approximately twice that reported by Herbst but admit that the rate could plausibly be higher. Our hyperparameter choices for the sole parameter in our toxicity model are shown in Table 5.5 and the generated toxicity rates and CIs, assumed the same in each cohort, are shown in Table 5.6.

For illustration, the predictive event densities under our informative priors are

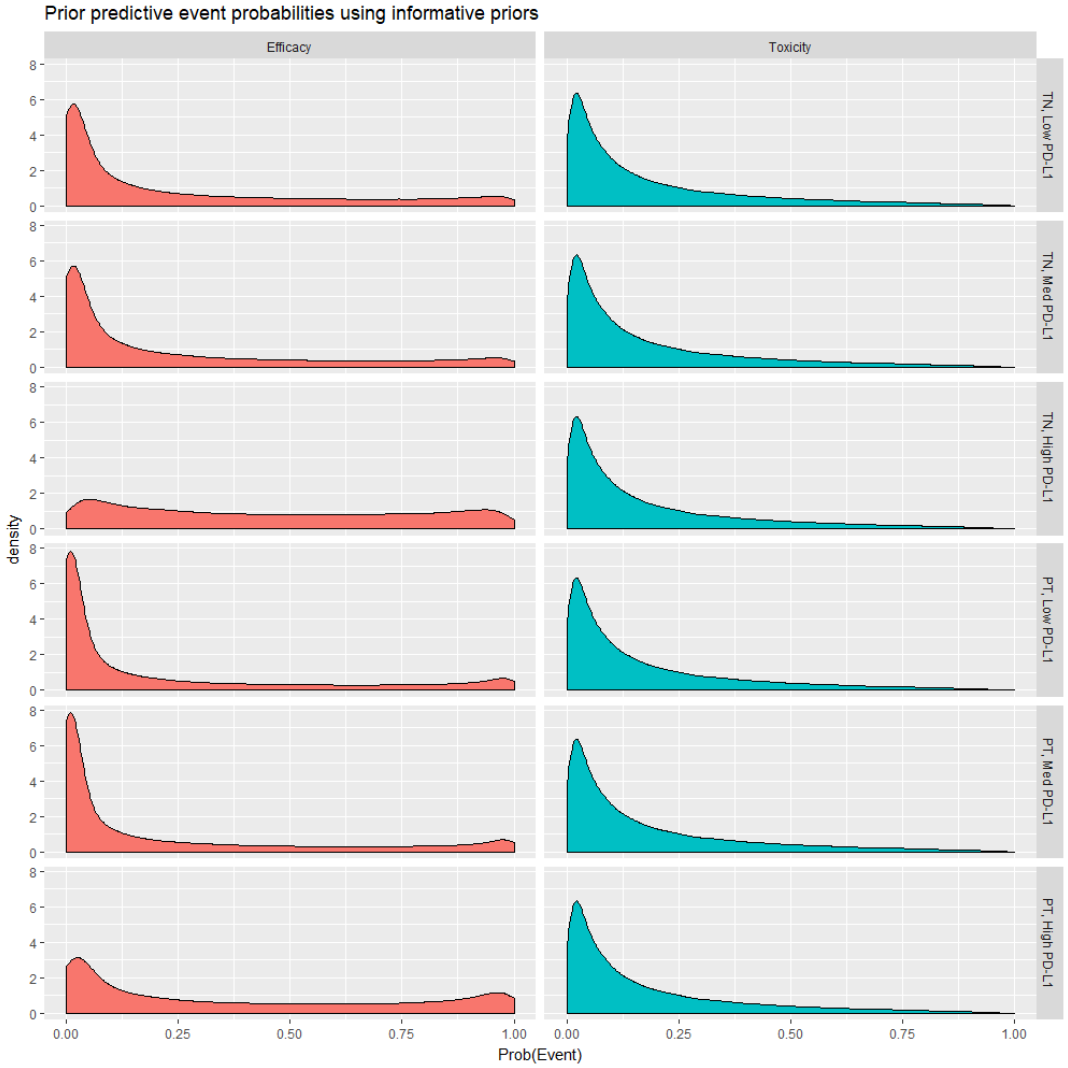



FIGURE 5.3: Prior predictive distributions of the probabilities of efficacy and toxicity in all cohorts under our informative priors.

shown in Figure 5.3. Contrast the high PD-L1 cohorts to the others. The efficacy distribution for the TN, high PD-L1 cohort, for instance, relocates a lot of the probability mass along the entire range of estimates, producing a quasi-uniform distribution that admits the potential for very high efficacy rates. There is much more probability mass in the left-hand tail at the lower efficacy rates in the other cohorts, leading to lower estimated means. We describe two further sets of priors in the following sections.

#### 5.3.3.2 Regularising priors

Informative priors have the benefit of encapsulating beliefs based on some body of knowledge. However, they can be contentious in clinical trial settings, and elsewhere, because of their ability to influence posterior beliefs in ways not reflected in the data. The PePS2 results will ultimately be published in a journal for the benefit of the medical community. Reviewers will have to be satisfied that the data have been analysed and reported in a fair way and in this regard, informative priors may hinder publication. We anticipate resistance and provide alternative analyses under different prior schemes.

TABLE 5.7: Regularising normal prior distributions for the elements of $\boldsymbol{\theta}$.

| | $\mu$ | $\sigma^2$ |
|---|---|---|
| $\alpha$ | -2.2 | 4 |
| $\beta$ | -0.5 | 4 |
| $\gamma$ | -0.5 | 4 |
| $\zeta$ | -0.5 | 4 |
| $\lambda$ | -2.2 | 4 |
| $\psi$ | 0 | 1 |

As a measure against the charge of providing a favourable analysis, we consider in this section *regularising* priors, chosen to prevent over-fitting. Listed in Table 5.7 and shown in Figure 5.4, these priors anticipate the same efficacy and toxicity event rates of approximately 20% in each cohort. This efficacy rate is close to that seen in the overall population in Garon *et al* and the toxicity rate is slightly higher. These priors could be interpreted as being sceptical with respect to the effect of our covariates, anticipating no benefit to having a higher PD-L1 score or being TN. These priors generate fairly wide credible intervals, as shown in Table 5.8.

These priors generate the types of outcomes we expect when all patients are analysed together without adjustment for covariates, so can only be considered *generative* under certain circumstances. However, they do not bias the analysis towards accepting the treatment in high PD-L1 cohorts, for instance, by assuming a high efficacy rate.

Figure 5.5 shows that the priors support a wide range of outcomes, evocative

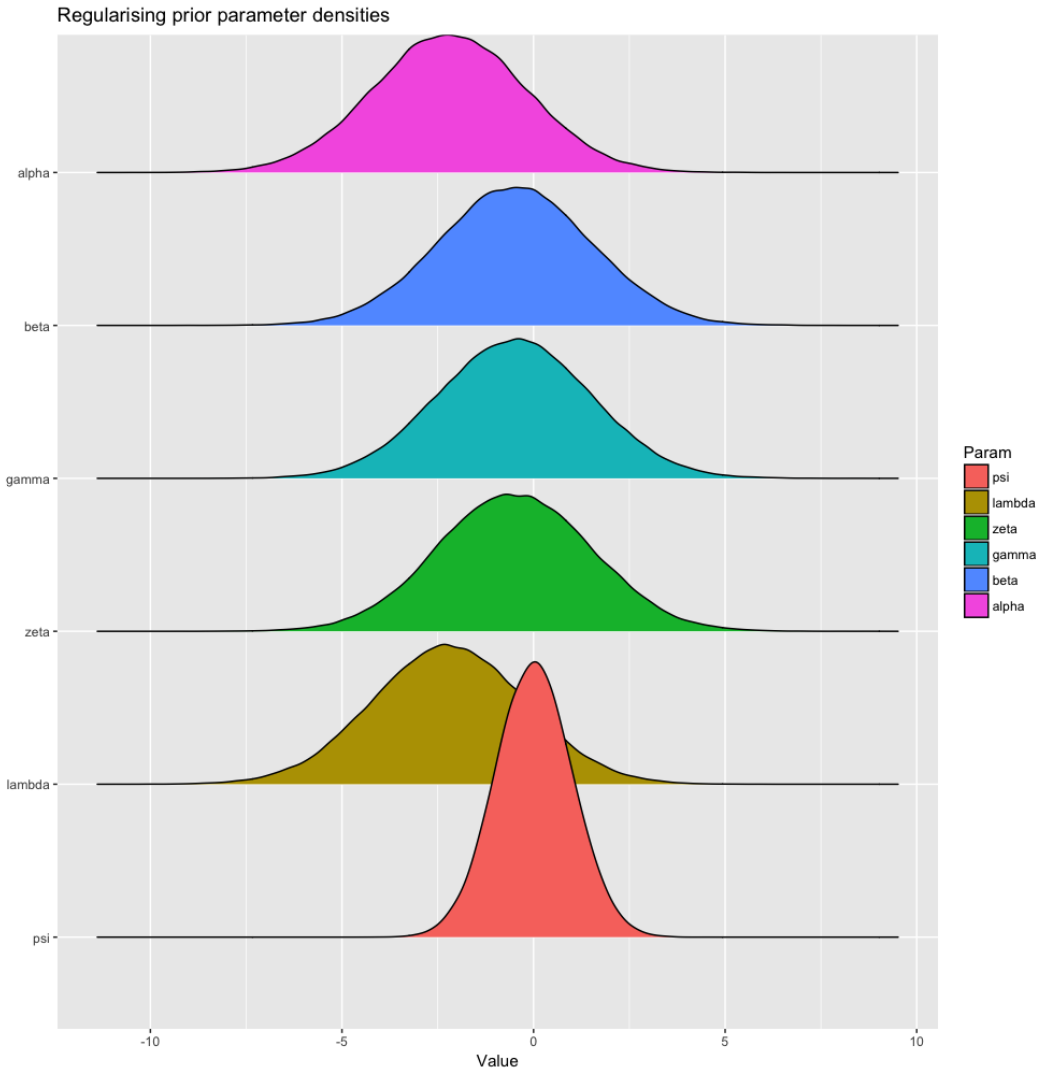



FIGURE 5.4: Regularising prior distributions on the parameters.

of so-called *spike-and-slab* and *horseshoe* priors. There is a large probability mass allocated to low event rates, reflecting the sceptical belief that most treatments are ineffective. However, the relatively flat, wide right tail of the prior facilitates the possibility of high event rates if the data are strong enough to overcome the prior scepticism. Regularising priors dissuade the model from over-fitting to small, chance events, but do not categorically rule out large effects. In our situation, we expect the priors to allow baseline effects like association of high efficacy with high PD-L1 scores to manifest through the data via the likelihood.

| Treatment status | PD-L1 | EffL | Effl | Eff | Effu | EffU |
|---|---|---|---|---|---|---|
| TN | Low | 0.00 | 0.01 | 0.21 | 0.31 | 0.87 |
| TN | Med | 0.00 | 0.01 | 0.21 | 0.31 | 0.87 |
| TN | High | 0.00 | 0.03 | 0.20 | 0.30 | 0.75 |
| PT | Low | 0.00 | 0.00 | 0.21 | 0.30 | 0.92 |
| PT | Med | 0.00 | 0.00 | 0.21 | 0.30 | 0.92 |
| PT | High | 0.00 | 0.01 | 0.21 | 0.32 | 0.87 |
| Treatment status | PD-L1 | ToxL | Toxl | Tox | Toxu | ToxU |
| TN/PT | Low-High | 0.00 | 0.03 | 0.20 | 0.30 | 0.75 |

TABLE 5.8: Credible intervals for events rates drawn from the prior predictive distribution of the regularising priors in Table 5.7. Eff and Tox show the probability of efficacy and toxicity, respectively. Lower-case l and u show the central 50% credible interval and upper-case L and U show the central 90% credible interval.

#### 5.3.3.3 Diffuse priors

Despite the encouragement for researchers to use priors that truly reflect their beliefs, it is still fairly common for diffuse priors to be used. This could be motivated by the desire that *the data should speak for themselves*. To convey the performance of our P2TNE model with very diffuse prior information, we also consider the prior parameters listed in Table 5.9 and shown in Figure 5.6.

TABLE 5.9: Diffuse normal prior distributions on $\boldsymbol{\theta}$.

| | $\mu$ | $\sigma^2$ |
|---|---|---|
| $\alpha$ | 0 | 100 |
| $\beta$ | 0 | 100 |
| $\gamma$ | 0 | 100 |
| $\zeta$ | 0 | 100 |
| $\lambda$ | 0 | 100 |
| $\psi$ | 0 | 100 |

The notable flaw with such diffuse priors is that they rarely reflect the researchers' beliefs. The statement $X \sim N(0, \sigma^2)$ generates the inference that $\text{Prob}(|X| > \frac{\sigma}{2}) > \text{Prob}(|X| < \frac{\sigma}{2})$. In the context of the priors in Table 5.9, this implies that the absolute value of each parameter is more likely to exceed 5 than to reside in the interval (-5, 5). The phenomenon is exacerbated with larger values of $\sigma$.

This folly is illustrated by the generated CIs in Table 5.10 and the prior predictive densities shown in Figure 5.7. The prior predictive distributions are horseshoe-shaped. The interaction of the normal prior with very wide tails, and the logit likelihood, which maps continuous real numbers to $(0, 1)$, puts an inordinate amount of

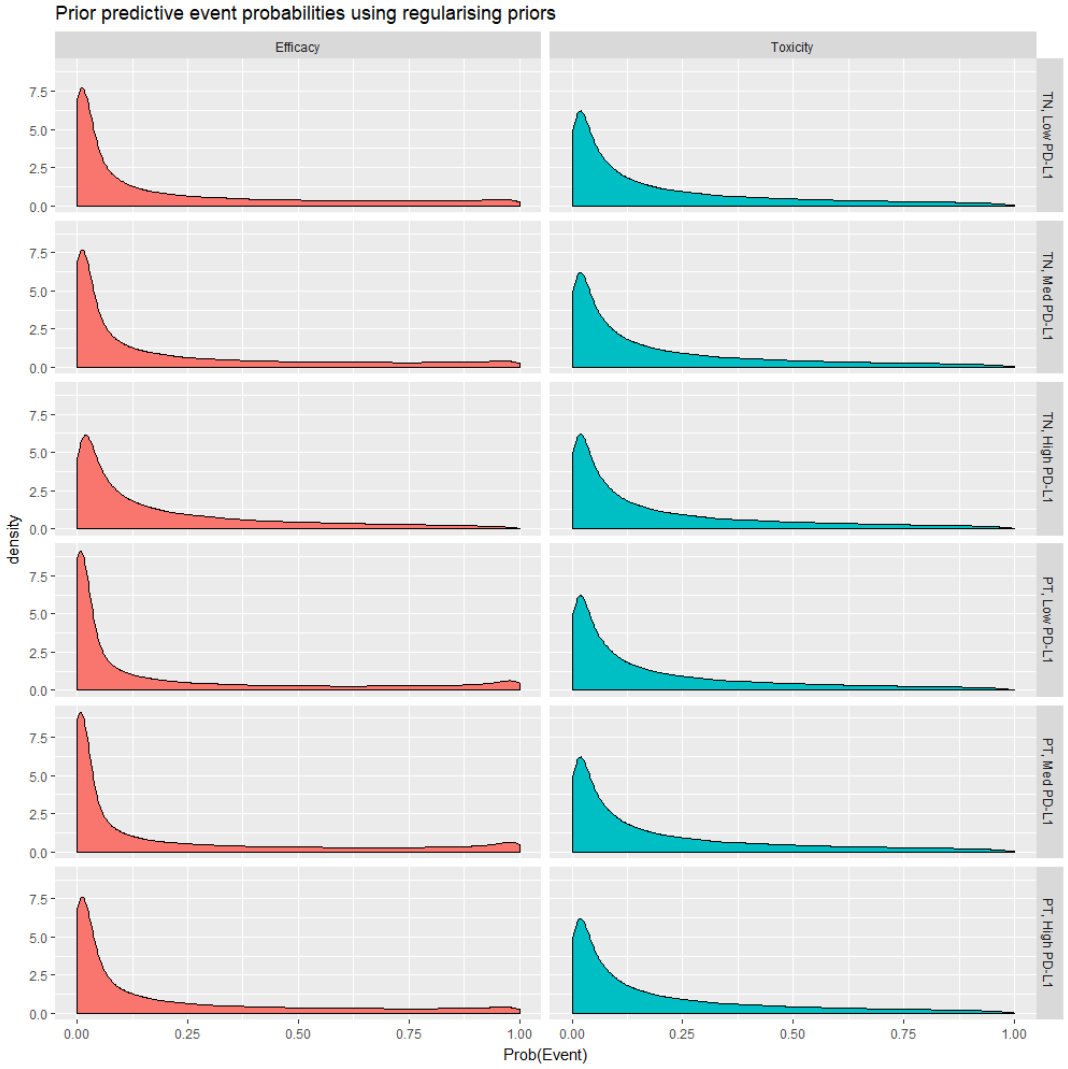



FIGURE 5.5: Prior predictive distributions of the probabilities of efficacy and toxicity in all cohorts under our regularising priors.

prior mass at the extremes. For each event rate in each cohort, 40% of the probability mass resides in the extremely narrow intervals very close to 0 and 1. Ultra-diffuse priors may sometimes be described as *uninformative*, but this example shows that the name is a misnomer when a sigmoidal link function is used. These priors absolutely do not reflect our genuine prior expectations.

In the next section, we detail how we came to choose $p_E$ and $p_T$ using the method described in 5.3.1.1.

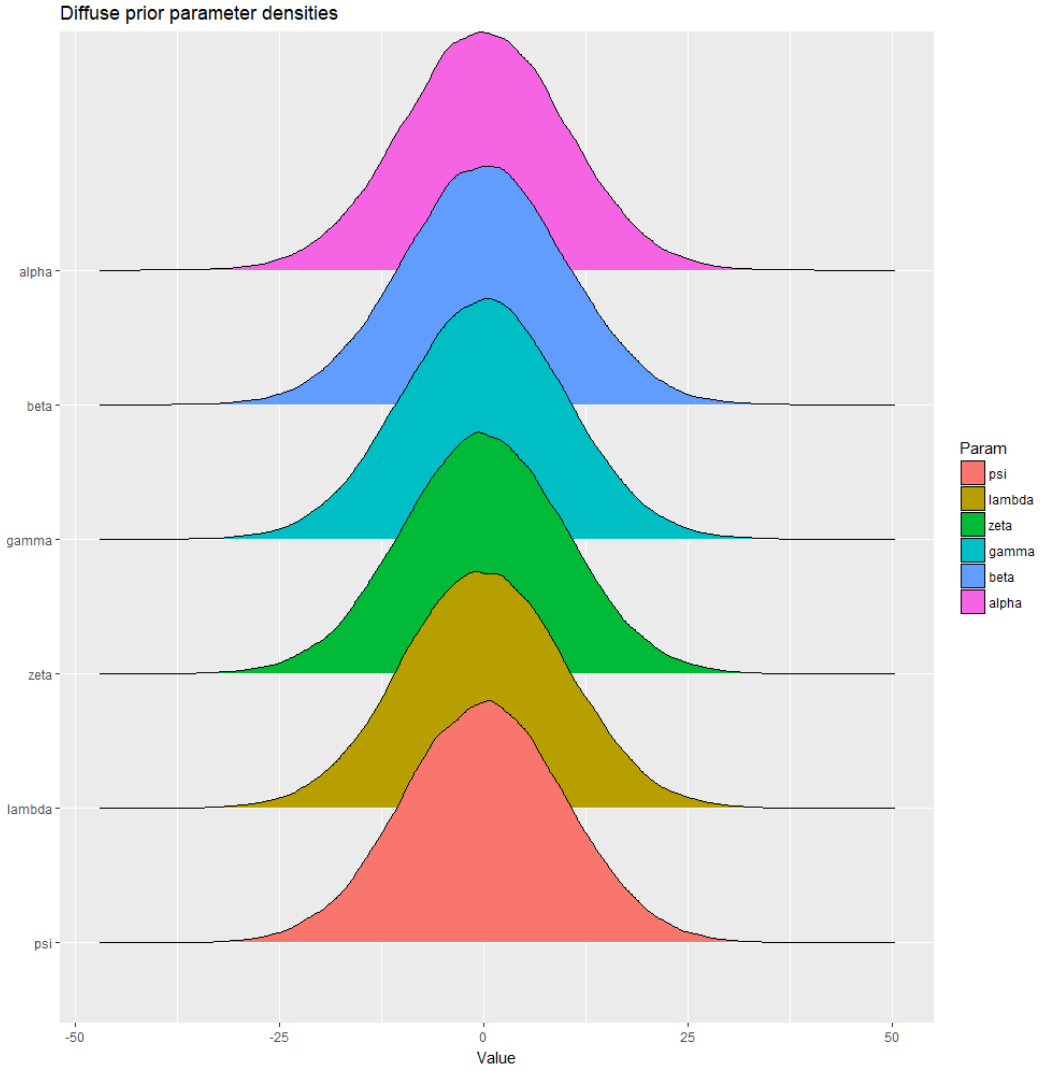



FIGURE 5.6: Diffuse prior distributions on the parameters.

## 5.4 Simulation Study

We conduct simulation studies to assess the operating characteristics of P2TNE implementations. The parameters chosen will affect performance so they should be driven by the clinical scenario, wherever possible. Sample size will naturally play a large part in determining performance. Increasing the number of patients is the typical method for providing more information to a clinical trial design with which to appraise treatments. In PePS2, the sample size is fixed at 60 patients because that is felt to be the most we could feasibly recruit in one year, so we demonstrate simulations at that level of accrual. However, we demonstrate further model embellishments and higher sample sizes in the following chapter.

At the very minimum in a simulation study, we are interested in estimating the

| Treatment status | PD-L1 | EffL | Effl | Eff | Effu | EffU |
|---|---|---|---|---|---|---|
| TN | Low | 0.000 | 0.000 | 0.499 | 1.000 | 1.000 |
| TN | Med | 0.000 | 0.000 | 0.500 | 1.000 | 1.000 |
| TN | High | 0.000 | 0.001 | 0.500 | 0.999 | 1.000 |
| PT | Low | 0.000 | 0.000 | 0.501 | 1.000 | 1.000 |
| PT | Med | 0.000 | 0.000 | 0.502 | 1.000 | 1.000 |
| PT | High | 0.000 | 0.000 | 0.501 | 1.000 | 1.000 |
| Treatment status | PD-L1 | ToxL | Toxl | Tox | Toxu | ToxU |
| TN/PT | Low-High | 0.000 | 0.001 | 0.498 | 0.999 | 1.000 |

TABLE 5.10: Credible intervals for events rates drawn from the prior predictive distribution of the diffuse priors in Table 5.9. Eff and Tox show the probability of efficacy and toxicity, respectively. Lower-case l and u show the central 50% credible interval and upper-case L and U show the central 90% credible interval.

probability that a design will correctly approve a treatment in a favourable scenario (analogous to power in a frequentist analysis) and incorrectly approve a treatment in an adverse scenario (analogous to statistical significance). Building on this minimum, we will be interested to estimate the performance of a design over a range of scenarios appropriate for the clinical setting.

### 5.4.1 Simulating cohort membership and outcomes

In the PePS2 trial, patients will belong to one of the six cohorts enumerated 1,..,6 in Table 5.3. For brevity and clarity, when discussing the parameterisation of cohorts, we present parameters that have different values for each cohort in that order. For example, a true efficacy vector $(0.1, 0.2, 0.3, 0.4, 0.5, 0.6)$ represents an efficacy rate of 0.2 in cohort 2, the cohort of patients that have not previously received treatment and have a medium PD-L1 score.

In our simulations, we will randomly sample cohort membership and this requires estimates of the cohort prevalences. In Table S9 of the supplementary information to Garon *et al.*[37], 39% of the 824 patients screened with evaluable tumour sample had low PD-L1 expression, 38% medium and 23% high. Amongst TN patients, these percentages were 31%, 44% and 25%. Amongst PT patients, they were 41%, 36% and 23%. Testing for association between the two categories by chi-squared test yields $p = 0.049$, so there is reasonable evidence to suspect that PD-L1 expression level is not identically distributed for TN and PT patients. Low PD-L1 scores appear to be less prevalent amongst TN patients.

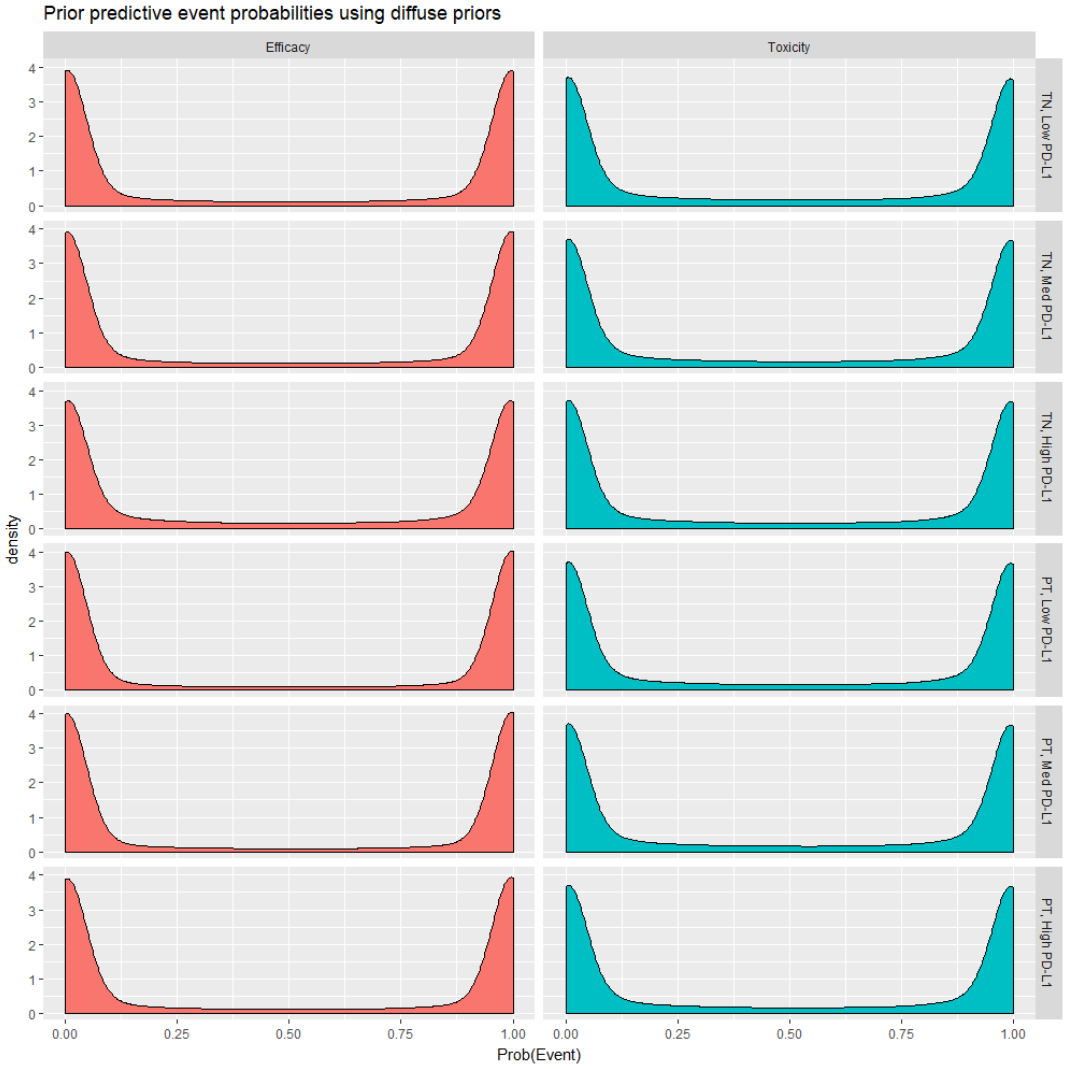



FIGURE 5.7: Prior predictive distributions of the probabilities of efficacy and toxicity in all cohorts under our diffuse priors.

The chief investigator of PePS2 expects approximately 50% of patients to have been previously treated, based on their experience with the patient population. Scaling the PD-L1 prevalences observed by Garon *et al.* in the TN and PT groups, we expect cohort membership probabilities

$$\tilde{\rho} = (0.157, 0.218, 0.124, 0.207, 0.180, 0.114)$$

For iteration $j$, we randomly sampled cohort membership probabilities, $\rho_j \sim Dirichlet(\hat{\rho})$, for $j = 1, ..., J$, where $\hat{\rho} = (15.7, 21.8, 12.4, 20.7, 18.0, 11.4)$ and $J$ is the number of simulated trial iterations. In Appendix C.1, we investigate the effect of alternative cohort prevalences.

This yielded 95% confidence intervals for $\rho_j$ given in Table 5.11. For each $j$, patient-level allocations to cohorts $1, ..., 6$ were randomly sampled from multinomial distributions with probability vector $\rho_j$. The mean cohort sizes and 95% confidence intervals are also shown in Table 5.11. These statistics are based on 100,000 random samples. The distribution of these cohort sizes approximately concurred with our expectations.

| | $\rho$ | Num patients | |
|---|---|---|---|
| Cohort | 95% CI | Mean | 95% CI |
| 1 | (0.093, 0.234) | 9.4 | (3, 17) |
| 2 | (0.143, 0.304) | 13.1 | (6, 17) |
| 3 | (0.067, 0.195) | 7.4 | (2, 14) |
| 4 | (0.134, 0.291) | 12.4 | (5, 21) |
| 5 | (0.111, 0.261) | 10.8 | (4, 19) |
| 6 | (0.060, 0.182) | 6.8 | (2, 14) |

TABLE 5.11: Simulated cohort prevalences and cohort sizes, based on 100,000 replicates.

The variance of Dirichlet random variables is determined by the size of the elements of the parameter vector, $\rho$. To consider alternatives and verify that we were using approximately the correct order of magnitude of randomness in our cohort allocations, we repeated the same exercise with Dirichlet parameter vectors $\hat{\rho}/10$ and $10\hat{\rho}$. The vector $\hat{\rho}/10$ yielded cohort sizes that were too wide, e.g. cohort sizes of zero were observed too frequently. The vector $10\hat{\rho}$ yielded cohort sizes that exhibited less variation, but looked plausible nonetheless. It is conservative to prepare for more variability rather than less, so we resolved to use $\hat{\rho}$.

A simulation scenario requires the specification of true efficacy and toxicity probabilities in each cohort, and the level of association between efficacy and toxicity events. In each scenario, we simulated $J = 10,000$ iterations. In each iteration, we randomly sampled $N = 60$ patients belonging to the six PePS2 cohorts using the method described above. We then randomly sampled binary efficacy and toxicity events with probabilities driven by the cohort memberships. We simulated correlated efficacy and toxicity events mirroring the method used in the R package `binarySimCLF` [21], itself based on the work of Qaqish[75]. The level of association is measured by odds-ratio. At the null value 1.0, efficacy is no more or less likely in the presence of toxicity. With values less than 1, the events are negatively associated,

i.e. the presence of one event makes the other event less likely.

### 5.4.2 Using simulation to select $p_E$ and $p_T$

We desire that the design approve the treatment in each cohort: (i) with at least 80% probability in each cohort in a benchmark favourable scenario where $\pi_E = 0.3$ and $\pi_T = 0.1$ throughout; and (ii) with no more than 5% probability in any cohort in a benchmark adverse scenario where $\pi_E = 0.1$ and $\pi_T = 0.3$ throughout. To demonstrate the process of choosing $p_E$ and $p_T$ for use in (5.12), Table 5.12 shows the combinations of $p_E$ and $p_T$ that we considered using the *regularising priors*.

| Parameters | Scenario | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| $p_E = 0.7, p_T = 0.7$ | Favourable | 0.90 | 0.92 | 0.91 | 0.91 | 0.91 | 0.89 |
| | Adverse | 0.06 | 0.06 | 0.07 | 0.06 | 0.06 | 0.07 |
| $p_E = 0.7, p_T = 0.8$ | Favourable | 0.90 | 0.92 | 0.91 | 0.91 | 0.91 | 0.89 |
| | Adverse | 0.04 | 0.05 | 0.05 | 0.04 | 0.04 | 0.05 |
| $p_E = 0.7, p_T = 0.9$ | Favourable | 0.90 | 0.92 | 0.91 | 0.91 | 0.91 | 0.89 |
| | Adverse | 0.02 | 0.03 | 0.03 | 0.02 | 0.02 | 0.02 |

TABLE 5.12: Probabilities of approving treatment in two key benchmark scenarios under different values for $p_E$ and $p_T$ using the regularising priors, based on 10,000 simulated trial runs. The favourable and adverse scenarios eventually became 1 & 2 in Table 5.13.

Initially, we tried $p_E = p_T = 0.7$. In the favourable scenario, the design reliably approves in all cohorts with a margin of at least 9% above our required probability of 80%. However, in the adverse scenario, the design does not reject often enough. PePS2 is an early-phase study and patients are potentially near end-of-life so we wanted to be quite confident when we say a treatment is tolerable. We can be slightly less stringent in our choice of $p_E$ because of the relative dearth of alternative treatments. To systematically arrive at an acceptable pair, we held one of $p_E$ and $p_T$ fixed, and adjusted the other. We increased $p_T$ to 0.8 so that the design would be more demanding when it infers the treatment is tolerable, and held $p_E$ constant. This reduced the probability of wrongly accepting in the adverse scenario to approximately 5% in all cohorts and did not perceptibly change the probability of accepting in the favourable scenario. The apparent ability to improve the probability in the adverse scenario without impacting the favourable scenario motivated investigation of a further increase in $p_T$ to 0.9, yielding the final two rows in Table 5.12. Again, probabilities in the favourable scenario did not change using 2 decimal places. With

$p_E = 0.7$ and $p_T = 0.9$, the design rejects in all cohorts in the adverse scenario at least 97% of the time, and approves in all cohorts in the favourable scenario at least 89% of the time. This was deemed a desirable compromise that addresses the two competing goals represented by the two scenarios.

The different sets of priors required their own values for $p_E$ and $p_T$. Similar processes showed that those same values would achieve the same under the diffuse priors; and that a small adjustment to $p_E = 0.7$ and $p_T = 0.95$ would achieve the same under the informative priors. Even though $p_E$ and $p_T$ took different values for different priors, the values for $\pi_E$ and $\pi_T$ were the same throughout.

In the next section, we assess performance over a wider range of scenarios. The favourable scenario above became scenario 1 and the adverse scenario became scenario 2 in the broad simulation study described below.

### 5.4.3 Main simulation study

Table 5.13 shows simulated performance in six scenarios. The scenarios chosen broadly reflect our expectations, driven by the Garon study, and the range of scenarios over which the design should perform well. The simulated mean number of patients, and efficacy and toxicity events, are presented for each cohort. The probabilities of approving treatment using P2TNE models under the informative, regularising, and diffuse priors are shown.

In scenarios 1 - 3, the rates of efficacy and toxicity are uniform across the cohorts. Scenario 1 is our benchmark favourable scenario. It shows that if the true probability of efficacy is 30% and toxicity is 10%, we can expect all designs to approve the treatment with at least 80% probability in all cohorts, irrespective the priors used. The cohorts have different approval probabilities because the average cohort sizes are different. Under the diffuse priors, cohorts 3 and 6 have the smallest approval probabilities because they have the fewest patients. In contrast, those same probabilities are very high under the informative priors because the observed data concur with the prior expectation, confirming that efficacy is good throughout. A key benefit of the P2TNE design is the sharing of information across cohorts via the Bayesian regression model. For instance, designs will quite reliably approve the treatment in scenario 1 in cohorts 3 and 6, even though they each only receive approximately 7

| Sc | Coh | PrEff | PrTox | Odds | N | Eff | Tox | Inf | Reg | Diffuse | BetaBin |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0.300 | 0.1 | 1.0 | 9.3 | 2.8 | 0.9 | 0.883 | 0.896 | 0.878 | 0.540 |
| | 2 | 0.300 | 0.1 | 1.0 | 13.1 | 3.9 | 1.3 | 0.906 | 0.920 | 0.905 | 0.658 |
| | 3 | 0.300 | 0.1 | 1.0 | 7.5 | 2.3 | 0.8 | 0.980 | 0.909 | 0.816 | 0.473 |
| | 4 | 0.300 | 0.1 | 1.0 | 12.5 | 3.7 | 1.2 | 0.875 | 0.912 | 0.896 | 0.635 |
| | 5 | 0.300 | 0.1 | 1.0 | 10.8 | 3.2 | 1.1 | 0.873 | 0.909 | 0.890 | 0.590 |
| | 6 | 0.300 | 0.1 | 1.0 | 6.8 | 2.0 | 0.7 | 0.959 | 0.893 | 0.819 | 0.459 |
| 2 | 1 | 0.100 | 0.3 | 1.0 | 9.3 | 0.9 | 2.8 | 0.012 | 0.025 | 0.019 | 0.035 |
| | 2 | 0.100 | 0.3 | 1.0 | 13.1 | 1.3 | 3.9 | 0.013 | 0.028 | 0.023 | 0.032 |
| | 3 | 0.100 | 0.3 | 1.0 | 7.5 | 0.8 | 2.3 | 0.038 | 0.029 | 0.021 | 0.034 |
| | 4 | 0.100 | 0.3 | 1.0 | 12.5 | 1.2 | 3.7 | 0.009 | 0.024 | 0.021 | 0.034 |
| | 5 | 0.100 | 0.3 | 1.0 | 10.8 | 1.1 | 3.2 | 0.009 | 0.024 | 0.022 | 0.032 |
| | 6 | 0.100 | 0.3 | 1.0 | 6.8 | 0.7 | 2.0 | 0.027 | 0.025 | 0.019 | 0.041 |
| 3 | 1 | 0.300 | 0.1 | 0.2 | 9.3 | 2.8 | 0.9 | 0.884 | 0.897 | 0.879 | 0.562 |
| | 2 | 0.300 | 0.1 | 0.2 | 13.1 | 3.9 | 1.3 | 0.906 | 0.920 | 0.904 | 0.667 |
| | 3 | 0.300 | 0.1 | 0.2 | 7.5 | 2.3 | 0.8 | 0.981 | 0.909 | 0.818 | 0.494 |
| | 4 | 0.300 | 0.1 | 0.2 | 12.5 | 3.7 | 1.2 | 0.877 | 0.913 | 0.897 | 0.652 |
| | 5 | 0.300 | 0.1 | 0.2 | 10.8 | 3.2 | 1.1 | 0.874 | 0.908 | 0.889 | 0.605 |
| | 6 | 0.300 | 0.1 | 0.2 | 6.8 | 2.0 | 0.7 | 0.960 | 0.893 | 0.820 | 0.478 |
| 4 | 1 | 0.167 | 0.1 | 1.0 | 9.3 | 1.5 | 0.9 | 0.408 | 0.451 | 0.398 | 0.293 |
| | 2 | 0.192 | 0.1 | 1.0 | 13.1 | 2.5 | 1.3 | 0.651 | 0.690 | 0.633 | 0.432 |
| | 3 | 0.500 | 0.1 | 1.0 | 7.5 | 3.8 | 0.8 | 0.993 | 0.981 | 0.974 | 0.622 |
| | 4 | 0.091 | 0.1 | 1.0 | 12.5 | 1.1 | 1.3 | 0.208 | 0.277 | 0.215 | 0.131 |
| | 5 | 0.156 | 0.1 | 1.0 | 10.8 | 1.7 | 1.1 | 0.405 | 0.493 | 0.419 | 0.298 |
| | 6 | 0.439 | 0.1 | 1.0 | 6.8 | 3.0 | 0.7 | 0.961 | 0.930 | 0.931 | 0.581 |
| 5 | 1 | 0.167 | 0.3 | 1.0 | 9.3 | 1.5 | 2.8 | 0.027 | 0.063 | 0.039 | 0.071 |
| | 2 | 0.192 | 0.3 | 1.0 | 13.1 | 2.5 | 3.9 | 0.046 | 0.099 | 0.066 | 0.084 |
| | 3 | 0.500 | 0.3 | 1.0 | 7.5 | 3.8 | 2.3 | 0.071 | 0.141 | 0.102 | 0.159 |
| | 4 | 0.091 | 0.3 | 1.0 | 12.5 | 1.1 | 3.7 | 0.014 | 0.037 | 0.021 | 0.028 |
| | 5 | 0.156 | 0.3 | 1.0 | 10.8 | 1.7 | 3.2 | 0.030 | 0.071 | 0.045 | 0.065 |
| | 6 | 0.439 | 0.3 | 1.0 | 6.8 | 3.0 | 2.0 | 0.070 | 0.135 | 0.099 | 0.163 |
| 6 | 1 | 0.167 | 0.1 | 0.2 | 9.3 | 1.5 | 0.9 | 0.408 | 0.451 | 0.396 | 0.308 |
| | 2 | 0.192 | 0.1 | 0.2 | 13.1 | 2.5 | 1.3 | 0.651 | 0.689 | 0.633 | 0.447 |
| | 3 | 0.500 | 0.1 | 0.2 | 7.5 | 3.8 | 0.8 | 0.993 | 0.981 | 0.974 | 0.627 |
| | 4 | 0.091 | 0.1 | 0.2 | 12.5 | 1.1 | 1.3 | 0.208 | 0.278 | 0.212 | 0.139 |
| | 5 | 0.156 | 0.1 | 0.2 | 10.8 | 1.7 | 1.1 | 0.402 | 0.493 | 0.415 | 0.313 |
| | 6 | 0.439 | 0.1 | 0.2 | 6.8 | 3.0 | 0.7 | 0.962 | 0.929 | 0.930 | 0.589 |

TABLE 5.13: A summary of simulated trials. Sc is scenario number and Coh the cohort number. Patient cohorts are defined in Table 5.3. PrEff and PrTox are the true probabilities of efficacy and toxicity. Odds denotes the ratio of odds of efficacy in patients that experience toxicity to those that do not. Odds=1 corresponds to no association; values less than one convey that efficacy is less likely when toxicity is observed; and vice-versa. N is the mean number of patients in a cohort; Eff and Tox the mean number of events. Inf is the probability the treatment is approved by the P2TNE model using informative priors; Reg and Diffuse are probabilities of approval under the regularising and diffuse priors. BetaBin is the approval probability using cohort-specific beta-binomial models. 10,000 iterations were used in each scenario.

patients who experience 2 efficacies. The high efficacy rate observed in other cohorts informs the posterior inference.

To give measure to the benefit of information sharing in P2TNE, we also present in the final column in Table 5.13 the approval probabilities under cohort-specific beta-binomial Bayesian conjugate models that assume the efficacy and toxicity events are independent. With prior $\pi \sim Beta(\alpha, \beta)$, the posterior beliefs are $\pi|r, n \sim Beta(\alpha + r, \beta + n - r)$ where $n$ is the number of patients in a cohort and $r$ is the number of events observed. Inferences are made on the posterior distribution. Re-implementing the same decision criteria, the beta-binomial models approve the treatment in a given cohort if $Pr(\pi_E > 0.1|r_E, n) > 0.7$ and $Pr(\pi_T < 0.3|r_T, n) > 0.9$. Diffuse $Beta(0.001, 0.001)$ priors on the rates of efficacy and toxicity were used. The beta-binomials models make decisions in each cohort singly and do not share information. Comparing the performances of the P2TNE model with diffuse priors to the beta-binomial models shows the benefit to sharing information.

In scenario 1, for instance, the P2TNE model with diffuse priors outperforms the beta-binomial model by at least 25% in each cohort. The beta-binomial model would approve the treatment less than 50% of the time in cohort 3 in scenario 1. With exactly 7 patients, being the median size of this cohort, the beta-binomial model must observe at least $r_E = 2$ efficacy events and exactly $r_T = 0$ toxicities to conclude that the treatment is acceptable. With event probabilities 0.3 and 0.1 respectively, the joint probability of this occurring is 0.32 using exact binomial probabilities, assuming independence. The individual cohorts are simply too small with the overall sample size $n = 60$ to achieve in a cohort-by-cohort analysis error rates typically used in clinical trials. However, information sharing can have adverse consequences too. The chances of erroneously approving in subgroups with poor efficacy will be inflated when positive effects are observed in other subgroups. We demonstrate an example of this in scenario 4 below.

Scenario 2 is our benchmark adverse scenario, where toxicity is 30% and efficacy is 10% in all cohorts. The designs should reject because the efficacy probability is undesirably low and the toxicity probability is undesirably high. As required, we see that all designs are very likely to reject. This was ensured by the selection of $p_E$ and $p_T$, as described above. Once again, it is revealing to compare P2TNE to the

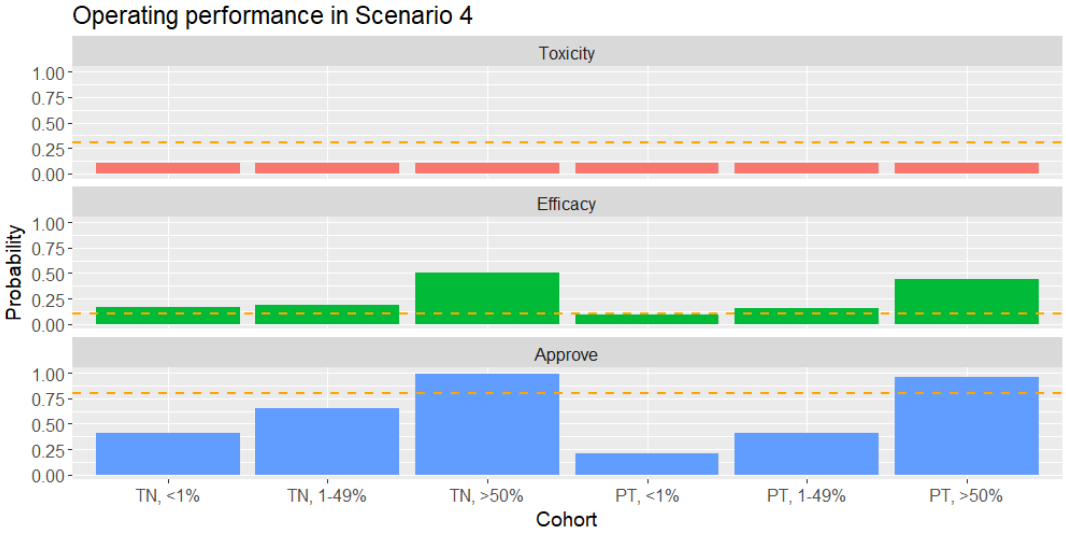



FIGURE 5.8: Performance of the P2TNE model in scenario 4 under informative priors.

beta-binomial alternative. Despite leveraging information to approve the treatment with small cohort sizes when performance is good, it does not show a predisposition to approve the treatment when outcomes are poor. In fact, the P2TNE design is generally more likely than the cohort-specific beta-binomial models to reject the treatment in scenario 2, irrespective the priors, because it uses information from all 60 patients.

One of the features of P2TNE is that it models the association between efficacy and toxicity. Scenario 3 shows performance when efficacy events are highly negatively associated with toxicity. Here, the ability of patients to achieve efficacious outcomes are strongly hindered if they experience a toxicity event. In every other regard, the parameterisation of scenario 3 is the same as scenario 1. We see that performance barely differs. This calls into question the benefit of modelling associated co-primary outcomes, a topic we return to in the Discussion and investigate further in the next chapter.

Scenario 4 uses efficacy probabilities that match the response rates observed in Garon *et al.*[37], with a uniform toxicity probability of 10%. This reflects the type of scenario we expect to observe in PePS2. A notable aspect is the apparent interaction yielded by simultaneous low PD-L1 and pre-treated status so that the PD-L1-efficacy curves are not piecewise-parallel, as depicted in Figure 5.1. Performance is shown in Figure 5.8 under the informative priors. Our design is overwhelmingly likely to

approve treatment in cohorts 3 and 6 where efficacy is high.

Cohort 2 in scenario 4 is an intermediate case. Whilst the toxicity rate is manageable at only 10%, the efficacy rate of 19.2% is attractive. However, 30% is the efficacy rate that we would not like to miss, so we do not necessarily demand that the designs offer 80% approval probabilities here. We see that the P2TNE designs are 60-70% likely to approve treatment here, an improvement over the beta-binomial model of at least 20%. P2TNE manages this, despite an average cohort size of 13.1 patients and efficacy rate only 9.2% above the 10% threshold by leveraging the information observed in other cohorts. Naturally, the opposite effect occurs too. The design is 20-30% likely to approve in cohort 4, and 7-15% more likely than the beta-binomial analyses, even though the true efficacy probability is insufficient. The model has inflated expectations of the efficacy probability because of the good efficacy observed in other cohorts. The natural solution to this flaw is to introduce interaction terms between the independent variables, a topic we develop in the next chapter

Scenario 5 shows the same efficacy probabilities as scenario 4 combined with a high toxicity probability of 30%. All approaches are now much less likely to approve the treatment. All analyses except the P2TNE method with informative priors show approval probabilities in excess of 10% in cohorts 3 and 6 where efficacy is highest. Here, the methods are overwhelmingly inclined to accept the treatment from an efficacy stance, but lack sufficient information on toxicity in a noteworthy percentage of simulations to correctly reject the treatment. This can be addressed by using the informative priors.

Finally, scenario 6 shows the same efficacy and toxicity rates as scenario 4, where the events are now strongly negatively associated. Once again, we see that performance under P2TNE barely changes, challenging the benefit of modelling associated co-primary outcomes in this trial. We see from Figure 5.9 that the estimation of $\psi$ adapts to the prevailing scenario. For instance, the estimates are clustered around zero when efficacy and toxicity are genuinely independent in scenario 4, but overwhelming negative in scenario 6. We revisit this in the next chapter.

The trial design yields simple dichotomous decisions on whether there exists sufficient evidence to warrant further study. To make this decision, the underlying statistical model produces estimates of the probabilities of efficacy and toxicity in

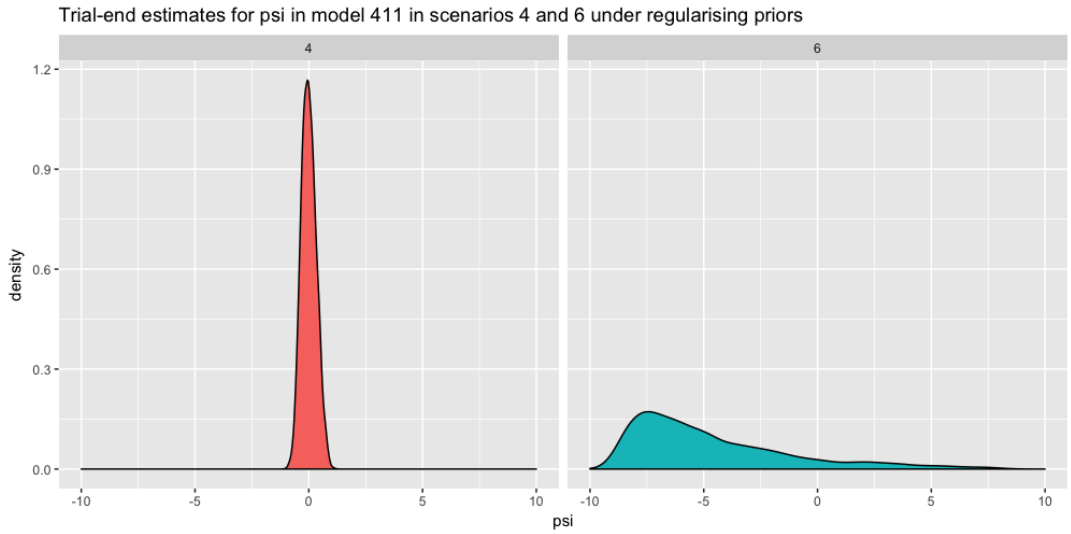



FIGURE 5.9: Simulated end-of-trial estimates of $\psi$ in scenarios 4 & 6.

each cohort. In simulations, we know the underlying values that generated the hypothetical trial outcomes so it is possible to assess the numerical performance of the model. The following definitions are adapted from Morris *et al.*[66].

For estimates $\hat{\theta}_1, ..., \hat{\theta}_K$ of estimand taking true value $\theta$, the *bias* of the estimator process is

$$\frac{1}{K}\sum_{k=1}^{K}\hat{\theta}_k - \theta \tag{5.14}$$

Bias measures whether the estimator targets the true value, on average, and an unbiased estimator has bias equal to zero. Let $\bar{\theta}$ be the sample mean of the $\hat{\theta}_k$. The *empirical standard error* is

$$\sqrt{\frac{1}{K-1}\sum_{k=1}^{K}(\hat{\theta}_k - \bar{\theta})^2} \tag{5.15}$$

and measures the standard deviation of the estimates.

Finally, the *coverage* of an estimator reflects the percentage of iterations in which an uncertainty interval of given width contains the true value. Let the equal-tailed 90% credible interval of estimate $\hat{\theta}_k$ have lower bound $\hat{\theta}_{low,k}$ and upper bound $\hat{\theta}_{upp,k}$. Then, the coverage is estimated by

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{I}(\hat{\theta}_{low,k} \leq \theta \leq \hat{\theta}_{upp,k}) \tag{5.16}$$

By our definition, we expect coverage values of 90%.

| Scenario | Cohort | EffBias | EffEmpSE | EffCov | ToxBias | ToxEmpSE | ToxCov |
|---|---|---|---|---|---|---|---|
| 1 | 1 | -0.006 | 0.113 | 0.882 | 0.001 | 0.037 | 0.910 |
| | 2 | -0.008 | 0.104 | 0.879 | 0.001 | 0.037 | 0.910 |
| | 3 | -0.004 | 0.108 | 0.912 | 0.001 | 0.037 | 0.910 |
| | 4 | -0.007 | 0.107 | 0.884 | 0.001 | 0.037 | 0.910 |
| | 5 | -0.004 | 0.109 | 0.882 | 0.001 | 0.037 | 0.910 |
| | 6 | 0.001 | 0.116 | 0.905 | 0.001 | 0.037 | 0.910 |
| 2 | 1 | 0.004 | 0.069 | 0.865 | -0.006 | 0.058 | 0.894 |
| | 2 | 0.001 | 0.063 | 0.864 | -0.006 | 0.058 | 0.894 |
| | 3 | 0.015 | 0.063 | 0.951 | -0.006 | 0.058 | 0.894 |
| | 4 | -0.007 | 0.064 | 0.827 | -0.006 | 0.058 | 0.894 |
| | 5 | -0.004 | 0.065 | 0.834 | -0.006 | 0.058 | 0.894 |
| | 6 | 0.012 | 0.070 | 0.909 | -0.006 | 0.058 | 0.894 |
| 3 | 1 | -0.007 | 0.113 | 0.882 | 0.000 | 0.037 | 0.913 |
| | 2 | -0.009 | 0.104 | 0.879 | 0.000 | 0.037 | 0.913 |
| | 3 | -0.004 | 0.107 | 0.912 | 0.000 | 0.037 | 0.913 |
| | 4 | -0.008 | 0.106 | 0.885 | 0.000 | 0.037 | 0.913 |
| | 5 | -0.005 | 0.109 | 0.883 | 0.000 | 0.037 | 0.913 |
| | 6 | 0.000 | 0.116 | 0.905 | 0.000 | 0.037 | 0.913 |
| 4 | 1 | -0.008 | 0.085 | 0.866 | 0.001 | 0.037 | 0.908 |
| | 2 | 0.009 | 0.089 | 0.892 | 0.001 | 0.037 | 0.908 |
| | 3 | -0.072 | 0.126 | 0.867 | 0.001 | 0.037 | 0.908 |
| | 4 | 0.030 | 0.066 | 0.902 | 0.001 | 0.037 | 0.908 |
| | 5 | 0.007 | 0.082 | 0.887 | 0.001 | 0.037 | 0.908 |
| | 6 | -0.075 | 0.132 | 0.844 | 0.001 | 0.037 | 0.908 |
| 5 | 1 | -0.008 | 0.085 | 0.864 | -0.005 | 0.058 | 0.892 |
| | 2 | 0.009 | 0.089 | 0.892 | -0.005 | 0.058 | 0.892 |
| | 3 | -0.072 | 0.126 | 0.867 | -0.005 | 0.058 | 0.892 |
| | 4 | 0.030 | 0.066 | 0.901 | -0.005 | 0.058 | 0.892 |
| | 5 | 0.007 | 0.082 | 0.887 | -0.005 | 0.058 | 0.892 |
| | 6 | -0.075 | 0.132 | 0.842 | -0.005 | 0.058 | 0.892 |
| 6 | 1 | -0.009 | 0.084 | 0.866 | 0.001 | 0.037 | 0.911 |
| | 2 | 0.009 | 0.088 | 0.893 | 0.001 | 0.037 | 0.911 |
| | 3 | -0.072 | 0.126 | 0.867 | 0.001 | 0.037 | 0.911 |
| | 4 | 0.030 | 0.066 | 0.903 | 0.001 | 0.037 | 0.911 |
| | 5 | 0.007 | 0.082 | 0.886 | 0.001 | 0.037 | 0.911 |
| | 6 | -0.075 | 0.132 | 0.842 | 0.001 | 0.037 | 0.911 |

TABLE 5.14: The prefixes Eff and Tox denote the estimates of the probability of efficacy and toxicity under regularising priors. The suffix Bias denotes bias; EmpSE denotes the empirical standard error; and Cov denotes 90% credible interval coverage.

The bias, empirical standard error and coverage of the P2TNE estimators of the probabilities of efficacy and toxicity under our regularising, informative and diffuse priors are given in Tables 5.14, 5.15 and 5.16. Furthermore, Table 5.17 summarises the mean performance by each measure for each model.

Firstly, we observe that the bias in the estimated probabilities of toxicity are low, the empirical standard errors are relatively low, and the coverages are close to 90% throughout. This is not particularly surprising given that the toxicity probabilities do not vary by group and the analysis model uses only an intercept parameter. That single parameter is estimated well with $n = 60$ and there is no cohort heterogeneity to contend with.

Under the regularising priors, Table 5.14 shows noteworthy negative bias in the estimation of $\pi_E$ in scenarios 4-6 in the high PD-L1 cohorts, i.e. the model estimates efficacies in cohorts 3 and 6 that are habitually lower than the underlying truth. Correspondingly, we see under-coverage in these cohorts. As we saw in Table 5.13, however, this does not adversely impair the approval probabilities. In contrast, the model with regularising priors correctly approves with high probability in these cohorts in scenario 4 despite the negative bias, and correctly rejects in scenario 5. Downward bias in cohorts 3 and 6 in scenarios 4-6 is associated with modest upward bias in the efficacy estimate in cohort 4. This stems from the absence of interaction terms in the efficacy sub-model. We investigate this further in the next chapter. Table 5.14 shows that the other cohorts are largely unaffected. The regularising priors anticipated efficacy probabilities of 20% in all cohorts. The *shrinkage* in cohorts 3 and 6 demonstrates an attractive aspect of regularisation. The model is dissuaded from fitting values that diverge substantially from the population mean, particularly when sample sizes are small, as a conservative measure to guard against over-fitting. We see that the overall analysis objective is not impaired by this shrinkage

Table 5.15 essentially shows the complementary phenomenon that arises from using our informative priors. Here, there is little bias in the estimation of efficacy in scenario 4-6. There is, however, material upward bias in the high PD-L1 cohorts in scenarios 1-3, as the observed efficacy rates diverge from those generated by the priors. Once again, Table 5.13 confirms that this does not adversely affect operating performance, with the error rates in the desired range in attractive and adverse scenarios.

Table 5.16 shows numerical performance under the diffuse priors. We see that bias is typically low throughout. Comparing to Tables 5.14 and 5.15, the standard errors are greater with less prior information available to keep the estimates in the

| Scenario | Cohort | EffBias | EffEmpSE | EffCov | ToxBias | ToxEmpSE | ToxCov |
|---|---|---|---|---|---|---|---|
| 1 | 1 | -0.012 | 0.113 | 0.876 | 0.001 | 0.036 | 0.911 |
| | 2 | -0.014 | 0.104 | 0.873 | 0.001 | 0.036 | 0.911 |
| | 3 | 0.073 | 0.115 | 0.889 | 0.001 | 0.036 | 0.911 |
| | 4 | -0.026 | 0.105 | 0.859 | 0.001 | 0.036 | 0.911 |
| | 5 | -0.023 | 0.107 | 0.861 | 0.001 | 0.036 | 0.911 |
| | 6 | 0.062 | 0.123 | 0.892 | 0.001 | 0.036 | 0.911 |
| 2 | 1 | -0.001 | 0.068 | 0.843 | -0.008 | 0.058 | 0.885 |
| | 2 | -0.004 | 0.062 | 0.846 | -0.008 | 0.058 | 0.885 |
| | 3 | 0.078 | 0.078 | 0.897 | -0.008 | 0.058 | 0.885 |
| | 4 | -0.019 | 0.060 | 0.769 | -0.008 | 0.058 | 0.885 |
| | 5 | -0.017 | 0.060 | 0.782 | -0.008 | 0.058 | 0.885 |
| | 6 | 0.056 | 0.086 | 0.910 | -0.008 | 0.058 | 0.885 |
| 3 | 1 | -0.012 | 0.113 | 0.877 | 0.000 | 0.036 | 0.915 |
| | 2 | -0.015 | 0.104 | 0.873 | 0.000 | 0.036 | 0.915 |
| | 3 | 0.072 | 0.114 | 0.890 | 0.000 | 0.036 | 0.915 |
| | 4 | -0.026 | 0.105 | 0.859 | 0.000 | 0.036 | 0.915 |
| | 5 | -0.024 | 0.107 | 0.862 | 0.000 | 0.036 | 0.915 |
| | 6 | 0.061 | 0.122 | 0.893 | 0.000 | 0.036 | 0.915 |
| 4 | 1 | -0.017 | 0.085 | 0.841 | 0.001 | 0.036 | 0.909 |
| | 2 | 0.002 | 0.089 | 0.883 | 0.001 | 0.036 | 0.909 |
| | 3 | 0.010 | 0.127 | 0.917 | 0.001 | 0.036 | 0.909 |
| | 4 | 0.015 | 0.063 | 0.887 | 0.001 | 0.036 | 0.909 |
| | 5 | -0.010 | 0.079 | 0.852 | 0.001 | 0.036 | 0.909 |
| | 6 | -0.020 | 0.137 | 0.894 | 0.001 | 0.036 | 0.909 |
| 5 | 1 | -0.017 | 0.085 | 0.839 | -0.007 | 0.058 | 0.885 |
| | 2 | 0.002 | 0.089 | 0.884 | -0.007 | 0.058 | 0.885 |
| | 3 | 0.010 | 0.127 | 0.915 | -0.007 | 0.058 | 0.885 |
| | 4 | 0.015 | 0.063 | 0.886 | -0.007 | 0.058 | 0.885 |
| | 5 | -0.010 | 0.079 | 0.852 | -0.007 | 0.058 | 0.885 |
| | 6 | -0.020 | 0.137 | 0.893 | -0.007 | 0.058 | 0.885 |
| 6 | 1 | -0.017 | 0.084 | 0.839 | 0.001 | 0.036 | 0.912 |
| | 2 | 0.002 | 0.089 | 0.885 | 0.001 | 0.036 | 0.912 |
| | 3 | 0.010 | 0.126 | 0.918 | 0.001 | 0.036 | 0.912 |
| | 4 | 0.015 | 0.063 | 0.886 | 0.001 | 0.036 | 0.912 |
| | 5 | -0.010 | 0.078 | 0.852 | 0.001 | 0.036 | 0.912 |
| | 6 | -0.021 | 0.137 | 0.894 | 0.001 | 0.036 | 0.912 |

TABLE 5.15: The prefixes Eff and Tox denote the estimates of the probability of efficacy and toxicity under informative priors. The suffix Bias denotes bias; EmpSE denotes the empirical standard error; and Cov denotes 90% credible interval coverage.

expected ranges. Given the low bias, we might be surprised to see that coverage of the efficacy probabilities is typically lowest under the diffuse priors. Reasons for under-coverage are given by Morris *et al.*[66], with the applicable explanation here being the excess variability in the estimates.

These general observations are reflected in the mean measures presented in Table

| Scenario | Cohort | EffBias | EffEmpSE | EffCov | ToxBias | ToxEmpSE | ToxCov |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 0.002 | 0.125 | 0.866 | 0.001 | 0.039 | 0.898 |
| | 2 | -0.002 | 0.113 | 0.866 | 0.001 | 0.039 | 0.898 |
| | 3 | 0.003 | 0.140 | 0.861 | 0.001 | 0.039 | 0.898 |
| | 4 | -0.002 | 0.114 | 0.873 | 0.001 | 0.039 | 0.898 |
| | 5 | 0.001 | 0.119 | 0.867 | 0.001 | 0.039 | 0.898 |
| | 6 | 0.004 | 0.142 | 0.861 | 0.001 | 0.039 | 0.898 |
| 2 | 1 | 0.004 | 0.085 | 0.785 | 0.000 | 0.060 | 0.890 |
| | 2 | -0.000 | 0.074 | 0.804 | 0.000 | 0.060 | 0.890 |
| | 3 | 0.005 | 0.095 | 0.734 | 0.000 | 0.060 | 0.890 |
| | 4 | -0.003 | 0.076 | 0.787 | 0.000 | 0.060 | 0.890 |
| | 5 | 0.003 | 0.081 | 0.790 | 0.000 | 0.060 | 0.890 |
| | 6 | 0.008 | 0.099 | 0.748 | 0.000 | 0.060 | 0.890 |
| 3 | 1 | 0.000 | 0.123 | 0.869 | -0.000 | 0.038 | 0.901 |
| | 2 | -0.004 | 0.112 | 0.869 | -0.000 | 0.038 | 0.901 |
| | 3 | 0.001 | 0.139 | 0.864 | -0.000 | 0.038 | 0.901 |
| | 4 | -0.004 | 0.113 | 0.875 | -0.000 | 0.038 | 0.901 |
| | 5 | -0.001 | 0.117 | 0.870 | -0.000 | 0.038 | 0.901 |
| | 6 | 0.002 | 0.141 | 0.862 | -0.000 | 0.038 | 0.901 |
| 4 | 1 | -0.017 | 0.098 | 0.803 | 0.000 | 0.039 | 0.898 |
| | 2 | 0.005 | 0.099 | 0.863 | 0.000 | 0.039 | 0.898 |
| | 3 | 0.009 | 0.162 | 0.867 | 0.000 | 0.039 | 0.898 |
| | 4 | 0.014 | 0.071 | 0.841 | 0.000 | 0.039 | 0.898 |
| | 5 | -0.005 | 0.090 | 0.834 | 0.000 | 0.039 | 0.898 |
| | 6 | -0.016 | 0.163 | 0.860 | 0.000 | 0.039 | 0.898 |
| 5 | 1 | -0.017 | 0.098 | 0.803 | 0.000 | 0.060 | 0.889 |
| | 2 | 0.005 | 0.100 | 0.860 | 0.000 | 0.060 | 0.889 |
| | 3 | 0.009 | 0.163 | 0.865 | 0.000 | 0.060 | 0.889 |
| | 4 | 0.014 | 0.072 | 0.839 | 0.000 | 0.060 | 0.889 |
| | 5 | -0.005 | 0.090 | 0.831 | 0.000 | 0.060 | 0.889 |
| | 6 | -0.016 | 0.163 | 0.857 | 0.000 | 0.060 | 0.889 |
| 6 | 1 | -0.018 | 0.097 | 0.805 | 0.000 | 0.039 | 0.901 |
| | 2 | 0.004 | 0.098 | 0.864 | 0.000 | 0.039 | 0.901 |
| | 3 | 0.007 | 0.161 | 0.871 | 0.000 | 0.039 | 0.901 |
| | 4 | 0.013 | 0.071 | 0.843 | 0.000 | 0.039 | 0.901 |
| | 5 | -0.006 | 0.089 | 0.834 | 0.000 | 0.039 | 0.901 |
| | 6 | -0.018 | 0.161 | 0.863 | 0.000 | 0.039 | 0.901 |

TABLE 5.16: The prefixes Eff and Tox denote the estimates of the probability of efficacy and toxicity under diffuse priors. The suffix Bias denotes bias; EmpSE denotes the empirical standard error; and Cov denotes 90% credible interval coverage.

5.17. The diffuse priors produce the most variable estimates and coverage suffers as a result. Bias is greater under our regularising and informative priors, but only noteworthy in an isolated number of instances. The cases of noteworthy bias can be understood with reference to the priors or the sub-model forms.

| Priors | EffBias | EffEmpSE | EffCov | ToxBias | ToxEmpSE | ToxCov |
|---|---|---|---|---|---|---|
| Sceptical | -0.010 | 0.096 | 0.881 | -0.001 | 0.044 | 0.905 |
| Informative | 0.004 | 0.097 | 0.871 | -0.002 | 0.043 | 0.903 |
| Diffuse | -0.001 | 0.113 | 0.840 | 0.000 | 0.046 | 0.896 |

TABLE 5.17: Means of numerical performance measures from Tables 5.14, 5.15 and 5.16 to 3 d.p.

## 5.5 Discussion

The proposed P2TNE design has many benefits.

Firstly, it makes efficient use of the available information because the predictive variables contribute to the modelling of the trial outcomes and ultimately to the approval decision. The effect of each variable is refined by regression. By adjusting for these sources of variability that are predictive of patient outcomes, the trial analysis gains in accuracy. We see this when comparing the P2TNE model to a simple beta-binomial alternative that makes decisions cohort by cohort.

Another key feature is that this design allows an acceptance / rejection decision for each permutation of the predictive variables via (5.12). Thus, it is feasible to approve the treatment in only the cohorts where it is shown to work. Without this facility, the undesirable risk is that the treatment is approved in cohorts where it is not appropriate or the treatment is rejected universally because the poor performance in some cohorts obscures the good performance in others. For instance, Thatcher [95] studied the effect on survival of Gefitinib in non-small cell lung cancer patients. Overall, they found that the treatment was not associated with a significant improvement in survival in the general population but that there was pronounced heterogeneity in survival in patient subgroups. In particular, there was evidence of benefit in patients of Asian origin and in those that have never smoked. Ultimately, it was determined that EGFR mutation was the factor that predicted benefit of the drug. If these predictive factors are known or even just suspected a priori, it is advantageous to be able to incorporate this information and retain the ability to tailor the acceptance / rejection decision using predictive variables. P2TNE provides this facility.

P2TNE explicitly models the association between the efficacy and toxicity outcomes. In real trials, as with PePS2, it is too simplistic to assume that efficacy and

toxicity are independent because severe toxicity partially precludes the scope for therapeutic benefit. In a model with co-primary outcomes of efficacy and toxicity, it is desirable that this important relationship be modelled. Asserting that there is no relationship (either explicitly or implicitly) risks spurious inference. We saw that the performance of P2TNE does not degrade when efficacy and toxicity are associated. We should stress that the level of association in our model is assumed to be constant amongst the cohorts. A more general (and complicated) model might allow each cohort to have its own association parameter but we do not consider that scenario here. In the next chapter, we look at removing the association parameter.

P2TNE is Bayesian and thus admits prior information. In a clinical trial, we usually want the data to speak for itself. However, in phase II trials with limited time and patients, we can gain efficiency by incorporating prior information. We saw here the benefit to using informative priors because overall the performance of the design was enhanced compared to the regularising and diffuse alternatives. However, each of the priors considered was sufficiently weak to be overwhelmed by the information in the data when the trial decision was clear as in scenarios 1 and 2.

Efficacy is seen in Garon to increase with PD-L1 score. It is a limitation of our analysis that we have implemented PD-L1 status as a categorical variable rather than an ordinal one. There are a number of ways that an ordinal PD-L1 variable could have been used in our model, each amounting to fixing the sign of coefficients in the efficacy sub-model. The signs of parameters can be fixed using exponential transforms or priors. For instance, a Gamma prior does not admit negative values, effectively guaranteeing a positive posterior estimate. We do not investigate this further here. More generally, we investigate the use of continuous PD-L1 in the next chapter.

In PePS2, we have not sought to model how the rate of toxicity might vary from cohort to cohort. We have omitted this potential complexity because we do not expect it to manifest in our clinical setting and expect no reward for the extra computational burden. However, the labels “efficacy" and “toxicity" are arbitrary so cohort-varying toxicity could easily be achieved in the same manner we have analysed efficacy here. In general, the principle of parsimony suggests not including too many parameters in $\boldsymbol{\theta}$. However, cohort-varying toxicity could easily be incorporated via

extra terms in (5.13) and $\boldsymbol{\theta}$ in a more fully-specified model. In the following chapter, we consider a fully-specified model that allows cohort-varying efficacy and toxicity and how a Bayesian information criterion can be used to choose amongst our default 'modestly-specified' model and the fully-specified model.

We have not considered an interim analysis in the simulation study because it is not required in our trial. The expected cohort size at final analysis is already small at 60 / 6 = 10 patients. Nevertheless, the P2TNE design easily facilitates an arbitrary number of interim analyses by repeated invocations of (5.12), potentially with different values for $\pi_E^*, \pi_T^*, p_E$ and $p_T$. Indeed, in a larger trial than we have considered, interim analyses would be preferable to allow early rejection of treatments that are inactive or excessively toxic in certain cohorts. If interim analyses are used, the statistician should choose values of $p_E$ and $p_T$ mindful of the effect of repeated testing, that lead to attractive operating characteristics overall.

Naturally, this design presents its challenges too. In some scenarios, the sharing of information could lead to questionable behaviour. For instance, when the P2TNE model specified observes requisite efficacies across the cohorts, it sometimes approves the treatment in a cohort that happened to observe zero efficacy events. When this happens in cohort $l$, the number of responses observed in cohorts $i = 1, .., 6, i \neq l$ compensate the lack of efficacy in $l$ to nevertheless yield a view that the true efficacy rate in $l$ is probably greater than $\pi_E^* = 0.10$. This is less likely to occur for larger values of $\pi_E^*$. If we wish to avoid this behaviour, we can use more stringent criteria than (5.12). For example, we could additionally require that the number of efficacies in a cohort must be greater than zero for the treatment to be approved.

Our P2TNE model is implemented in the Bayesian statistical language *Stan*[22] and made available in the trialr package[16] of Bayesian clinical trial designs. It uses Hamiltonian Monte Carlo to obtain samples from the posterior distributions. The calculations are reasonably demanding and a computer simulation takes approximately 3-4 hours to perform 10,000 iterations. More complicated specifications with more parameters will likely take longer.

It is problematic that there is no way to determine the required sample size without running computer simulations. A pragmatic solution to this specific problem is

to calculate a initial estimate of sample size using something like a Bryant & Day design or simple beta-binomial models and refining as the situation demands. Sample size, $p_E$ and $p_T$ are chosen to achieve acceptable operating characteristics. If truly predictive variables are introduced, the performance of P2TNE should be superior to the beta-binomial method, as demonstrated, and this will improve the statistical efficiency for the selected sample size..

When considering the predictive variables to include, there might appear to be a problem of circularity. It could be considered unreasonable to expect trialists to have knowledge of predictive variables at the start of a trial. Whilst this is sometimes true, often it is not. Trials are inherently sequential, each building on what is already known. P2TNE is a phase II design and phase II trials build on the results garnered in other early phase trials. For instance, we believe PD-L1 score to be predictive in PePS2 because it was demonstrated so by Garon *et al.*[37] in a closely-related patient population. However, this remains to be demonstrated in the PS2 population and this is the purpose of our trial.

Lastly, selecting sensible, modestly-informative joint priors is not a trivial task. Thall *et al.*[94] provide a general method for equating the amount of information in a multivariate normal prior to an hypothetical equivalent number of patient observations, a quantity they call the *effective sample size*. Priors can be as informative as the existing data allows, but sufficiently vague to justify the clinical trial under consideration.

In summary, we feel that the many benefits provided by P2TNE are attractive enough to warrant the effort to overcome the challenges, as we have in PePS2.

### 5.5.1 Further Development

Our predictive variables in PePS2 are binary. We have not considered a continuous predictive variable in PePS2 but the method described here needs only minimal changes to accommodate it. This makes sense in scenarios like PePS2 where the major predictive variable is a categorised mapping of an underlying continuous variable, PD-L1 score. It is likely that information is lost in the dichotomisation process and that using the continuous PD-L1 variable in 5.3 enhances the performance of the design. As with the binary variables, the effect of a continuous predictive variable $x_{.i}$

would be governed by coefficient $\eta$, for instance, in $\boldsymbol{\theta}$. The posterior mean of $\exp(\eta)$ would be the odds ratio for an event given a one unit increase in $x_{.i}$. One potential complication is that (5.12) would need to be resolved for each distinct value of $x_{.i}$. For a truly continuous explanatory variable, this would be the same as the number of patients. Although this may sound prohibitively costly, the posterior parameter samples provided by Stan make this trivial. We develop this idea in the following chapter.

We have presented P2TNE in a single arm setting but there is no reason why it could not be immediately applied in a multi-arm trial using dummy variables in (5.3) and (5.4) to reflect allocation to treatment arms. We have discussed the problem of having many components in $\boldsymbol{\theta}$ and adding variables for treatment arms would seem to exacerbate that problem. However, the inclusion of randomisation would abrogate the need for some other explanatory markers. For instance, if PePS2 were a randomised trial, we would have less need to include pre-treatment status as a predictive variable because the proportions of previously treated patients would be broadly balanced between the treatment arms. In a randomised controlled trial, the decision criteria (5.12) would instead accept the experimental treatment if it is likely that efficacy and/or toxicity are superior (or not-inferior) to the reference treatment, a posteriori.

Lastly, P2TNE uses a binary efficacy outcome that is dichotomised from the underlying continuous tumour size ratio variable. Wason *et al.*[100–102] have shown that the efficiency of clinical trials can be significantly increased by using all the information in a continuous response variable. In place of RECIST[35], a preferable design would use tumour size ratio and binary variables for non-shrinkage failures (e.g. the appearance of new lesions) to measure efficacy. This would require an analogue of (2.4) in the continuous setting.

## 5.6 Conclusions

It is a tremendous advantage to be able to tailor the clinical trial decision to patient cohorts where there is evidence that efficacy and/or toxicity is associated with predictive variables, especially when separate trials in cohorts are infeasible. This is

one of the primary goals of stratified medicine. The design presented, P2TNE, satisfies an unmet need by incorporating predictive information to jointly model efficacy and toxicity and selectively approve a treatment only in the patient cohorts where it is shown to be efficacious and tolerable, a posteriori. We demonstrate the method in the context of PePS2, a phase II trial of pembrolizumab in performance status 2 patients with non-small cell lung cancer. Our predictive variables are PD-L1 expression level and pretreatment status. The model described is flexible enough to admit arbitrary binary and continuous predictive variables. We demonstrate that model performance is good across a wide range of scenarios. Key to this is that P2TNE shares information across related cohorts to improve statistical performance. In contrast, benchmark beta-binomial designs that operate on cohorts singly perform relatively poorly because they use the available information less efficiently. In PePS2, P2TNE has allowed us to avoid the unappealing prospect of running parallel trials in cohorts. The main limitation of our method is that it is computationally intensive. The P2TNE phase II design provides researchers with an early opportunity to evaluate potentially predictive variables for stratified medicine that can be important to better inform phase III trials and future treatments for patients.

# Chapter 6

# Further Embellishments to the Statistical Design in PePS2

**Background:** We noted various assumptions in our P2TNE efficacy and toxicity sub-models used in Chapter 5 that had the potential to yield poor inferences. Furthermore, we observed that our model could potentially be simplified by removing the association parameter.

**Notable methods in this chapter:** We investigated various adaptations to the models proposed for PePS2, including more complexity in the efficacy model, and the facility of cohort-varying toxicity.

**The implications on efficiency:** We learned that additional parameters in the efficacy and toxicity sub-models require greater trial sample size to maintain the statistical performance presented in Chapter 5. Competing goals exist to conduct trials quickly and accurately, and the motivation to meet the extra resource burden of a more complex model should be appraised in the clinical context, in light of existing information and alternative treatments. In PePS2, the modest enhancement to inference in peripheral areas is unlikely to warrant recruiting at least 40 extra patients and materially delaying the dissemination of trial results.

## 6.1 Introduction

In the previous chapter, we presented a novel adaptation of the TNE trial design[89] for studying associated co-primary binary outcomes in the presence of predictive covariates. We did this in the context of PePS2, a phase II trial of pembrolizumab in

performance status 2 non-small-cell lung cancer patients. Our co-primary outcomes were efficacy and toxicity. Our predictive variables were PD-L1 score (Low, Medium or High) and pre-treatment status (TN or PT), each categorical in nature. We demonstrated that the P2TNE method with predictive baseline covariates was far more efficient than a model-free technique that used a conjugate beta-binomial approach in cohorts separately. Overall, we demonstrated that typical phase II error rates were possible in six cohorts using only 60 patients in total. Nevertheless, we noted a number of limitations in our chosen model specifications, (5.13). Specifically, our efficacy model lacked interactions and our toxicity model was very simplistic, containing only an intercept term. Our model choices were motivated by the published data, and sought a balance of efficiency and realism. We anticipate that more elaborate models with extra parameters will be more flexible but will require greater sample sizes. In this chapter, we consider the impact of embellishments to the model forms, specifically with respect to the trade-off between performance and sample size.

The lack of interactions in the efficacy model in (5.13) effectively assumes that the log-odds of efficacy in each PD-L1 cohort for PT patients was a common linear shift of that for TN patients, an assumption we referred to as *piecewise parallelism*. This assumption was pertinent because the data presented by Garon *et al.*[37] suggested that the log-odds of objective response in a closely-related patient population were perhaps not strictly piecewise parallel with respect to these covariates, as shown in Figure 5.1. In Section 6.2, we relax this assumption by adding interaction terms to our efficacy model.

Our toxicity model too was simplistic in that it assumed a common probability of toxicity in all cohorts, despite the acknowledgements that efficacy varied and toxicity and efficacy were plausibly associated. Our justification, again, was the published data[37, 44], collectively reporting the outcomes of over 1,000 NSCLC patients on pembrolizumab monotherapy, without noting toxicity that varied with pretreatedness or PD-L1 status. This does not, however, rule out heterogeneity in toxicity in the lower performance status population in PePS2. Furthermore, in other trial scenarios, the information to anticipate homogeneity in toxicity may not be available. In Section 6.3, we examine the effect on overall design performance by allowing toxicity to vary across cohorts by adding extra terms to the toxicity model.

TNE use a bivariate model that associates efficacy and toxicity outcomes. They consider the Gumbel model and a Gaussian copula. We use the Gumbel model (2.4) with association parameter $\psi$. We demonstrated in Table 5.13 that model performance barely differs when comparing scenarios that vary only in the presence of a strong negative association between efficacy and toxicity. This questions the usefulness of the association parameter. In Section 6.4, we examine the effect of removing it.

Finally, we have hitherto treated PD-L1 as a categorical variable because the three PD-L1 cohorts were defined and validated in Garon *et al.*[37]. This is perhaps regrettable, given that PD-L1 proportion score, the variable that determines PD-L1 category, is effectively continuous on $[0, 1]$. Dichotomising a continuous variable leads to loss of efficiency in analysis[2]. In this regard, we are motivated to research how the continuous PD-L1 proportion score, hitherto referred to as *PD-L1 score*, can be used in place of the categorised alternative variable to further enhance efficiency in the PePS2 setting. This ongoing work is introduced in Appendix D.

## 6.2 Interaction-terms in the efficacy model

Adding interaction terms for PD-L1 and pre-treatedness to the marginal efficacy model yields:

$$\begin{aligned} \text{logit}\, \pi_E(x_i, \boldsymbol{\theta}) &= \alpha + \beta x_{1i} + \gamma x_{2i} + \zeta x_{3i} + \eta x_{1i} x_{2i} + \kappa x_{1i} x_{3i} \\ \text{logit}\, \pi_T(x_i, \boldsymbol{\theta}) &= \lambda \end{aligned} \tag{6.1}$$

As before, $x_{1i}, ..., x_{3i}$ are the baseline covariates for patient $i$ as described in Table 5.3, and $\boldsymbol{\theta}$ is the vector of all parameters in the model. Here, $\boldsymbol{\theta} = (\alpha, \beta, \gamma, \zeta, \eta, \kappa, \lambda, \psi)$, and (2.4) is used to model the joint probability of $\pi_E$ and $\pi_T$.

We refer to this as *model 611*, because there are 6 parameters in the efficacy model, one in the toxicity model, and a single extra parameter for association in the joint model. Using this nomenclature, the model in the previous chapter is *model 411*.

Table 6.1 shows the operating performance of model 611 in our previous scenarios 1, 4 and 5, with increasing sample size, hitherto italicised and referred to simply

| Sc | TotN | Coh | PrEff | PrTox | N | Eff | Tox | Diffuse | EffBias | EffEmpSE | EffCov | ToxBias | ToxEmpSE | ToxCov |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 60 | 1 | 0.300 | 0.1 | 9.3 | 2.8 | 0.9 | 0.754 | 0.001 | 0.162 | 0.836 | 0.000 | 0.039 | 0.892 |
| | | 2 | 0.300 | 0.1 | 13.1 | 3.9 | 1.3 | 0.839 | -0.002 | 0.132 | 0.856 | 0.000 | 0.039 | 0.892 |
| | | 3 | 0.300 | 0.1 | 7.5 | 2.3 | 0.8 | 0.716 | 0.006 | 0.171 | 0.839 | 0.000 | 0.039 | 0.892 |
| | | 4 | 0.300 | 0.1 | 12.5 | 3.7 | 1.2 | 0.826 | -0.002 | 0.136 | 0.862 | 0.000 | 0.039 | 0.892 |
| | | 5 | 0.300 | 0.1 | 10.8 | 3.2 | 1.1 | 0.797 | -0.000 | 0.147 | 0.852 | 0.000 | 0.039 | 0.892 |
| | | 6 | 0.300 | 0.1 | 6.8 | 2.0 | 0.7 | 0.683 | 0.003 | 0.185 | 0.816 | 0.000 | 0.039 | 0.892 |
| 1 | 80 | 1 | 0.300 | 0.1 | 12.6 | 3.8 | 1.2 | 0.829 | -0.001 | 0.138 | 0.856 | 0.000 | 0.034 | 0.866 |
| | | 2 | 0.300 | 0.1 | 17.4 | 5.2 | 1.7 | 0.904 | 0.001 | 0.114 | 0.875 | 0.000 | 0.034 | 0.866 |
| | | 3 | 0.300 | 0.1 | 9.9 | 3.0 | 1.0 | 0.788 | 0.003 | 0.151 | 0.856 | 0.000 | 0.034 | 0.866 |
| | | 4 | 0.300 | 0.1 | 16.5 | 4.9 | 1.7 | 0.886 | -0.001 | 0.118 | 0.871 | 0.000 | 0.034 | 0.866 |
| | | 5 | 0.300 | 0.1 | 14.4 | 4.3 | 1.4 | 0.859 | -0.003 | 0.127 | 0.861 | 0.000 | 0.034 | 0.866 |
| | | 6 | 0.300 | 0.1 | 9.1 | 2.7 | 0.9 | 0.753 | 0.001 | 0.161 | 0.843 | 0.000 | 0.034 | 0.866 |
| 1 | 100 | 1 | 0.300 | 0.1 | 15.6 | 4.7 | 1.6 | 0.879 | 0.000 | 0.123 | 0.865 | 0.001 | 0.030 | 0.898 |
| | | 2 | 0.300 | 0.1 | 21.9 | 6.5 | 2.2 | 0.936 | -0.001 | 0.101 | 0.878 | 0.001 | 0.030 | 0.898 |
| | | 3 | 0.300 | 0.1 | 12.4 | 3.7 | 1.3 | 0.833 | 0.002 | 0.137 | 0.862 | 0.001 | 0.030 | 0.898 |
| | | 4 | 0.300 | 0.1 | 20.7 | 6.2 | 2.1 | 0.927 | -0.001 | 0.105 | 0.877 | 0.001 | 0.030 | 0.898 |
| | | 5 | 0.300 | 0.1 | 18.0 | 5.4 | 1.8 | 0.906 | 0.002 | 0.113 | 0.873 | 0.001 | 0.030 | 0.898 |
| | | 6 | 0.300 | 0.1 | 11.4 | 3.4 | 1.1 | 0.808 | 0.001 | 0.144 | 0.859 | 0.001 | 0.030 | 0.898 |
| 4 | 60 | 1 | 0.167 | 0.1 | 9.3 | 1.5 | 0.9 | 0.406 | 0.000 | 0.130 | 0.751 | 0.000 | 0.039 | 0.890 |
| | | 2 | 0.192 | 0.1 | 13.1 | 2.5 | 1.3 | 0.560 | -0.001 | 0.112 | 0.847 | 0.000 | 0.039 | 0.890 |
| | | 3 | 0.500 | 0.1 | 7.5 | 3.8 | 0.8 | 0.929 | -0.003 | 0.190 | 0.854 | 0.000 | 0.039 | 0.890 |
| | | 4 | 0.091 | 0.1 | 12.5 | 1.1 | 1.3 | 0.175 | 0.001 | 0.083 | 0.646 | 0.000 | 0.039 | 0.890 |
| | | 5 | 0.156 | 0.1 | 10.8 | 1.7 | 1.1 | 0.408 | 0.001 | 0.114 | 0.773 | 0.000 | 0.039 | 0.890 |
| | | 6 | 0.439 | 0.1 | 6.8 | 3.0 | 0.7 | 0.863 | -0.005 | 0.202 | 0.840 | 0.000 | 0.039 | 0.890 |
| 4 | 80 | 1 | 0.167 | 0.1 | 12.6 | 2.1 | 1.3 | 0.469 | 0.001 | 0.110 | 0.810 | 0.000 | 0.033 | 0.866 |
| | | 2 | 0.192 | 0.1 | 17.4 | 3.4 | 1.7 | 0.640 | 0.002 | 0.098 | 0.855 | 0.000 | 0.033 | 0.866 |
| | | 3 | 0.500 | 0.1 | 9.9 | 5.0 | 1.0 | 0.962 | -0.003 | 0.167 | 0.869 | 0.000 | 0.033 | 0.866 |
| | | 4 | 0.091 | 0.1 | 16.5 | 1.5 | 1.6 | 0.176 | 0.001 | 0.072 | 0.738 | 0.000 | 0.033 | 0.866 |
| | | 5 | 0.156 | 0.1 | 14.4 | 2.2 | 1.4 | 0.444 | -0.000 | 0.100 | 0.826 | 0.000 | 0.033 | 0.866 |
| | | 6 | 0.439 | 0.1 | 9.1 | 4.0 | 0.9 | 0.917 | -0.003 | 0.177 | 0.856 | 0.000 | 0.033 | 0.866 |
| 4 | 180 | 1 | 0.167 | 0.1 | 28.4 | 4.7 | 2.8 | 0.637 | 0.001 | 0.073 | 0.864 | 0.000 | 0.022 | 0.875 |
| | | 2 | 0.192 | 0.1 | 39.3 | 7.5 | 3.9 | 0.819 | -0.000 | 0.065 | 0.879 | 0.000 | 0.022 | 0.875 |
| | | 3 | 0.500 | 0.1 | 22.3 | 11.2 | 2.2 | 0.999 | -0.001 | 0.112 | 0.879 | 0.000 | 0.022 | 0.875 |
| | | 4 | 0.091 | 0.1 | 37.2 | 3.4 | 3.7 | 0.189 | 0.000 | 0.049 | 0.856 | 0.000 | 0.022 | 0.875 |
| | | 5 | 0.156 | 0.1 | 32.2 | 5.0 | 3.2 | 0.595 | 0.000 | 0.066 | 0.869 | 0.000 | 0.022 | 0.875 |
| | | 6 | 0.439 | 0.1 | 20.6 | 9.1 | 2.0 | 0.991 | -0.001 | 0.116 | 0.880 | 0.000 | 0.022 | 0.875 |
| 4 | 300 | 1 | 0.167 | 0.1 | 47.5 | 7.8 | 4.7 | 0.749 | -0.002 | 0.053 | 0.903 | -0.000 | 0.018 | 0.868 |
| | | 2 | 0.192 | 0.1 | 65.6 | 12.6 | 6.5 | 0.934 | -0.000 | 0.047 | 0.898 | -0.000 | 0.018 | 0.868 |
| | | 3 | 0.500 | 0.1 | 36.7 | 18.4 | 3.7 | 1.000 | 0.001 | 0.091 | 0.872 | -0.000 | 0.018 | 0.868 |
| | | 4 | 0.091 | 0.1 | 62.0 | 5.6 | 6.2 | 0.157 | -0.000 | 0.036 | 0.884 | -0.000 | 0.018 | 0.868 |
| | | 5 | 0.156 | 0.1 | 53.9 | 8.5 | 5.4 | 0.738 | 0.002 | 0.050 | 0.889 | -0.000 | 0.018 | 0.868 |
| | | 6 | 0.439 | 0.1 | 34.4 | 15.0 | 3.3 | 1.000 | -0.003 | 0.092 | 0.886 | -0.000 | 0.018 | 0.868 |
| 5 | 60 | 1 | 0.167 | 0.3 | 9.3 | 1.5 | 2.8 | 0.041 | 0.000 | 0.130 | 0.752 | 0.000 | 0.060 | 0.887 |
| | | 2 | 0.192 | 0.3 | 13.1 | 2.5 | 3.9 | 0.059 | -0.001 | 0.112 | 0.844 | 0.000 | 0.060 | 0.887 |
| | | 3 | 0.500 | 0.3 | 7.5 | 3.8 | 2.3 | 0.099 | -0.003 | 0.190 | 0.849 | 0.000 | 0.060 | 0.887 |
| | | 4 | 0.091 | 0.3 | 12.5 | 1.1 | 3.7 | 0.018 | 0.001 | 0.084 | 0.648 | 0.000 | 0.060 | 0.887 |
| | | 5 | 0.156 | 0.3 | 10.8 | 1.7 | 3.2 | 0.044 | 0.001 | 0.115 | 0.772 | 0.000 | 0.060 | 0.887 |
| | | 6 | 0.439 | 0.3 | 6.8 | 3.0 | 2.0 | 0.093 | -0.005 | 0.202 | 0.836 | 0.000 | 0.060 | 0.887 |
| 5 | 180 | 1 | 0.167 | 0.3 | 28.4 | 4.7 | 8.5 | 0.071 | 0.001 | 0.073 | 0.864 | -0.000 | 0.034 | 0.891 |
| | | 2 | 0.192 | 0.3 | 39.3 | 7.5 | 11.8 | 0.092 | -0.001 | 0.065 | 0.878 | -0.000 | 0.034 | 0.891 |
| | | 3 | 0.500 | 0.3 | 22.3 | 11.2 | 6.7 | 0.111 | -0.001 | 0.113 | 0.876 | -0.000 | 0.034 | 0.891 |
| | | 4 | 0.091 | 0.3 | 37.2 | 3.4 | 11.2 | 0.022 | 0.000 | 0.049 | 0.854 | -0.000 | 0.034 | 0.891 |
| | | 5 | 0.156 | 0.3 | 32.2 | 5.0 | 9.6 | 0.066 | 0.000 | 0.066 | 0.868 | -0.000 | 0.034 | 0.891 |
| | | 6 | 0.439 | 0.3 | 20.6 | 9.1 | 6.2 | 0.110 | -0.001 | 0.116 | 0.877 | -0.000 | 0.034 | 0.891 |

TABLE 6.1: Operating performance of model 611 in selected scenarios from Chapter 5 using increasing total sample sizes, $TotN$. Columns Sc to Tox have the same definitions as Table 5.3. Diffuse shows approval probability under diffuse priors. Columns EffBias to ToxCov have the same definitions as Table 5.14.

as $N$. For parsimony, we do not show performance in all scenarios. Scenario 1 shows performance in our benchmark scenario analogous to an analysis of power. Scenarios 4 and 5 illustrate performance in plausible efficacy scenarios where toxicity is acceptable and not, respectively. Scenarios 3 and 6 were excluded because Table 5.13 revealed that they add little beyond scenarios 1 and 4. Scenario 2 was excluded for

brevity but simulations revealed that the false approval rate was not increased. Cohort memberships are simulated using the same method described in Section 5.4.1.

Diffuse $N(0, 10^2)$ priors on each element of $\boldsymbol{\theta}$ were used in Table 6.1. We did this so that the performance mainly reflects the observed data and the specified model forms, and not the prior information. As such, we compare to the 'Diffuse' performance under model 411 in Table 5.13. This is purely to aid comparison. Were we to use model 611 in a trial, we would specify prior distributions less diffuse than $N(0, 10^2)$.

We see in Table 6.1 that the two extra parameters are associated with a reduction in the probability of approving the treatment with $N = 60$. The probabilities of approval fall to 68.3 - 83.8% in scenario 1. These are absolute reductions of 6 - 13% compared to model 411, but still compare favourably to the performance of the beta-binomial analyses with $n = 60$ in Table 5.13. The extra parameters have reduced performance because there are now more ways the model can err when fitting the data. There is also less opportunity to borrow information across cohorts in the efficacy model. We see that bias is not generally a problem. However, comparing to Table 5.16, efficacy coverage of the 90% CIs has now fallen below 85% in three of the six cohorts. Furthermore, empirical standard error has increased in each cohort by approximately 2 to 4%, in absolute terms. Standard error of efficacy estimates is particularly high in the small high-PD-L1 cohorts.

Generally, models that estimate more parameters require a greater sample size than models that estimate few parameters. As $N$ increases to 100 in scenario 1, the probability of approval is generally within a few percent of that under model 411 and $N = 60$, with performance slightly better in some cohorts and slightly worse in others. The performance of the beta-binomial models also improves as $n$ increases (data not shown), but the performance of 611 is always superior. Likewise, the empirical standard errors of efficacy estimates have fallen to approximately the same levels, and coverages are now all between 85% and 90%. These qualities were not achieved with $N = 80$. We may regard the extra 40 patients required by model 611 as the approximate cost of supporting 2 extra parameters in the efficacy model in this scenario and maintaining a similar level of operating performance and accuracy.

In scenario 4, the efficacy probabilities show a more plausible relationship with

the baseline covariates. The observations highlighted above are again evident. At $N = 60$, the probabilities of approval have generally fallen because the model is under-informed. The efficacy coverages are particularly low in the small cohorts. Once again, approval probabilities and coverages improve as $N$ increases. To match the approval probability of model 411 in the medium and high PD-L1 cohorts, a sample size of $N = 80$ is approximately sufficient in this scenario. However, even with this increase in $N$, we see that efficacy coverage is still poor.

As Figure 5.1 shows, Scenario 4 contains an interaction between PD-L1 and pre-treatedness. The correct decision in cohort 4 is to reject the treatment because the efficacy rate is marginally below the minimum threshold, 10%. The 411 model incorrectly approved with probability 21.5%. We attributed some of this failure to the high efficacy seen in other cohorts and the lack of interaction terms required to precisely estimate efficacy in this particular cohort. The approval rate is only slightly lower in model 611 with the interaction terms. Increasing $N$ from 60 to 180, we see that the approval probability actually increases from 17.5% to 18.9% despite the fact that coverage improves from 64.6% to 85.6% and empirical standard error falls. The interaction terms in 611 have not overcome one of the notable shortcomings of model 411, even with triple the sample size. Increasing sample size further to $N = 300$, the 90% efficacy coverage moves to over 88%, the variability of estimates falls, and the approval probability in cohort 4 falls to 15.7%. However, the benefits are very small. The approval decision is made with reference to the model-estimated rates of efficacy and toxicity using (5.12). With the interaction terms included, those efficacy and toxicity rates are estimated by (6.1).

Other authors have highlighted the great demands in sample size to estimate interaction effects. Schmoor *et al.*[80] demonstrate that at least four times the sample size is required to detect a prognostic effect via interaction of a binary covariate in a two-arm trial with a time-to-event outcome analysed by Cox model. The exact multiplier is affected by the nature and prevalence of the covariate, but this general result offers an insight into why model 611 performs relatively poorly despite a greatly-increased sample size.

In scenario 5, we see that model 611 correctly rejects approximately as often as model 411 with $N = 60$. However, coverage is particularly poor in some cohorts

and this is rectified by increased $N$. With sample size as high as $N = 180$, coverage is above 85% in all cohorts.

The unifying theme from these simulations is that greater sample size is required to estimate the extra parameters in model 611 and to have confidence in the inference it yields. We saw that 20 - 40 additional patients would allow this model to perform similarly to model 411 in scenarios 1 and 4. However, merely replicating what we had before will not provide justification. We saw that far more patients would be required for model 611 to provide sufficient additional accuracy to materially improve the extent to which the analysis makes the correct decision in cohort 4. The potential to justify this drives at the heart of this thesis. The biological plausibility of the interaction efficacy model is far from clear, and the data represented in Figure 5.1 support piecewise parallelism as a reasonable working assumption. In PePS2, the extra time, money and effort that would be required to recruit these patients in a phase II clinical trial would almost certainly not be warranted given the need to conduct trials quickly to give patients a chance of tolerable and effective treatments. Rather than increase the sample size fivefold, when efficacy under a targeted therapy can be so profoundly associated with a biomarker, it would clearly be preferable to run multiple trials of modest size in a wide variety of treatments.

In each of the scenarios considered hitherto, the toxicity coverage has been stable and accurate. However, there are no variables beyond an intercept in the toxicity model and there is no heterogeneity by cohort. We demand more from the toxicity model in the next section.

## 6.3 Covariate terms in the toxicity model

The extra terms in the efficacy model in the previous section did not materially improve operating performance but did incur material penalty in terms of greater required sample size. In this section, we revert to a four parameter efficacy model and add parameters to the toxicity model so that it has the same freedom to estimate a different toxicity rate in each cohort:

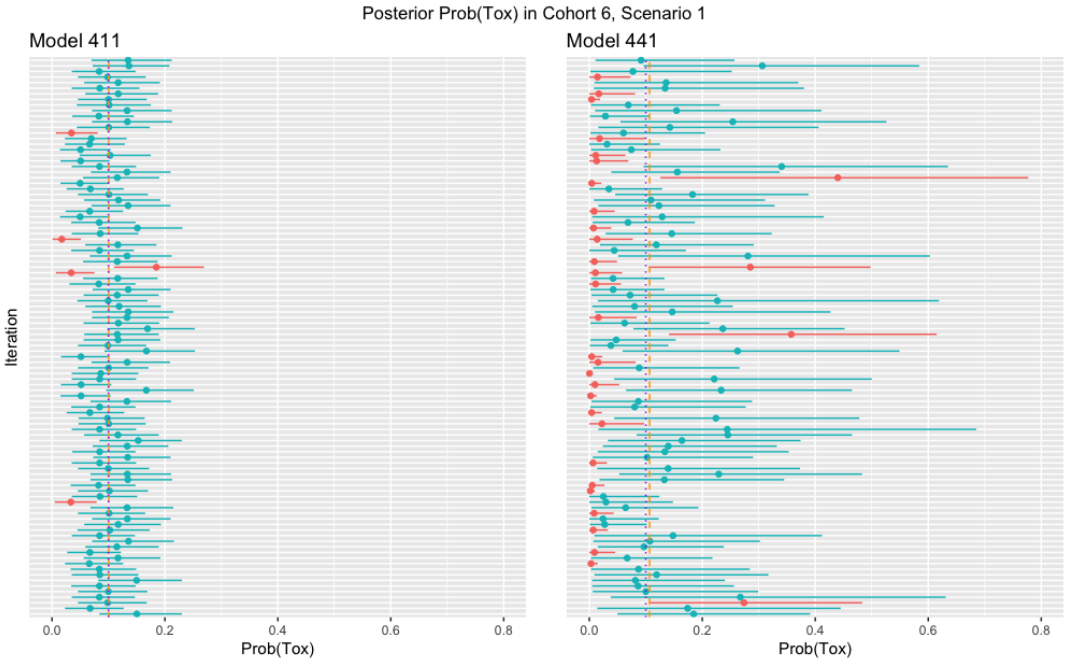



FIGURE 6.1: Posterior mean estimates of Prob(Tox) with 90% CIs for cohort 6 in scenario 1 using models 411 and 441 and a total $N = 60$. The red intervals exclude the true value, 10%, shown by purple dotted line.

$$\begin{aligned} \text{logit}\, \pi_E(x_i, \boldsymbol{\theta}) &= \alpha + \beta x_{1i} + \gamma x_{2i} + \zeta x_{3i} \\ \text{logit}\, \pi_T(x_i, \boldsymbol{\theta}) &= \lambda + \mu x_{1i} + \nu x_{2i} + \xi x_{3i} \end{aligned} \tag{6.2}$$

In this model that we will refer to as model 441, we are effectively assuming that the log-odds of efficacy and toxicity are each piecewise parallel across the cohorts.

Once again, an indicative subset of scenarios is shown and to aid comparability, diffuse $N(0, 10^2)$ priors on each element of $\boldsymbol{\theta}$ were used.

Table 6.2 shows that, as with model 611, the extra terms in model 441 have eroded operating performance when $N = 60$ such that we can no longer expect at least 80% approval probabilities in scenario 1. Operating performance improves with 40 extra patients but not enough to attain 80% approval in all cohorts. Performance still lags in the high PD-L1 cohorts, where the empirical standard error of the estimated efficacy probabilities is highest.

Coverage is particularly poor in the estimated toxicity probabilities. The extra terms in the toxicity model have increased the opportunity for the model to overfit

| Sc | TotN | Coh | PrEff | PrTox | N | Eff | Tox | Diffuse | EffBias | EffEmpSE | EffCov | ToxBias | ToxEmpSE | ToxCov |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 60 | 1 | 0.300 | 0.1 | 9.3 | 2.8 | 0.9 | 0.677 | 0.002 | 0.125 | 0.868 | 0.006 | 0.087 | 0.776 |
| | | 2 | 0.300 | 0.1 | 13.1 | 3.9 | 1.3 | 0.762 | -0.002 | 0.113 | 0.865 | -0.001 | 0.074 | 0.799 |
| | | 3 | 0.300 | 0.1 | 7.5 | 2.3 | 0.8 | 0.611 | 0.003 | 0.140 | 0.860 | 0.005 | 0.093 | 0.740 |
| | | 4 | 0.300 | 0.1 | 12.5 | 3.7 | 1.2 | 0.752 | -0.002 | 0.114 | 0.873 | -0.002 | 0.076 | 0.786 |
| | | 5 | 0.300 | 0.1 | 10.8 | 3.2 | 1.1 | 0.721 | 0.001 | 0.119 | 0.866 | 0.002 | 0.080 | 0.790 |
| | | 6 | 0.300 | 0.1 | 6.8 | 2.0 | 0.7 | 0.597 | 0.004 | 0.142 | 0.864 | 0.007 | 0.097 | 0.745 |
| 1 | 100 | 1 | 0.300 | 0.1 | 15.6 | 4.7 | 1.6 | 0.853 | 0.000 | 0.094 | 0.882 | 0.003 | 0.064 | 0.847 |
| | | 2 | 0.300 | 0.1 | 21.9 | 6.5 | 2.2 | 0.913 | -0.001 | 0.086 | 0.887 | -0.000 | 0.057 | 0.845 |
| | | 3 | 0.300 | 0.1 | 12.4 | 3.7 | 1.3 | 0.785 | 0.001 | 0.108 | 0.876 | 0.003 | 0.072 | 0.822 |
| | | 4 | 0.300 | 0.1 | 20.7 | 6.2 | 2.1 | 0.906 | -0.001 | 0.088 | 0.881 | -0.002 | 0.058 | 0.839 |
| | | 5 | 0.300 | 0.1 | 18.0 | 5.4 | 1.8 | 0.884 | 0.002 | 0.091 | 0.880 | 0.001 | 0.061 | 0.840 |
| | | 6 | 0.300 | 0.1 | 11.4 | 3.4 | 1.1 | 0.770 | 0.002 | 0.109 | 0.882 | 0.004 | 0.074 | 0.819 |
| 2 | 100 | 1 | 0.100 | 0.3 | 15.6 | 1.6 | 4.7 | 0.029 | 0.003 | 0.064 | 0.846 | 0.001 | 0.095 | 0.877 |
| | | 2 | 0.100 | 0.3 | 21.9 | 2.2 | 6.6 | 0.025 | -0.002 | 0.056 | 0.849 | 0.001 | 0.088 | 0.875 |
| | | 3 | 0.100 | 0.3 | 12.4 | 1.2 | 3.7 | 0.029 | 0.003 | 0.071 | 0.816 | 0.002 | 0.108 | 0.874 |
| | | 4 | 0.100 | 0.3 | 20.7 | 2.1 | 6.2 | 0.030 | -0.000 | 0.059 | 0.837 | -0.001 | 0.088 | 0.881 |
| | | 5 | 0.100 | 0.3 | 18.0 | 1.8 | 5.4 | 0.028 | 0.002 | 0.061 | 0.842 | 0.002 | 0.091 | 0.882 |
| | | 6 | 0.100 | 0.3 | 11.4 | 1.2 | 3.5 | 0.031 | 0.004 | 0.073 | 0.819 | 0.004 | 0.109 | 0.879 |
| 4 | 60 | 1 | 0.167 | 0.1 | 9.3 | 1.5 | 0.9 | 0.310 | -0.017 | 0.098 | 0.803 | 0.005 | 0.086 | 0.783 |
| | | 2 | 0.192 | 0.1 | 13.1 | 2.5 | 1.3 | 0.531 | 0.005 | 0.099 | 0.861 | -0.000 | 0.074 | 0.801 |
| | | 3 | 0.500 | 0.1 | 7.5 | 3.8 | 0.8 | 0.727 | 0.009 | 0.162 | 0.864 | 0.005 | 0.092 | 0.744 |
| | | 4 | 0.091 | 0.1 | 12.5 | 1.1 | 1.3 | 0.179 | 0.014 | 0.071 | 0.841 | -0.003 | 0.076 | 0.786 |
| | | 5 | 0.156 | 0.1 | 10.8 | 1.7 | 1.1 | 0.339 | -0.005 | 0.090 | 0.831 | 0.002 | 0.081 | 0.788 |
| | | 6 | 0.439 | 0.1 | 6.8 | 3.0 | 0.7 | 0.681 | -0.016 | 0.162 | 0.860 | 0.006 | 0.096 | 0.751 |
| 4 | 110 | 1 | 0.167 | 0.1 | 17.3 | 2.9 | 1.7 | 0.460 | -0.018 | 0.071 | 0.826 | 0.003 | 0.061 | 0.846 |
| | | 2 | 0.192 | 0.1 | 24.0 | 4.6 | 2.4 | 0.763 | 0.006 | 0.073 | 0.878 | -0.001 | 0.054 | 0.850 |
| | | 3 | 0.500 | 0.1 | 13.7 | 6.8 | 1.4 | 0.871 | 0.013 | 0.120 | 0.875 | 0.001 | 0.067 | 0.830 |
| | | 4 | 0.091 | 0.1 | 22.8 | 2.1 | 2.3 | 0.239 | 0.015 | 0.051 | 0.874 | -0.002 | 0.055 | 0.841 |
| | | 5 | 0.156 | 0.1 | 19.8 | 3.1 | 2.0 | 0.504 | -0.006 | 0.066 | 0.854 | 0.001 | 0.057 | 0.853 |
| | | 6 | 0.439 | 0.1 | 12.5 | 5.5 | 1.2 | 0.851 | -0.014 | 0.120 | 0.871 | 0.002 | 0.069 | 0.825 |
| 4 | 180 | 1 | 0.167 | 0.1 | 28.4 | 4.7 | 2.8 | 0.587 | -0.019 | 0.055 | 0.836 | 0.002 | 0.047 | 0.867 |
| | | 2 | 0.192 | 0.1 | 39.3 | 7.5 | 3.9 | 0.897 | 0.006 | 0.057 | 0.882 | -0.000 | 0.042 | 0.868 |
| | | 3 | 0.500 | 0.1 | 22.3 | 11.2 | 2.2 | 0.956 | 0.013 | 0.093 | 0.879 | -0.000 | 0.052 | 0.854 |
| | | 4 | 0.091 | 0.1 | 37.2 | 3.4 | 3.7 | 0.282 | 0.015 | 0.040 | 0.874 | -0.001 | 0.043 | 0.868 |
| | | 5 | 0.156 | 0.1 | 32.2 | 5.0 | 3.2 | 0.621 | -0.008 | 0.051 | 0.863 | 0.001 | 0.045 | 0.866 |
| | | 6 | 0.439 | 0.1 | 20.6 | 9.1 | 2.0 | 0.949 | -0.016 | 0.093 | 0.875 | 0.000 | 0.053 | 0.857 |
| 5 | 60 | 1 | 0.167 | 0.3 | 9.3 | 1.5 | 2.8 | 0.059 | -0.017 | 0.098 | 0.802 | 0.002 | 0.125 | 0.868 |
| | | 2 | 0.192 | 0.3 | 13.1 | 2.5 | 3.9 | 0.093 | 0.005 | 0.099 | 0.863 | -0.000 | 0.114 | 0.870 |
| | | 3 | 0.500 | 0.3 | 7.5 | 3.8 | 2.3 | 0.144 | 0.009 | 0.162 | 0.867 | 0.003 | 0.138 | 0.861 |
| | | 4 | 0.091 | 0.3 | 12.5 | 1.1 | 3.7 | 0.032 | 0.014 | 0.072 | 0.841 | -0.003 | 0.115 | 0.868 |
| | | 5 | 0.156 | 0.3 | 10.8 | 1.7 | 3.2 | 0.061 | -0.005 | 0.090 | 0.833 | 0.002 | 0.119 | 0.867 |
| | | 6 | 0.439 | 0.3 | 6.8 | 3.0 | 2.0 | 0.144 | -0.016 | 0.162 | 0.860 | 0.004 | 0.141 | 0.862 |
| 5 | 110 | 1 | 0.167 | 0.3 | 17.3 | 2.9 | 5.2 | 0.066 | -0.018 | 0.071 | 0.826 | 0.003 | 0.090 | 0.880 |
| | | 2 | 0.192 | 0.3 | 24.0 | 4.6 | 7.2 | 0.099 | 0.006 | 0.072 | 0.878 | -0.000 | 0.082 | 0.885 |
| | | 3 | 0.500 | 0.3 | 13.7 | 6.8 | 4.1 | 0.132 | 0.013 | 0.120 | 0.874 | 0.000 | 0.102 | 0.879 |
| | | 4 | 0.091 | 0.3 | 22.8 | 2.1 | 6.9 | 0.032 | 0.015 | 0.051 | 0.874 | 0.001 | 0.085 | 0.879 |
| | | 5 | 0.156 | 0.3 | 19.8 | 3.1 | 5.9 | 0.067 | -0.006 | 0.066 | 0.855 | 0.002 | 0.088 | 0.881 |
| | | 6 | 0.439 | 0.3 | 12.5 | 5.5 | 3.8 | 0.133 | -0.014 | 0.120 | 0.872 | 0.001 | 0.103 | 0.878 |
| 7 | 60 | 1 | 0.167 | 0.1 | 9.3 | 1.5 | 0.9 | 0.332 | -0.017 | 0.098 | 0.805 | 0.005 | 0.075 | 0.822 |
| | | 2 | 0.192 | 0.1 | 13.1 | 2.5 | 1.3 | 0.553 | 0.005 | 0.099 | 0.864 | 0.000 | 0.068 | 0.825 |
| | | 3 | 0.500 | 0.1 | 7.5 | 3.8 | 0.8 | 0.765 | 0.009 | 0.162 | 0.864 | 0.006 | 0.082 | 0.808 |
| | | 4 | 0.091 | 0.3 | 12.5 | 1.1 | 3.7 | 0.035 | 0.014 | 0.071 | 0.839 | -0.003 | 0.125 | 0.864 |
| | | 5 | 0.156 | 0.3 | 10.8 | 1.7 | 3.2 | 0.066 | -0.005 | 0.090 | 0.832 | 0.001 | 0.132 | 0.860 |
| | | 6 | 0.439 | 0.3 | 6.8 | 3.0 | 2.0 | 0.150 | -0.016 | 0.162 | 0.861 | 0.002 | 0.160 | 0.844 |
| 7 | 110 | 1 | 0.167 | 0.1 | 17.3 | 2.9 | 1.7 | 0.478 | -0.018 | 0.071 | 0.826 | 0.002 | 0.053 | 0.859 |
| | | 2 | 0.192 | 0.1 | 24.0 | 4.6 | 2.4 | 0.778 | 0.006 | 0.073 | 0.878 | -0.001 | 0.049 | 0.859 |
| | | 3 | 0.500 | 0.1 | 13.7 | 6.8 | 1.4 | 0.906 | 0.013 | 0.120 | 0.875 | 0.003 | 0.060 | 0.853 |
| | | 4 | 0.091 | 0.3 | 22.8 | 2.1 | 6.9 | 0.034 | 0.015 | 0.051 | 0.874 | 0.001 | 0.092 | 0.881 |
| | | 5 | 0.156 | 0.3 | 19.8 | 3.1 | 5.9 | 0.068 | -0.006 | 0.066 | 0.853 | 0.002 | 0.097 | 0.878 |
| | | 6 | 0.439 | 0.3 | 12.5 | 5.5 | 3.8 | 0.142 | -0.014 | 0.120 | 0.872 | -0.000 | 0.118 | 0.870 |

TABLE 6.2: Operating performance of model 441 in selected scenarios from Chapter 5, and the new scenario 7. Diffuse shows probability of approval under diffuse priors. Bias, SE and coverage columns retain the definitions already given.

chance occurrences at small sample sizes. Figure 6.1 illustrates this. The dots show the posterior mean probability of toxicity in cohort 6 of scenario 1 and the horizontal lines show 90% credible intervals. For clarity, we show only estimates from the first 100 simulated iterations. The vertical purple dotted lines show the true toxicity

probability, 10%, and the orange dashed lines show the mean of the posterior mean estimates. The blue intervals contain the true value but the red intervals do not. The left-hand panel shows those estimates by model 411 and the right-hand panel model 441. For model 411, the orange and purple lines are so close as to be barely distinguishable but there is a small amount of bias visible for model 441.

We see that the intervals for model 411 are generally narrower. This is not surprising because each uses the full sample size of 60 patients to estimate only the intercept parameter. In contrast, the extra parameters in model 441 admit much more uncertainty at this modest sample size. Incongruously, even though the CIs are generally wider in the 441 model, they are less likely to contain the true toxicity value. We see a relative abundance of very low estimates with unduly narrow CIs. The tenth iteration from the bottom is an extreme example. Having observed 0 toxicities in cohorts 4, 5 and 6 with sizes 14, 9 and 9 respectively, the posterior mean toxicity estimates are very low. The posterior mean probabilities of toxicity in these cohorts are all estimated to be less than 1% and 95th percentiles are each less than 4%. Faced with a chance occurrence, the model has produced parameter estimates that are not only erroneous, but also unjustifiably precise.

The priors are partly to blame. The horseshoe-shaped prior distributions on the event probability similar to those demonstrated in Figure 5.7, generated by the diffuse parameter priors, are having an excessively adverse effect. In the chance negative example identified above, the data agree strongly with the large prior mass placed close to the probability zero, pinning undue posterior mass to that boundary. The combined effect of the priors and logit likelihood is to have removed too much uncertainty from the posterior estimate. In contrast, a regularising effect could have been provided by modestly informative priors, like our regularising priors in Chapter 5, that prevent the model from over-fitting to chance events[64]. Unfortunately, this particular example illustrates how diffuse priors can be inadvertently informative, and reflects what a misnomer 'uninformative' can be.

Whilst the example highlighted above leads to an erroneously low estimate of toxicity, an error that Figure 6.1 shows is relatively common in model 441 compared to model 411 under diffuse priors, we see from the locations and frequency of the red lines that erroneously *high* estimates of toxicity are relatively common too. It

is the coverage and variability of the toxicity estimates that have become impaired, rather than the introduction of bias.

An alternative to changing the priors, naturally, is that we may recruit so many patients that our posterior beliefs are dominated by the likelihood. However, once again we are mindful that this thesis concerns itself with efficiency in clinical trials. It is expedient to see priors as part of the overall model, chosen with similar motivation as the likelihood function and the explanatory variables, to provide the best inference possible. Priors that promote pathological behaviour like that demonstrated above are a detriment to efficiency because they necessitate patients, who themselves necessitate time and money, in order to counteract the misleading information endowed in the posterior. Clinical trials provide a rare but costly opportunity to make a decision that will impact the lives of many patients. It is preferable to use priors that help the analysis to achieve its objectives. This does not mean forcing the model to produce estimates that we expect a-priori. It means producing estimates from the trial data that contain appropriate uncertainty, and stopping the model from over-fitting.

It is interesting in scenario 2 to examine the coverages of estimates. In this scenario, efficacy is low and toxicity is high in every cohort, and the correct decision is to reject throughout. Coverage is now relatively poor for efficacy and acceptable in toxicity. This is the opposite of what we saw in scenario 1, even though each submodel uses four parameters in each scenario. This suggests that bias is a greater risk in logistic models when the true event rate is close to 0 or 1, as contraction in the credible interval will occur at the boundary. As discussed above, this is exacerbated by the diffuse priors.

In scenario 4, we see that performance is again relatively poor at $N = 60$. In this scenario, raising $N$ as high as 110 yields approval probabilities at least as high as model 411 in cohorts 1, 2, 4 and 5. However, model 441 remains less likely to approve in the high PD-L1 cohorts, 3 and 6. The toxicity coverage remains low in these cohorts despite the comparatively high sample size. Overall sample size of the order of $N = 180$ is required to restore the approval probabilities in these cohorts close to those seen in model 411.

In scenario 5 with $N = 60$, model 441 inflates the probability of incorrectly approving in cohorts 3 and 6 by approximately 4%. Even with sample size as high as $N = 110$, this flaw is not completely rectified.

In Table 6.2, we have also added a new scenario 7 to test the discriminatory power of the toxicity model. It uses the same efficacy probabilities as scenarios 4 and 5, but the rate of toxicity is low in TN patients and high in PT patients. We see that the model is now very unlikely to approve in cohorts 4, 5, and 6, despite the high efficacy in cohort 6. We see that $N = 60$ does not yield our desired approval probability of 80% in cohort 3 but $N = 110$ raises the approval probability above 90%.

Overall, we have demonstrated that extra terms in the toxicity model will improve inference as expected when toxicity varies by cohort. However, the extra parameters demand a greater sample size. To reproduce in scenarios 1-6 the operating performance seen with model 411, up to $N = 180$ patients may be required. Once again, the appetite for satisfying this burden will be driven by the clinical scenario. Given the data already presented on pembrolizumab in a related patient group, the motivation will not exist in PePS2 to triple the sample size to use model 441.

The previous two sections have concerned adding parameters in the effort to produce a better model. Keeping all else constant, we saw that adding parameters increases the amount of uncertainty in a model and reduces the statistical efficiency of the method, a situation that can be rectified by recruiting more patients. This forces us to contemplate that fewer parameters might be appropriate if it increases statistical efficiency. We investigate that in the next section.

## 6.4 Removing the association between efficacy and toxicity

In the previous chapter, we saw evidence that questioned the benefit of modelling associated co-primary outcomes. Scenarios 3 and 6 in Table 5.13 simulated efficacy and toxicity events that were strongly negatively associated. These mirrored scenarios 1 and 4 respectively in every other regard with the exception that the outcomes in scenarios 1 and 4 were not associated, on average. Comparing scenario 1 to 3 and 4 to 6 in Table 5.13, model performance is practically unchanged. Here we investigate

further the benefit of $\psi$ in the joint model by considering an alternative model with no association parameter:

$$\begin{aligned} \text{logit}\, \pi_E(x_i, \boldsymbol{\theta}) &= \alpha + \beta x_{1i} + \gamma x_{2i} + \zeta x_{3i} \\ \text{logit}\, \pi_T(x_i, \boldsymbol{\theta}) &= \lambda \\ \pi_{a,b}(\pi_E, \pi_T) &= \pi_E^a (1-\pi_E)^{(1-a)} \pi_T^b (1-\pi_T)^{(1-b)} \end{aligned} \tag{6.3}$$

Once again, $a$ takes the value 1 for a patient if the efficacy event happened, else 0, and $b$ plays the equivalent role for the toxicity event. This joint model assumes the two events are independent. In Table 6.3, we simulate the effect of this assumption. To provide comparability, we have once again used diffuse $N(0, 10^2)$ priors on each element of $\boldsymbol{\theta}$ in Table 6.3.

| Sc | Coh | PrEff | PrTox | Odds | N | Eff | Tox | Diffuse | EffBias | EffEmpSE | EffCov | ToxBias | ToxEmpSE | ToxCov |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0.300 | 0.100 | 1.0 | 9.3 | 2.8 | 0.9 | 0.877 | 0.002 | 0.125 | 0.867 | 0.001 | 0.039 | 0.901 |
| | 2 | 0.300 | 0.100 | 1.0 | 13.1 | 3.9 | 1.3 | 0.904 | -0.002 | 0.113 | 0.866 | 0.001 | 0.039 | 0.901 |
| | 3 | 0.300 | 0.100 | 1.0 | 7.5 | 2.3 | 0.8 | 0.816 | 0.002 | 0.140 | 0.859 | 0.001 | 0.039 | 0.901 |
| | 4 | 0.300 | 0.100 | 1.0 | 12.5 | 3.7 | 1.2 | 0.897 | -0.003 | 0.114 | 0.876 | 0.001 | 0.039 | 0.901 |
| | 5 | 0.300 | 0.100 | 1.0 | 10.8 | 3.2 | 1.1 | 0.890 | 0.000 | 0.119 | 0.868 | 0.001 | 0.039 | 0.901 |
| | 6 | 0.300 | 0.100 | 1.0 | 6.8 | 2.0 | 0.7 | 0.818 | 0.003 | 0.142 | 0.862 | 0.001 | 0.039 | 0.901 |
| 2 | 1 | 0.100 | 0.300 | 1.0 | 9.3 | 0.9 | 2.8 | 0.019 | 0.004 | 0.085 | 0.785 | 0.000 | 0.060 | 0.892 |
| | 2 | 0.100 | 0.300 | 1.0 | 13.1 | 1.3 | 3.9 | 0.023 | -0.001 | 0.074 | 0.806 | 0.000 | 0.060 | 0.892 |
| | 3 | 0.100 | 0.300 | 1.0 | 7.5 | 0.8 | 2.3 | 0.021 | 0.005 | 0.094 | 0.737 | 0.000 | 0.060 | 0.892 |
| | 4 | 0.100 | 0.300 | 1.0 | 12.5 | 1.2 | 3.7 | 0.021 | -0.002 | 0.075 | 0.789 | 0.000 | 0.060 | 0.892 |
| | 5 | 0.100 | 0.300 | 1.0 | 10.8 | 1.1 | 3.2 | 0.023 | 0.003 | 0.080 | 0.792 | 0.000 | 0.060 | 0.892 |
| | 6 | 0.100 | 0.300 | 1.0 | 6.8 | 0.7 | 2.0 | 0.018 | 0.007 | 0.098 | 0.746 | 0.000 | 0.060 | 0.892 |
| 3 | 1 | 0.300 | 0.100 | 0.2 | 9.3 | 2.8 | 0.9 | 0.878 | 0.002 | 0.125 | 0.867 | 0.000 | 0.039 | 0.905 |
| | 2 | 0.300 | 0.100 | 0.2 | 13.1 | 3.9 | 1.3 | 0.904 | -0.002 | 0.113 | 0.868 | 0.000 | 0.039 | 0.905 |
| | 3 | 0.300 | 0.100 | 0.2 | 7.5 | 2.3 | 0.8 | 0.817 | 0.002 | 0.140 | 0.861 | 0.000 | 0.039 | 0.905 |
| | 4 | 0.300 | 0.100 | 0.2 | 12.5 | 3.7 | 1.2 | 0.897 | -0.003 | 0.114 | 0.877 | 0.000 | 0.039 | 0.905 |
| | 5 | 0.300 | 0.100 | 0.2 | 10.8 | 3.2 | 1.1 | 0.890 | 0.000 | 0.119 | 0.868 | 0.000 | 0.039 | 0.905 |
| | 6 | 0.300 | 0.100 | 0.2 | 6.8 | 2.0 | 0.7 | 0.817 | 0.003 | 0.142 | 0.863 | 0.000 | 0.039 | 0.905 |
| 4 | 1 | 0.167 | 0.100 | 1.0 | 9.3 | 1.5 | 0.9 | 0.397 | -0.017 | 0.098 | 0.806 | 0.000 | 0.039 | 0.900 |
| | 2 | 0.192 | 0.100 | 1.0 | 13.1 | 2.5 | 1.3 | 0.635 | 0.005 | 0.099 | 0.864 | 0.000 | 0.039 | 0.900 |
| | 3 | 0.500 | 0.100 | 1.0 | 7.5 | 3.8 | 0.8 | 0.974 | 0.009 | 0.162 | 0.869 | 0.000 | 0.039 | 0.900 |
| | 4 | 0.091 | 0.100 | 1.0 | 12.5 | 1.1 | 1.3 | 0.214 | 0.014 | 0.071 | 0.842 | 0.000 | 0.039 | 0.900 |
| | 5 | 0.156 | 0.100 | 1.0 | 10.8 | 1.7 | 1.1 | 0.417 | -0.005 | 0.090 | 0.833 | 0.000 | 0.039 | 0.900 |
| | 6 | 0.439 | 0.100 | 1.0 | 6.8 | 3.0 | 0.7 | 0.930 | -0.017 | 0.162 | 0.861 | 0.000 | 0.039 | 0.900 |
| 5 | 1 | 0.167 | 0.300 | 1.0 | 9.3 | 1.5 | 2.8 | 0.038 | -0.017 | 0.098 | 0.807 | 0.000 | 0.060 | 0.891 |
| | 2 | 0.192 | 0.300 | 1.0 | 13.1 | 2.5 | 3.9 | 0.064 | 0.005 | 0.099 | 0.864 | 0.000 | 0.060 | 0.891 |
| | 3 | 0.500 | 0.300 | 1.0 | 7.5 | 3.8 | 2.3 | 0.100 | 0.009 | 0.162 | 0.867 | 0.000 | 0.060 | 0.891 |
| | 4 | 0.091 | 0.300 | 1.0 | 12.5 | 1.1 | 3.7 | 0.020 | 0.014 | 0.071 | 0.842 | 0.000 | 0.060 | 0.891 |
| | 5 | 0.156 | 0.300 | 1.0 | 10.8 | 1.7 | 3.2 | 0.043 | -0.005 | 0.090 | 0.833 | 0.000 | 0.060 | 0.891 |
| | 6 | 0.439 | 0.300 | 1.0 | 6.8 | 3.0 | 2.0 | 0.098 | -0.017 | 0.162 | 0.863 | 0.000 | 0.060 | 0.891 |
| 6 | 1 | 0.167 | 0.100 | 0.2 | 9.3 | 1.5 | 0.9 | 0.396 | -0.017 | 0.098 | 0.806 | 0.001 | 0.039 | 0.902 |
| | 2 | 0.192 | 0.100 | 0.2 | 13.1 | 2.5 | 1.3 | 0.633 | 0.005 | 0.099 | 0.866 | 0.001 | 0.039 | 0.902 |
| | 3 | 0.500 | 0.100 | 0.2 | 7.5 | 3.8 | 0.8 | 0.974 | 0.009 | 0.162 | 0.869 | 0.001 | 0.039 | 0.902 |
| | 4 | 0.091 | 0.100 | 0.2 | 12.5 | 1.1 | 1.3 | 0.214 | 0.014 | 0.071 | 0.841 | 0.001 | 0.039 | 0.902 |
| | 5 | 0.156 | 0.100 | 0.2 | 10.8 | 1.7 | 1.1 | 0.417 | -0.005 | 0.090 | 0.831 | 0.001 | 0.039 | 0.902 |
| | 6 | 0.439 | 0.100 | 0.2 | 6.8 | 3.0 | 0.7 | 0.929 | -0.017 | 0.162 | 0.861 | 0.001 | 0.039 | 0.902 |

TABLE 6.3: Operating performance of model 410 in scenarios from Chapter 5 with total sample size 60. Diffuse shows probability of approval under diffuse priors. Bias, SE and coverage columns retain the definitions already given.

Comparing Table 6.3 to Tables 5.13 and 5.16, we see that performance is virtually

identical. The differences in approval probabilities are of the order of 0.1%. Likewise, the coverages of the 90% CIs are practically identical, as are the estimates of bias and standard error. When appraising the model by its ability to correctly approve or reject a treatment, there is no benefit in estimating the association parameter. We note that $\psi$ does not appear in either of the marginal models for $\pi_E$ or $\pi_T$ (6.3), and thus does not determine the marginal probability of either event or affect the approval determined by (5.12). However, as we noted above, performance is driven not naively by the number of parameters, but by the scope for borrowing and how this is impacted by the arrangement of parameters. We know from Figure 5.9 that model 411 received information on the prevailing association between efficacy and toxicity and adapted its estimation of $\psi$.

The $\psi$ parameter appears in the general joint-likelihood (2.4) and thus also in the conditional density of $\pi_E$ given $\pi_T$, for instance. In the remainder of this section, we demonstrate how $\psi$ would affect inference on unobserved efficacy after confirmed presence or absence of toxicity.

Imagine a trial scenario where a set of complete patient outcomes $\boldsymbol{X}$ has been observed. Suppose also that the toxicity status for a further patient with baseline covariate vector $x_i$ is known but their efficacy status is not. Let $T$ be a Bernoulli random variable taking values $\in \{0, 1\}$, representing the patient's toxicity outcome.

Using the identity that links conditional and joint probability for events $A$ and $B$:

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} \tag{6.4}$$

we can estimate the posterior conditional probability of efficacy for this patient given that toxicity occurred as:

$$\pi_E(x_i, \boldsymbol{\theta}|\boldsymbol{X}, T = 1) = \frac{\pi_{1,1}(\pi_E(x_i, \boldsymbol{\theta}|\boldsymbol{X}), \pi_T(x_i, \boldsymbol{\theta}|\boldsymbol{X}), \psi)}{\pi_T(x_i, \boldsymbol{\theta}|\boldsymbol{X})} \tag{6.5}$$

and given that toxicity has not occurred as:

$$\pi_E(x_i, \boldsymbol{\theta}|\boldsymbol{X}, T = 0) = \frac{\pi_{1,0}(\pi_E(x_i, \boldsymbol{\theta}|\boldsymbol{X}), \pi_T(x_i, \boldsymbol{\theta}|\boldsymbol{X}), \psi)}{1 - \pi_T(x_i, \boldsymbol{\theta}|\boldsymbol{X})} \tag{6.6}$$

To illustrate this, we sampled a single trial dataset with $N = 60$ using event rates

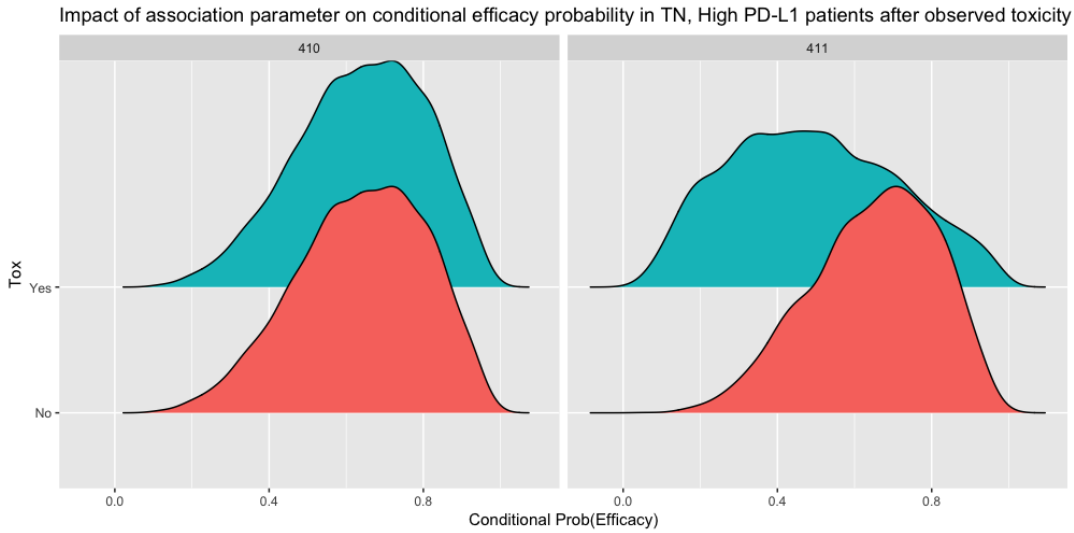



FIGURE 6.2: Posterior conditional Prob(Eff) under two models using diffuse priors, where toxicity status is categorically observed but efficacy is unknown. The wider observed dataset that forms the posterior distributions is described in the text.

and associations from scenario 6 in Table 6.3, i.e. efficacy is very likely with high PD-L1, but efficacy and toxicity are strongly negatively associated.

In our simulated data, there were 3 efficacy events from 5 patients in cohort 3, the cohort of TN patients with high PD-L1, so the observed efficacy rate slightly exceeds the underlying true rate of 50%. This dataset was then fit using models 411 and 410, each using their sets of diffuse priors.

Suppose that we wish to perform inference on an extra patient that has covariate vector that would put them in cohort 3. *A posteriori*, model 411 estimates the efficacy rate to be 63% and model 410 estimates it to be 64% in cohort 3. The posterior densities for the probability of efficacy conditional on the confirmed presence and absence of toxicity using both models are shown in Figure 6.2.

As expected, the densities under model 410 are identical because this model lacks the association parameter and cannot reflect the partial information. The expected probability of efficacy in this scenario is 64%, irrespective the observed toxicity outcome.

In contrast, the posterior density for the efficacy probability is shifted to the left under model 411 if toxicity is observed. This accurately reflects the underlying reality that efficacy and toxicity are negatively correlated, so efficacy is less likely when toxicity is observed. Here, the expected efficacy probability has fallen by 14% to 49%.

When the absence of toxicity is observed, the expected efficacy probability increases very modestly to 65%. The probability has increased to correctly reflect the negative association. However, the increase is relatively small because toxicity is rare - the true underlying rate is 10%. Confirmation of the absence of toxicity provides relatively little additional information on efficacy.

In Chapter 2 we introduced the notion of outcome ambiguity where only one of the two co-primary outcomes is known. Our motivation then was dose selection but this method of conditional inference is useful for making judgements when one of the outcomes is missing.

We see that $\psi$ performs a useful role in model 411 that will improve inference in situations with partially observed outcomes. Table 6.3 shows that performance is not improved by removing the parameter so it seems natural to retain it in the model. However, it is important that performance of the model is appraised to carry out the intended inference. If it is desired that conditional inference be reliable, simulations should be used to assess the performance of the model at this particular task. For instance, the sample size and the priors need to be sufficiently informative to identify $\psi$ and provide reliable conditional inference. We stress this because with small sample sizes, chance spurious associations could adversely impact inference. In this situation, a regularising prior on $\psi$ would likely be beneficial. This was not desired in PePS2 and further investigation of parameterisation for reliable conditional inference is beyond the scope of this thesis.

We advocate retaining the association parameter.

## 6.5 Discussion

In this chapter, we have considered notable embellishments to the P2TNE model presented for PePS2 in Chapter 5. These embellishments all sought to alter the model forms. By adding interaction terms to the efficacy sub-model to abrogate the piecewise parallel assumption, and terms to the toxicity sub-model to handle heterogeneity therein, we showed that more discriminative inference is possible. However, we saw that materially greater sample size is required to inform estimation of the extra parameters and maintain the sought level of overall statistical performance. A

unifying conclusion to these experiments in the context of PePS2 is that the potential inferential benefits were not particularly valuable and did not warrant greater accrual and a longer trial.

The incremental information to estimate the extra parameters could have come from the priors rather than the outcomes. In the examples presented in this chapter, we used very diffuse priors for comparability. The information these priors and the diffuse priors in the previous chapter contain can be legitimately said to be minimal, allowing us to ascribe difference in performance to the model specifications. This satisfies the objective of this chapter. If used in a real trial situation, we would use more informative priors that generate outcomes that genuinely reflect our expectation.

Having observed the cost of parameters to performance, we then considered whether our model could be improved by removing the association parameter that appeared to provide little benefit. We found that performance did not improve when this parameter was removed. Furthermore, we demonstrated how this parameter could be used to improve conditional inference when the co-primary outcome measures were partly observed. In these circumstances, it is natural to retain the association parameter.

An undesirable curiosity of the categorical model is that it offers the same approval probabilities to patients with PD-L1 equal to 50% and 100%. In ongoing work in Appendix D, we explore the use of continuous PD-L1 as a baseline covariate, in place of the categorical variable.

# Chapter 7

# Conclusion

In this thesis, we have considered methods that enhance efficiency in clinical trials with limited sample size. These methods have been operational and statistical, facilitating efficient clinical trial conduct and analysis. They broadly imply two overarching goals: i) use more outcomes to answer questions in trials; and ii) use all available information. We provide concluding remarks below.

## 7.1 Use more outcomes to answer questions in trials

Seamless phase I/II designs like EffTox[92] allow us to add efficacy to the typical toxicity outcome when selecting doses. This could be necessary if there is doubt about the monotonicity of the dose-efficacy relationship. By additionally evaluating short-term efficacy, a dose-finding trial can address the traditional phase I objective of identifying a dose suitable for further research, and an objective typical of phase II trials in assessing whether there is sufficient activity to warrant a randomised study. It is likely that achieving both of these objectives in a single clinical trial will be faster and cheaper than running separate trials.

In Chapter 2, we gave an in-depth account of our use of EffTox in Matchpoint. Since writing the chapter, this work has been published by Brock *et al.*[17]. Based on our experience implementing this infrequently-used design, we advocated practical measures such as phase I/II dose transition pathways (DTPs) to check in advance that a parameterisation behaves in a desirable and consistent way, and gave an illustration of latent undesirable behaviour in our original parameterisation. We introduced the phenomena of dose ambivalence, where different doses can be recommended in response to identical outcomes because of the uncertainty inherent in the

analysis and the imperfect calculation method. We overcame this challenge with repeated calculation of the dose decision. We also introduced the challenge presented by outcome ambiguity and how it can be overcome using DTPs. Phase I/II trials are efficient because they allow the objectives of two trial phases to be addressed at once. However, the described phenomena can erode that efficiency by allowing sub-optimal doses to be selected and causing delays in trial conduct. The methods we introduce show how those complications can be managed and overcome.

In Chapter 3, we introduced a novel hybrid of EffTox and Wages & Tait's (WT) design for phase I/II trials[98]. We compared the hybrid to variants of EffTox and WT in a simulation study. A version of WT that does not use randomisation showed superior statistical performance, achieving our operational efficiency objective without compromising statistical efficiency. Our hybrid achieved the same operational objective but offered slightly inferior statistical performance and greater heterogeneity, whilst allocating marginally fewer patients at attractive doses.

We remarked that additional innovation of WT would allow further streamlining of the trial objectives. In phase I/II trials, we already require that efficacy and toxicity can be evaluated over a similar, acceptable time horizon. In situations where randomisation to either a trivial or non-trivial dose is ethical, the *zero-dose* cohort may serve as a control arm. This potentially elevates a dose-finding trial to randomised controlled trial status, thus facilitating causal inference. In such scenarios, a single large dose-finding trial could be used to find the most promising dose of an experimental treatment and provide a randomised comparison to a control, thus achieving the objectives of phases I and II, requiring only a subsequent phase III trial to compare long term clinical efficacy. This takes to the extreme the potential benefits of *answering more questions in clinical trials*. Potential applications include where an experimental agent is optionally added to the standard of care. Crucially, giving an experimental treatment *instead of* a standard therapy is unlikely to be ethical when efficacy evidence exists for the latter but not the former. Thus, this method is unlikely to yield a comparison of a novel therapy to a completely distinct standard therapy.

These examples are part of a wider trend to combine the traditional phases in search of efficiency. Beyond this thesis, multi-arm multi-stage (MAMS) designs[77]

allow researchers to conduct seamless phase II/III trials. Further so-called platform trials, such as *basket trials* that assess a single treatment in many diseases, and *umbrella trials* that assess many treatments in a single disease, are becoming increasingly common. These elaborate different aspects of the traditional *Patient, Intervention, Comparator, Outcome* (PICO) paradigm that has formed the foundation of innumerable historic clinical trials. The innovations reflect the desire to answer more questions under one protocol, augment traditional trial methods to accommodate the abundance of potential treatments and molecular stratifiers, and conduct trials that accommodate the likely biological nature of contemporary experimental treatments.

## 7.2 Use all available information

The second of our overarching conclusions advocates the use of all available information to maximise statistical efficiency.

In Chapter 4, we presented a design to conduct a pivotal randomised clinical trial in an ultra-rare disease. We demonstrated by simulation that we could expect to achieve conventional error rates using 70 patients assessed seven times throughout the trial. Critical to our parameterisation of the simulations and our proposal to analyse the repeated outcome measures by hierarchical model was the data on the cohort of patients with Wolfram syndrome from St Louis. We were able to specify a model that would plausibly test the presence of a treatment effect based on the St Louis outcomes, and investigate the impact of longitudinal missing data on statistical efficiency. We did not incorporate any information from the St Louis data into the model, via priors, for example, because the trial seeks to be pivotal.

Efficiency was a factor in the choice of analysis method in TreatWolfram. We demonstrated that an analysis of final visual acuity values alone was inefficient, requiring an infeasibly high sample size. We expect to conduct a trial that will achieve acceptable power with our constrained sample size by using a hierarchical model to analyse the repeated measures data. We learned that patient-specific intercepts and gradients would likely be required to account for patient heterogeneity. That the outcome measures are subject to modest variability and are likely to be highly

correlated within patient means that inflation factors to account for missing data are relatively low, promoting an efficient analysis if up to 15% of data is missing.

There are other examples of structured outcomes that arise in clinical trials beyond repeated measures. Outcomes nested within individuals arise in crossover trials, and in scenarios where randomisation can be conducted on experimental units within individuals. Examples of the latter include trials of topical treatments like dressings, drops, and ointments that can be randomly allocated to wounds, burns, or diseased eyes within patients. Experimental designs that facilitate within-patient randomisation are particularly efficient because they control for a practically unbounded and uncountable number of potential confounding variables. For instance, lifestyle and environmental factors, concomitant medications, and every known *and unknown* gene expression are generally controlled by comparing outcomes within individual. We noted that there was further opportunity to increase expected power in TreatWolfram by analysing the repeated visual acuity outcomes within eye, nested within individuals. This scenario does not permit within patient randomisation because the oral medication is systemic. However, it will generate approximately twice the number of series, and increase statistical power. The expected gain will be much less than that notionally generated by doubling the sample size however, because of the anticipated high correlation between each individual's eyes.

Efficient use of the available information was also the theme of Chapter 5. We introduced a novel refinement of Thall, Nguyen & Estey's dose-finding design[89] that we called P2TNE. This design incorporates baseline covariates to assess co-primary binary efficacy and toxicity outcomes. Our motivation was PePS2, a trial of pembrolizumab in performance status 2 non-small-cell lung cancer patients. Based on the information reported in previous trials[37, 44] of the same treatment in closely-related patient groups, we anticipate that PD-L1 expression and pretreatedness will be predictive of efficacy but not toxicity in the PS2 patient population. We demonstrated that including the baseline predictive variables in the analysis model improved performance considerably compared to separate cohort-specific inferences provided by beta-binomial models. This improvement was apparent under diffuse, regularising and informative priors. We noted several assumptions implicit in our marginal model forms that had the potential to bias conclusions. By considering

more complex model specifications in Chapter 6, we learned that covariate terms in the toxicity sub-model and interaction terms in the efficacy sub-model required substantial increases in sample size. This was considered unjustifiable given the outcomes suggested by previous trials and the potential benefit to the PS2 population of conducting a fast phase II trial. This illustrates a restriction of statistical analysis pertinent to efficiency that is felt particularly strongly in clinical trials. The perfect analysis[1] does not exist: Inference can practically always be improved by collecting more data. However, trials are expensive, arduous and numerous. Achieving the trial objective in the agreed time-frame and avoiding the diversion of subsidiary questions and elaborate flourishes is instrumental in a successful trial.

Another important source of information in the P2TNE example was our parameter priors. It is a pervasive belief in biostatistics that priors should be diffuse to avoid biasing an analysis with external information. This view is seen to be questionable when we consider the opportunity for influence stemming from investigator degrees of freedom, like experimental design, choice of likelihood function, inclusion and exclusion of explanatory variables, choice of statistical test, and method of dealing with missing data. We demonstrated that the diffuse priors provided poor posterior coverage and empirical standard error of estimated event rates, particularly when the models contained many terms or underlying event rates were very low. Our statistical model was more efficient under our regularising and informative priors. We advocate that priors are seen as any other part of an analysis that is chosen to promote accuracy and efficiency, and requires justification in light of the alternatives.

In Chapter 6, we demonstrated the role played by the association parameter in conditional inference on partially observed outcomes in the EffTox, TNE and P2TNE models. This further reiterated the benefit of using all available information.

In Appendix D, we considered the underlying continuous PD-L1 covariate instead of the categorisation presented in [37]. We expected the continuous covariate to be more efficient and yield superior inference because it can discriminate between more cases than the categorisation. For instance, the continuous covariate can reflect that a PD-L1 score of 40% is superior to a score of 5%, but the categorical variable

[1] also, the perfect thesis

treats these as equal. Whilst we did improve inference using continuous PD-L1 in some scenarios, we noted the care needed when specifying the model form and priors to avoid inadvertently coercing undesirable information that is detrimental to analytical efficiency.

A topic largely absent from Chapters 5 and 6 is the potential benefit for the *response* variables to be continuous rather than binary. Efficacy is naturally continuous when we consider that tumour size underlies the response categories defined by RECIST[35]. The toxicity outcome may seem more naturally dichotomous in that a specific event either occurs or does not. However, methods have been proposed that analyse the total toxicity burden, informed by the frequency and severity of all adverse events sustained by patients. Evidence of the dominant culture treating efficacy as categorical or binary is that the novel methods in this thesis are refinements of previous methods[89, 92, 98] and that the RECIST[35] paper has been cited over 11,000 times (according to Google Scholar at 05-Sep-2018). Nevertheless, methods have been introduced by Wason *et al.*[100–102] that demonstrate the benefit to efficiency from retaining the continuous tumour measurements. More recently, joint modelling methods have been proposed[14] that use two-level hierarchical structure to analyse the repeated measurements of tumour lesions nested within individual (i.e. the constituent parts of the RECIST calculation) through time as an ongoing mediator of the hazard of some time-to-event endpoint like death. These methods present a desirable future direction for the methods presented herein in the pursuit of further efficiency.

## 7.3 Final conclusion

This thesis has demonstrated that efficiency in clinical trials comes from a blend of operational and statistical choices. Investigators seeking to improve efficiency should consider how they may use multiple outcomes to address the objectives of trials; and how they may use all available information, be that structured patient data, association in outcome measures, baseline covariates or priors. The settings for the contained methodological work have been the Matchpoint, TreatWolfram

and PePS2 trials of the University of Birmingham's Cancer Research UK Clinical Trials Unit.

The Matchpoint trial started recruiting patients in 2015. Shortly after it opened, the design parameterisation was slightly altered, as described in Section 2.3.2.6. Almost from the start, dose-transition pathways and repeated invocations of the dose-update decision were used to detect dose ambivalence and routinely reported to the data monitoring committee to help justify dose selections. By late 2018, the trial had evaluated 17 patients and remained open to recruitment.

The TreatWolfram trial opened in the UK at the beginning of 2019 and immediately started randomising patients. It seeks to recruit 70 patients and these are being randomised to sodium valproate or placebo at a ratio of 2:1. The team intends to open European sites in France, Spain and Poland in 2019. The primary outcome measure is visual acuity and assessments are being taken at baseline and then every six months for three years. The intended analysis model will use population-level effects for time and the interaction of time and treatment allocation, with patient-level terms for intercepts and gradients with respect to time.

PePS2 opened for recruitment at the beginning of 2017 and recruited 63 patients between then and February 2018. Baseline pretreatedness and PD-L1 status were sought at registration for all patients. The intended primary analysis will use the 411 model using categorical PD-L1 and the Gumbel association function, expanded at length in Chapter 5. This will be used to present the evidence on efficacy and toxicity in the six cohorts in Table 5.3. Sensitivity analyses may be conducted using the 611 and 441 models from Chapter 6. The association parameter is performing an important role in allowing the trialists to study the association between joint outcomes.

# Appendix A

# Published form of Chapter 2

The paper appears overleaf.

# Appendix B

# Supplementary material for Chapter 4

## B.1 Literature review search strategy

On 08-March-2017, we searched PubMed with the search phrase "visual acuity sample size trial". The PubMed search engine generalised this search string to be:

> ("visual acuity"[MeSH Terms] OR ("visual"[All Fields] AND "acuity"[All Fields]) OR "visual acuity"[All Fields]) AND ("sample size"[MeSH Terms] OR ("sample"[All Fields] AND "size"[All Fields]) OR "sample size"[All Fields]) AND ("clinical trials as topic"[MeSH Terms] OR ("clinical"[All Fields] AND "trials"[All Fields] AND "topic"[All Fields]) OR "clinical trials as topic"[All Fields] OR "trial"[All Fields])

The search returned 109 results.

Abstracts were reviewed for all results and full-texts were sought in all cases. 25 manuscripts were not immediately available through the University of Birmingham's package of journal subscriptions. Of these 25, it was evident from the content of the abstract alone in eight instances that the manuscript would not replicate our method. This was because the abstract identified another method for calculating sample size ($n = 3$), described a systematic review that did not require a prospective sample size estimate ($n = 2$), identified a non-longitudinal analysis method ($n = 2$) or listed outcomes but omitted visual acuity ($n = 1$). In the remaining 17 cases, 16 were found to be listed on ResearchGate and a full-text copy was requested

directly from the author(s). Six full-text manuscripts were obtained from ResearchGate in this way and formed part of the review described below. We resigned that the residual eleven manuscripts would remain unavailable but did not envisage that it would impact the generalisability of our results. Ninety full-text manuscripts were obtained and reviewed, summarised in Table 4.6.

## B.2 Grading our efforts by framework of Parmar *et al.*

Whilst the TreatWolfram trial was being designed, Parmar *et al.*[72] published guidance on sequential steps that may be taken to arrive at a defensible trial using a feasible sample size when conducting randomised controlled trials in rare diseases. Those steps, with the pertinent choice in TreatWolfram, are listed in Table B.1. We discuss the most noteworthy of those points now.

We addressed each item listed under the objective of *increasing what is feasible*. It became obvious relatively early on that repeated measures analysis and international recruitment would be necessary to increase the information content. Widening the eligibility criteria was not an option because the syndrome is genetically defined.

We also addressed most of the *commonly considered approaches*. As described, we sought the outcomes with the highest information content. It was perhaps fortuitous that the key outcome of visual acuity saw relatively large changes through time with relatively modest variability, making it conducive to study. The trial and intended analysis should have power of least approximately 80% to detect a treatment effect of approximately 50% in VA.

Having reached a feasible and defensible design, we had no reason to explore any of the *less common approaches*. It is reassuring to see that the decisions we took to arrive at a feasible design bear a high affinity for the advice of Parmar *et al.*[72]

| **Item** | **Wolfram** |
|---|---|
| 1. Increase what is feasible | |
| Increase accrual and / or follow-up time | We increased trial to 3 years and repeated measures to reach a feasible experiment. |
| Broaden eligibility criteria | This is not an option as the syndrome is monogenic |
| Extend collaboration nationally | We will recruit nationally in the UK |
| Extend collaboration internationally | We will use international sites |
| 2. Explore commonly-considered approaches to reducing sample size | |
| Identify experimental arm which starkly differs from control arm | We only had one experimental arm |
| Change outcome to one that is is more information-heavy | VA & VPV were selected for maximal information content. |
| Define target difference that is realistic and worthwhile, which might be larger | Targeting a 50% treatment effect in VA was deemed worthwhile. Slightly larger effect had to be used in VPV considering the invasive nature of frequent assessment. |
| Relax power by a small amount | Power stands at approximately 80% |
| 3. Explore less common approaches to reducing sample size | |
| Relax $\alpha$ a small amount | We managed to retain the conventional 5% |
| Move from two- to one-sided effects | For this pivotal study, conventional two-sided test used. We had no need to use a one-sided test. |
| Include covariate information | We considered some patient characteristics in this chapter but they appeared to yield little benefit. No further prognostic factors are known. |
| Re-randomise patients | The duration of the trial was already quite long. |
| Use external information | We had no need. |

TABLE B.1: Our choices on TreatWolfram summarised according to the framework published by Parmar *et al.*[72].

# Appendix C

# Supplementary material for Chapter 5

The following sections are included because they address potential questions that the reader might have about the methods in the main text.

## C.1 Alternative cohort prevalences in P2TNE simulations

In all of the simulations presented in the main body, we use the parameter vector $\hat{\rho} = (15.7, 21.8, 12.4, 20.7, 18.0, 11.4)$ to sample cohort memberships, for the reasons described. For clarity in this section, we refer to that set of prevalences derived for the PePS2 trial as $\hat{\rho}_P$. We investigate the sensitivity of our P2TNE implementation to the prevalences used by comparing performance under the alternative vector $\hat{\rho}_A = (16.67, 16.67, 16.67, 16.67, 16.67, 16.67)$, labelled with subscript A to denote it as an alternative. Under $\hat{\rho}_A$, patients are uniformly distributed amongst the six cohorts and the expected size of each is 10 patients.

Table C.1 compares the performance of P2TNE designs using cohort prevalences $\hat{\rho}_P$ and $\hat{\rho}_A$ in scenario 8 of our simulations in the main text. Once again, scenario 8 was chosen for its representativeness and variety. The first thing to note is that the probability of approving treatment has changed by no more than 4% in any cohort. In all cohorts except one, the probability of accepting treatment has increased where N has increased, and vice-versa. This is what we would expect.

Somewhat curiously, the performance in cohort 4 has actually improved with

TABLE C.1: Comparison of P2TNE performance in scenario 4 of Table 5.13 using the cohort prevalences derived in the main body and alternative, uniform prevalences. N is the expected cohort size. Pr(Approve) is the probability of the P2TNE design approving the treatment.

| | Scenario 8 | | $\hat{\rho}_P$ | | $\hat{\rho}_A$ | |
|---|---|---|---|---|---|---|
| Cohort | Pr(Eff) | Pr(Tox) | N | Pr(Approve) | N | Pr(Approve) |
| 1 | 0.167 | 0.1 | 9.5 | 0.457 | 10.0 | 0.483 |
| 2 | 0.192 | 0.1 | 13.1 | 0.681 | 10.0 | 0.654 |
| 3 | 0.500 | 0.1 | 7.4 | 0.979 | 10.0 | 0.989 |
| 4 | 0.091 | 0.1 | 12.4 | 0.299 | 10.0 | 0.330 |
| 5 | 0.156 | 0.1 | 10.8 | 0.493 | 10.0 | 0.511 |
| 6 | 0.439 | 0.1 | 6.8 | 0.924 | 10.0 | 0.961 |

fewer patients. This might seem counter-intuitive. However, in P2TNE, the probability of accepting a treatment in a cohort is affected by the outcomes in other cohorts. By transitioning from $\hat{\rho}_P$ to $\hat{\rho}_A$, we have effectively allocated more patients to the high PD-L1 cohorts, specifically the cohorts with the highest response rates. This has raised the expected baseline rate of response, i.e. efficacy is believed more likely in all cohorts. This is felt most sensitively in the cohort with the lowest response rate, namely cohort 4.

In the main text we showed via simulation that the P2TNE design has good operating characteristics in a wide range of scenarios. All these simulations used a common assumed set of cohort prevalences. In some trial scenarios where recruitment is stratified by the predictive variables, there will be no uncertainty about the realised cohort sizes. In "all-comers" trials like PePS2, where the cohort sizes will be randomly determined, Table C.1 shows that P2TNE is robust to reasonable deviations from the assumed cohort prevalences.

# Appendix D

# Supplementary material for Chapter 6

## D.1 Continuous PD-L1 as a covariate

The process of categorisation throws away information. This is familiar to statisticians[2]. For instance, with respect to a continuous outcome measure, the values for individual patients can be ordered from smallest to largest. When this measure is categorised to some scheme of ordered disjoint sets (e.g. "0-10", "10-20", etc), ties, where two or more patients have the same score, are necessarily more common. For instance, patients with scores 2 and 8 can be distinctly and unambiguously ordered with respect to the continuous measure, but are treated as equal in a categorisation scheme containing the set "0-10". This inability to resolve ties is the essence of information loss, and degrades the efficiency of a statistical analysis.

Much of the information that informed the design of PePS2 came from the KEYNOTE-001 study published by Garon *et al*[37]. We have hitherto used the three-level categorical PD-L1 variable that they introduced and validated. However, it would theoretically benefit us to eschew the categorisation in favour of the underlying continuous PD-L1 proportion, bounded on $[0, 1]$.

| | PD-L1 score | | Screened | | Treated | |
|---|---|---|---|---|---|---|
| | Limits (%) | Mid-point (%) | | Prevalence | | Prob(OR) |
| k | $l_k - u_k$ | $m_k$ | N | $\rho_k$ | N | $\omega_k$ |
| 1 | 0 | 0 | 323 | 0.392 | 87 | 0.081 |
| 2 | 1-24 | 12.5 | 255 | 0.310 | 147 | 0.129 |
| 3 | 25-49 | 37 | 55 | 0.067 | 27 | 0.194 |
| 4 | 50-74 | 62 | 71 | 0.086 | 39 | 0.296 |
| 5 | 75-100 | 87.5 | 120 | 0.146 | 72 | 0.454 |
| | | Total | 824 | | 372 | |

TABLE D.1: Distribution of PD-L1 scores in screened patients and probabilities of objective response (OR) in treated patients, reproduced from Figure S4 of the supplementary information of Garon *et al.*[37]. The authors label the lowest category as "$< 1$". For the purposes of modelling, we have interpreted this as 0.

In the supplementary appendix, Garon *et al.* provide the prevalence information in Table D.1, showing the distribution of the PD-L1 scores for all screened patients. It is more granular than the data in the main paper, using $k = 5$ categories instead of three. We infer from the boundaries that PD-L1 score is recorded to the nearest whole percent, else it would be ambiguous to which category a score of 24.5% would belong, for instance. For the purposes of modelling, we have given the mid-point of the PD-L1 categories, $m_k$ in Table D.1.

The PD-L1 frequencies are plotted in Figure D.1. We see that the distribution of scores is bimodal, non-normal and asymmetric. The most common category is the biomarker-negative cohort, PD-L1 < 1%. There is another local peak in the category PD-L1 > 75%, and relatively few patients in the third and fourth cohorts.

Garon *et al.* also present the observed probabilities of objective response (OR) in those same PD-L1 categories. These are given in Table D.1 and plotted in Figure D.2. As in the previous chapter, we see that the probability of response is convincingly associated with PD-L1 score.

Garon *et al.* do not provide patient-level PD-L1 scores and responses. If they did, we could fit a generalised linear model (GLM) using PD-L1 score as an independent variable to explain the chances of OR. With the resulting model, we could infer the

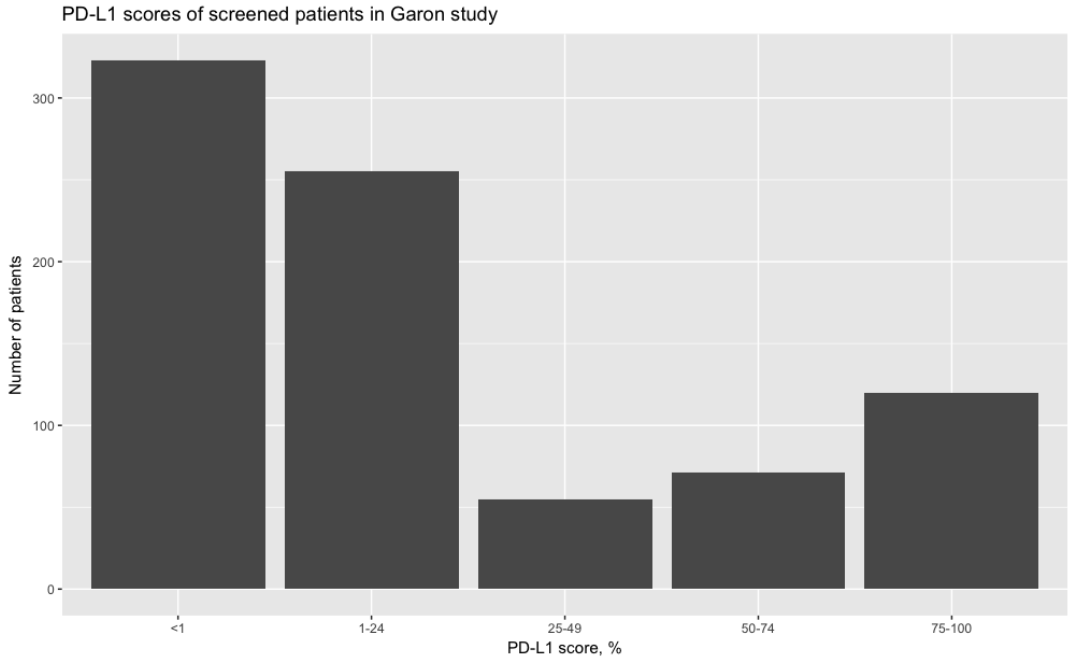



FIGURE D.1: Count of screened patients by PD-L1 score, reproduced from Figure S4 of the supplementary information of Garon *et al.*[37].

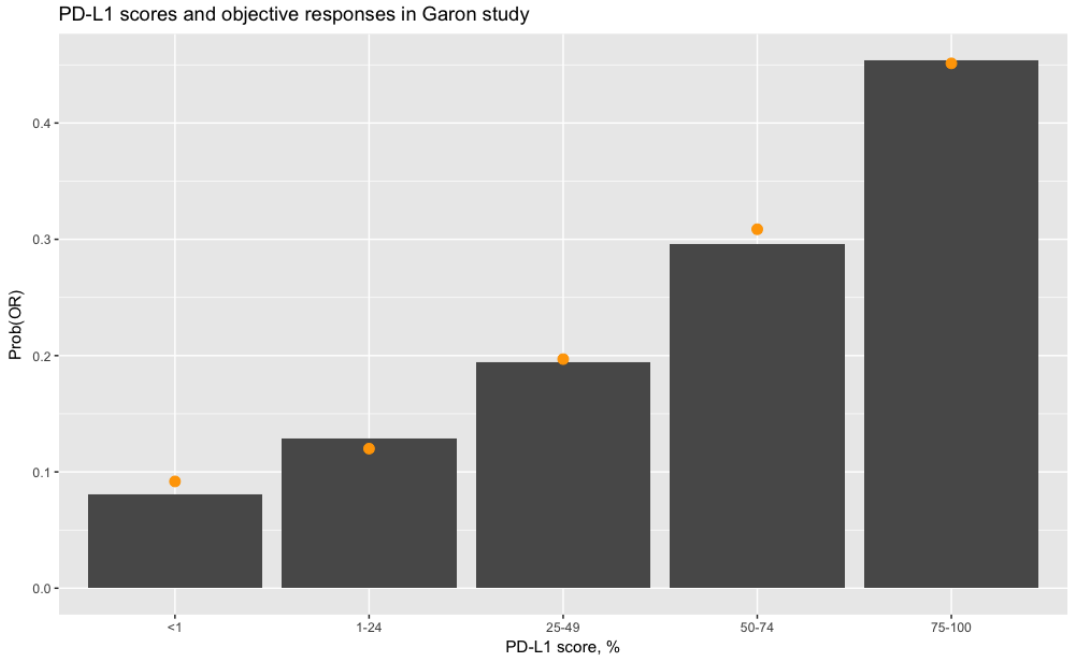



FIGURE D.2: Probability of objective response by PD-L1 score, reproduced from Figure S4 of the supplementary information of Garon *et al.*[37]. The orange dots are the values estimated by the model in (D.1).

probability of OR at any PD-L1 score. Instead, we will improvise with the data we have. The orange dots in Figure D.2 show the fitted values according to the GLM with a logit link function given by

$$\begin{aligned} \text{logit}\,\omega_k \mid m_k &= \alpha + \beta m_k, \quad k = 1, ..., 5 \\ &= -2.29 + 2.40 m_k \end{aligned} \tag{D.1}$$

when fit using the `glm` function in R, where $\omega_k$ and $m_k$ are given in Table D.1. We see from Figure D.2 that the resulting fit is good, and that interpreting the probability of response as a continuous function of PD-L1 is a plausible working model. There are errors on the probability scale of approximately 1% in cohorts 1, 2 and 4, and very small errors in cohorts 3 and 5. The intercept coefficient says that the probability of OR when PD-L1 = 0 is $\text{logit}^{-1}(-2.292) =$ expit(-2.292) $= 9.2\%$. The slope coefficient says that the odds of OR are scaled by $\exp(0.01 \times 2.40) = 1.024$, i.e. increases by 2.4%, for each 1% absolute increase in PD-L1 score.

In following sections, we use a similar method of regressing response probabilities against PD-L1 category mid-points, $m_k$, to produce data-generating models that match our simulation scenarios used hitherto.

### D.1.1 P2TNE models to use continuous PD-L1 in PePS2

In this section, we consider models that facilitate the analysis of PePS2 outcomes with continuous PD-L1 and binary pre-treatment status as baseline covariates. The general P2TNE model presented in the previous chapter used logit models for the marginal probabilities of efficacy and toxicity:

$$\pi_E(x, \boldsymbol{\theta}) = g(x, \boldsymbol{\theta}) \quad \text{and} \quad \pi_T(x, \boldsymbol{\theta}) = h(x, \boldsymbol{\theta}) \tag{D.2}$$

We will use that general form again to investigate models with different choices for $g$ and $h$ and re-use our shorthand to identify models by the number of parameters in the three model components.

Once again, let us use $x_{1i}$ to designate pre-treatment status, with $x_{1i} = 1$ signifying that patient $i$ has been previously treated, else $x_{1i} = 0$. Again, we refer to

those patients with $x_1 = 0$ as TN for *treatment naive*, and those with $x_1 = 1$ as PT for *pre-treated*. In the previous chapter, we used dummy variables $x_{2i}$ and $x_{3i}$ that determined membership of the high, medium and low PD-L1 categories. Henceforth, let us redefine $x_{2i}$ to be PD-L1 score for patient $i$, taking values on $[0, 1]$.

We demonstrated previously that overall approval probability of the ensemble model does not depend on the association parameter $\psi$. However, we identified theoretical benefits and no evidence of detriment to performance, so we retain $\psi$ and use the association model (2.4).

We use the following marginal models:

$$\text{logit}\, \pi_E(x_i, \boldsymbol{\theta}) = \alpha + \beta x_{1i} + \gamma x_{2i} + \zeta x_{1i} x_{2i} \tag{D.3}$$

and

$$\text{logit}\, \pi_T(x_i, \boldsymbol{\theta}) = \lambda \tag{D.4}$$

There is an interaction between pretreatedness and PD-L1 in the efficacy model. Toxicity is assumed uniform. We refer to this as model 411c, with the suffix *c* reflecting that the use of *continuous* PD-L1.

The posterior quantities (5.7) to (5.11), and the approval criteria (5.12) are re-used in this study. Posterior sampling is again performed using Stan[22].

### D.1.2 Randomly sampling covariates

Randomly sampling PD-L1 scores will allow us to assess the performance of our models on unseen data. We will sample PD-L1 scores to mimic the distribution presented in Table D.1 because we expect that this will reflect the distribution of scores in the PePS2 population. We only have frequencies in five categories, and no further information on the distribution of the individual scores. However, this is enough to create a functional and plausible method for simulating PD-L1 data. We propose to sample $n$ PD-L1 scores that mimic the distribution in Table D.1 using the following algorithm:

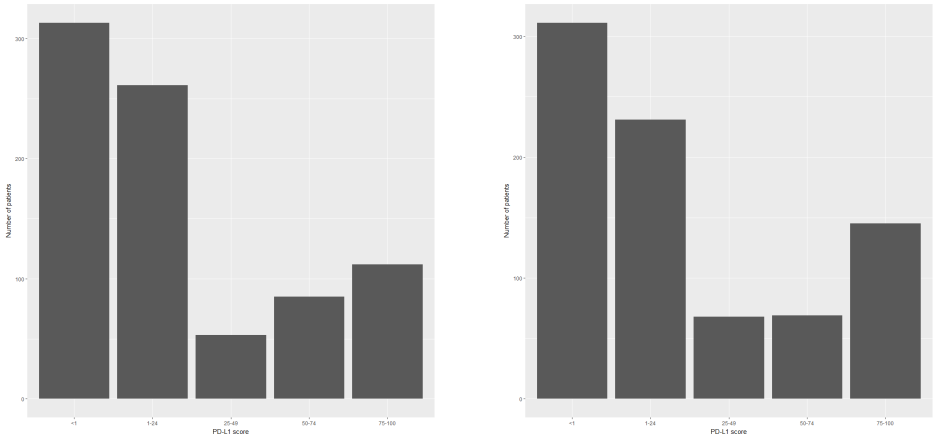


FIGURE D.3: Two examples of $n = 824$ simulated PD-L1 samples using the algorithm described above.

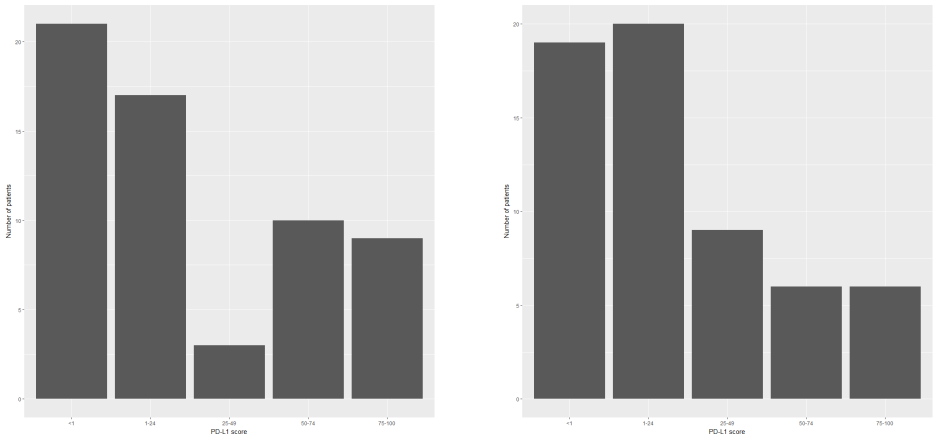


FIGURE D.4: Two examples of $n = 60$ simulated PD-L1 samples using the algorithm described above.

1. Sample cohort sizes $n_1, n_2, n_3, n_4, n_5$ from a multinomial distribution with probability parameter $\rho = (\rho_1, ..., \rho_5)$, so that $\sum_{i=1}^{5} n_i = n$;

2. Sample $n_k$ PD-L1 scores assumed to be uniformly distributed on $[l_k, u_k]$, for $k = 1, ..., 5$.

The values for $\rho_k$, $l_k$ and $u_k$ are given in Table D.1. PD-L1 scores in cohort 1 are taken simply to be 0.

Figure D.3 shows two samples of $n = 824$ PD-L1 scores simulated by this algorithm. The sample size was fixed to match that in Figure D.1. Many similarities and some modest differences are immediately recognisable. Two further examples are given in Figure D.4 with $n = 60$. Much more variability is now apparent.

Pretreatedness is sampled as a Bernoulli random variable with success parameter 0.5.

### D.1.3 Sampling efficacy and toxicity events

In simulations in Chapter 5, we assumed true probabilities of efficacy and toxicity in each of the six cohorts. With continuous PD-L1 scores, that simplicity is no longer present[1]. Instead, we now assume that models determine the probabilities of efficacy and toxicity from PD-L1 score and pre-treatment status. We refer to these as *event generating models* (EGMs). Generating events in this way presents a methodological challenge. The EGMs should be realistic in order to provide useful inferences on performance of the analysis model on unseen data. However, we should be mindful of the degree of similarity between the model used to simulate events and that used to analyse them. If the two models are unjustifiably similar, simulations will make the analysis method appear artificially good.

We address this potential hazard by using as EGMs saturated generalised linear regression models that exactly match the event probabilities for a PD-L1 score in the centre of the cohort range in the categorised setting. For example, the probability of efficacy for a PT patient with PD-L1 score of 75% will match that of a patient in cohort 6 in Chapter 5. The event probabilities at the other PD-L1 scores are interpolated by the model. This approach has the benefit of allowing us to compare model performance under categorised and continuous PD-L1.

| Scenario | Efficacy EGM | Toxicity EGM | Chapter 5 scenario |
|---|---|---|---|
| 1c | $\pi_E = 0.3$ | $\pi_T = 0.1$ | 1 |
| 2c | $\pi_E = 0.1$ | $\pi_T = 0.3$ | 2 |
| 4c | $\text{logit}\, \pi_E = -1.61 - 0.69x_1 - 0.05x_2 + 2.93x_2^2 + 2.36x_1x_2 - 2.35x_1x_2^2$ | $\pi_T = 0.1$ | 4 |
| 5c | $\text{logit}\, \pi_E = -1.61 - 0.69x_1 - 0.05x_2 + 2.93x_2^2 + 2.36x_1x_2 - 2.35x_1x_2^2$ | $\pi_T = 0.3$ | 5 |

TABLE D.2: Simulation scenarios used in Section D.1. Coefficients are expressed to two decimal places. Scenario descriptions are given in the text.

The simulation scenarios are summarised in Table D.2. Scenarios 1c and 2c mimic the benchmark favourable and adverse scenarios used to calibrate $p_E$ and $p_T$ with the categorical model in Section 5.4.2. Again, the suffix *c* is for *continuous* PD-L1.

[1]Measured to the nearest percent, we could regard the integer-valued PD-L1 scores as forming a stratification over the PD-L1 space with 101 mutually-exclusive and -exhaustive cohorts, but that would be an unfortunate way to address this problem, not least because the efficacy model would require more than 100 parameters.

The parameters for the efficacy EGM in Scenario 4c were calculated by fitting the following GLM model

$$\begin{aligned}\text{logit}\,\pi_E &= \theta_1 + \theta_2 x_1 + \theta_3 x_2 + \theta_4 x_2^2 + \theta_5 x_1 x_2 + \theta_6 x_1 x_2^2 \\ &= -1.61 - 0.69 x_1 - 0.05 x_2 + 2.93 x_2^2 + 2.36 x_1 x_2 - 2.35 x_1 x_2^2\end{aligned} \tag{D.5}$$

to the six points in Table D.3 using the `glm` function in R. The model uses 6 parameters to fit six points so that a perfect fit is guaranteed, as shown in Figure D.5. The orange dots represent the binding values in Table D.3. We see that in this scenario, efficacy in patients with high PD-L1 scores is highly likely, especially in TN patients.

| Pre-treatment status | PD-L1 score | Prob(Eff) | Prob(Tox) |
|---|---|---|---|
| $x_1$ | $x_2$ | $\pi_E$ | $\pi_T$ |
| 0 | 0 | 0.167 | 0.1 |
| 0 | 25 | 0.192 | 0.1 |
| 0 | 75 | 0.500 | 0.1 |
| 1 | 0 | 0.091 | 0.1 |
| 1 | 25 | 0.156 | 0.1 |
| 1 | 75 | 0.439 | 0.1 |

TABLE D.3: Efficacy probabilities in scenario 4c were fit to intersect these 6 points from scenario 4 in the Chapter 5.

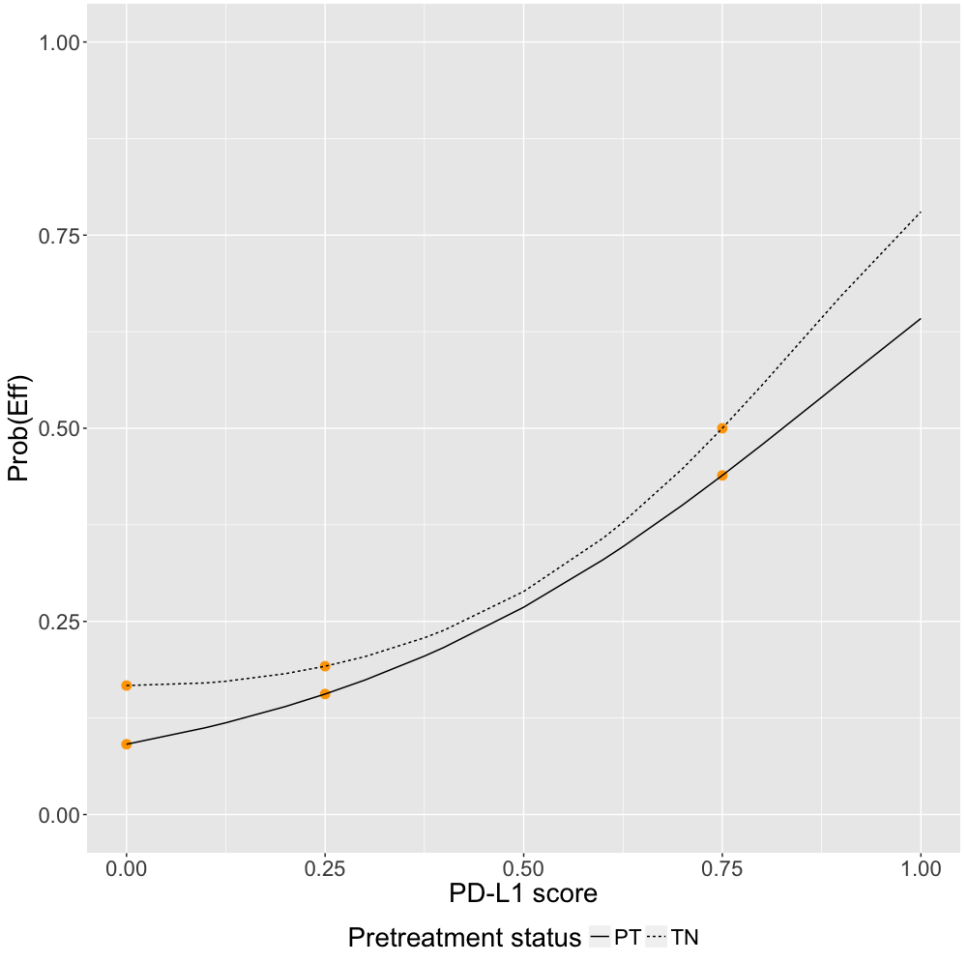



FIGURE D.5: Efficacy probabilities as functions of PD-L1 score in scenario 4c. The curves were fit to the points in scenario 4 from Chapter 5, reproduced in Table D.3.

Scenario 5c is similar to 4c, albeit with high toxicity.

### D.1.4 Simulations Analysing Continuous PD-L1

TABLE D.4: Regularising normal prior distributions for the elements of $\boldsymbol{\theta}$.

| | $\mu$ | $\sigma^2$ |
|---|---|---|
| $\alpha$ | -2.2 | 4.8 |
| $\beta$ | -0.5 | 4 |
| $\gamma$ | -0.5 | 4 |
| $\zeta$ | -0.5 | 4 |
| $\lambda$ | -2.2 | 4 |
| $\psi$ | 0 | 1 |

Regularising priors are shown in Table D.4 and the events rates that they generate are summarised in Table D.5. We see that the expected event rates and credible

intervals are similar to those generated by the regularising priors in the categorical model in Table 5.8 with only modest differences. Figure D.6 shows the prior predictive densities for TN patients with PD-L1 = 75%, chosen to allow comparison to the analogous cohort under the categorical specification, shown in Figure 5.5. The prior predictive distributions are very similar.

| PreTreat | PDL1 | ProbEffL | ProbEffl | ProbEff | ProbEffu | ProbEffU |
|---|---|---|---|---|---|---|
| TN | 0 | 0.00 | 0.02 | 0.22 | 0.33 | 0.80 |
| TN | 0.25 | 0.00 | 0.02 | 0.21 | 0.31 | 0.80 |
| TN | 0.75 | 0.00 | 0.01 | 0.21 | 0.32 | 0.86 |
| PT | 0 | 0.00 | 0.01 | 0.22 | 0.34 | 0.90 |
| PT | 0.25 | 0.00 | 0.01 | 0.20 | 0.29 | 0.89 |
| PT | 0.75 | 0.00 | 0.00 | 0.20 | 0.27 | 0.93 |
| PreTreat | PDL1 | ProbToxL | ProbToxl | ProbTox | ProbToxu | ProbToxU |
| TN | 0 | 0.00 | 0.03 | 0.20 | 0.30 | 0.75 |
| TN | 0.25 | 0.00 | 0.03 | 0.20 | 0.30 | 0.75 |
| TN | 0.75 | 0.00 | 0.03 | 0.20 | 0.30 | 0.75 |
| PT | 0 | 0.00 | 0.03 | 0.20 | 0.30 | 0.75 |
| PT | 0.25 | 0.00 | 0.03 | 0.20 | 0.30 | 0.75 |
| PT | 0.75 | 0.00 | 0.03 | 0.20 | 0.30 | 0.75 |

TABLE D.5: Credible intervals for events rates drawn from the prior predictive distribution of the regularising priors in Table D.4. Lower-case l and u show the central 50% credible interval and upper-case L and U show the central 90% credible interval.

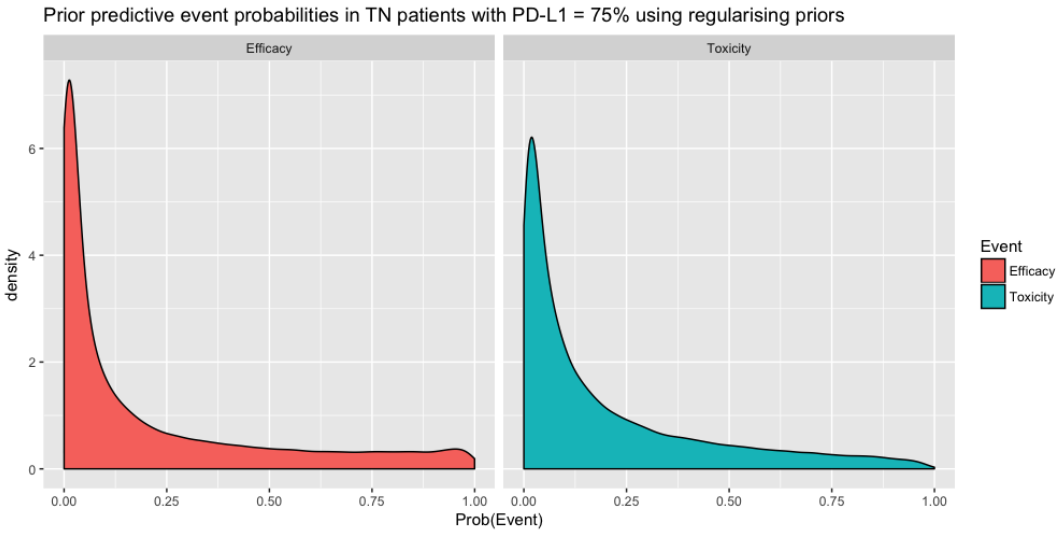



FIGURE D.6: Prior predictive distributions of the probabilities of efficacy and toxicity in TN patients with PD-L1 = 75% under regularising priors.

Table D.6 shows operating characteristics for model 411c with $n = 60$ patients.

| Sc | PreTreat | PDL1 | PrEff | PrTox | Reg | EffBias | EffEmpSE | EffCov | ToxBias | ToxEmpSE | ToxCov |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1c | 0 | 0.00 | 0.300 | 0.1 | 0.974 | -0.002 | 0.086 | 0.910 | 0.001 | 0.037 | 0.909 |
| | 0 | 0.25 | 0.300 | 0.1 | 0.977 | -0.010 | 0.081 | 0.889 | 0.001 | 0.037 | 0.909 |
| | 0 | 0.50 | 0.300 | 0.1 | 0.952 | -0.010 | 0.093 | 0.899 | 0.001 | 0.037 | 0.909 |
| | 0 | 0.75 | 0.300 | 0.1 | 0.883 | -0.003 | 0.114 | 0.917 | 0.001 | 0.037 | 0.909 |
| | 0 | 1.00 | 0.300 | 0.1 | 0.792 | 0.008 | 0.136 | 0.929 | 0.001 | 0.037 | 0.909 |
| | 1 | 0.00 | 0.300 | 0.1 | 0.959 | 0.005 | 0.095 | 0.894 | 0.001 | 0.037 | 0.909 |
| | 1 | 0.25 | 0.300 | 0.1 | 0.963 | -0.013 | 0.085 | 0.872 | 0.001 | 0.037 | 0.909 |
| | 1 | 0.50 | 0.300 | 0.1 | 0.899 | -0.018 | 0.103 | 0.877 | 0.001 | 0.037 | 0.909 |
| | 1 | 0.75 | 0.300 | 0.1 | 0.792 | -0.012 | 0.130 | 0.888 | 0.001 | 0.037 | 0.909 |
| | 1 | 1.00 | 0.300 | 0.1 | 0.684 | -0.001 | 0.157 | 0.891 | 0.001 | 0.037 | 0.909 |
| 2c | 0 | 0.00 | 0.100 | 0.3 | 0.029 | 0.009 | 0.051 | 0.921 | -0.004 | 0.057 | 0.903 |
| | 0 | 0.25 | 0.100 | 0.3 | 0.025 | -0.003 | 0.046 | 0.896 | -0.004 | 0.057 | 0.903 |
| | 0 | 0.50 | 0.100 | 0.3 | 0.020 | -0.004 | 0.051 | 0.893 | -0.004 | 0.057 | 0.903 |
| | 0 | 0.75 | 0.100 | 0.3 | 0.020 | 0.002 | 0.064 | 0.913 | -0.004 | 0.057 | 0.903 |
| | 0 | 1.00 | 0.100 | 0.3 | 0.021 | 0.014 | 0.081 | 0.936 | -0.004 | 0.057 | 0.903 |
| | 1 | 0.00 | 0.100 | 0.3 | 0.027 | 0.005 | 0.057 | 0.892 | -0.004 | 0.057 | 0.903 |
| | 1 | 0.25 | 0.100 | 0.3 | 0.017 | -0.013 | 0.047 | 0.830 | -0.004 | 0.057 | 0.903 |
| | 1 | 0.50 | 0.100 | 0.3 | 0.014 | -0.016 | 0.054 | 0.829 | -0.004 | 0.057 | 0.903 |
| | 1 | 0.75 | 0.100 | 0.3 | 0.016 | -0.006 | 0.072 | 0.845 | -0.004 | 0.057 | 0.903 |
| | 1 | 1.00 | 0.100 | 0.3 | 0.017 | 0.011 | 0.096 | 0.871 | -0.004 | 0.057 | 0.903 |
| 4c | 0 | 0.00 | 0.167 | 0.1 | 0.650 | 0.002 | 0.062 | 0.921 | 0.001 | 0.037 | 0.909 |
| | 0 | 0.25 | 0.192 | 0.1 | 0.921 | 0.049 | 0.075 | 0.852 | 0.001 | 0.037 | 0.909 |
| | 0 | 0.50 | 0.289 | 0.1 | 0.975 | 0.051 | 0.101 | 0.872 | 0.001 | 0.037 | 0.909 |
| | 0 | 0.75 | 0.500 | 0.1 | 0.981 | -0.048 | 0.132 | 0.899 | 0.001 | 0.037 | 0.909 |
| | 0 | 1.00 | 0.780 | 0.1 | 0.981 | -0.224 | 0.155 | 0.703 | 0.001 | 0.037 | 0.909 |
| | 1 | 0.00 | 0.091 | 0.1 | 0.226 | 0.017 | 0.051 | 0.928 | 0.001 | 0.037 | 0.909 |
| | 1 | 0.25 | 0.156 | 0.1 | 0.638 | 0.013 | 0.067 | 0.899 | 0.001 | 0.037 | 0.909 |
| | 1 | 0.50 | 0.268 | 0.1 | 0.882 | 0.002 | 0.104 | 0.883 | 0.001 | 0.037 | 0.909 |
| | 1 | 0.75 | 0.439 | 0.1 | 0.928 | -0.038 | 0.149 | 0.878 | 0.001 | 0.037 | 0.909 |
| | 1 | 1.00 | 0.642 | 0.1 | 0.938 | -0.115 | 0.182 | 0.870 | 0.001 | 0.037 | 0.909 |
| 5c | 0 | 0.00 | 0.167 | 0.3 | 0.083 | 0.002 | 0.063 | 0.920 | -0.004 | 0.057 | 0.901 |
| | 0 | 0.25 | 0.192 | 0.3 | 0.121 | 0.049 | 0.075 | 0.850 | -0.004 | 0.057 | 0.901 |
| | 0 | 0.50 | 0.289 | 0.3 | 0.129 | 0.051 | 0.101 | 0.870 | -0.004 | 0.057 | 0.901 |
| | 0 | 0.75 | 0.500 | 0.3 | 0.130 | -0.048 | 0.132 | 0.900 | -0.004 | 0.057 | 0.901 |
| | 0 | 1.00 | 0.780 | 0.3 | 0.130 | -0.224 | 0.155 | 0.704 | -0.004 | 0.057 | 0.901 |
| | 1 | 0.00 | 0.091 | 0.3 | 0.028 | 0.017 | 0.051 | 0.927 | -0.004 | 0.057 | 0.901 |
| | 1 | 0.25 | 0.156 | 0.3 | 0.081 | 0.013 | 0.067 | 0.900 | -0.004 | 0.057 | 0.901 |
| | 1 | 0.50 | 0.268 | 0.3 | 0.117 | 0.002 | 0.104 | 0.885 | -0.004 | 0.057 | 0.901 |
| | 1 | 0.75 | 0.439 | 0.3 | 0.122 | -0.038 | 0.149 | 0.878 | -0.004 | 0.057 | 0.901 |
| | 1 | 1.00 | 0.642 | 0.3 | 0.124 | -0.115 | 0.182 | 0.871 | -0.004 | 0.057 | 0.901 |

TABLE D.6: Operating performance of continuous PD-L1 model 411 in the scenarios in Table D.2 with total sample size 60. PreTreat reflects $x_1$ and PDL1 $x_2$. Reg shows approval probability under the regularising priors. Eff and Tox are abbreviations for efficacy and toxicity. EmpSE is empirical standard error and Cov is coverage of 90% posterior credible intervals.

In scenario 1c, we see that approval probability is unacceptable in some cohorts.

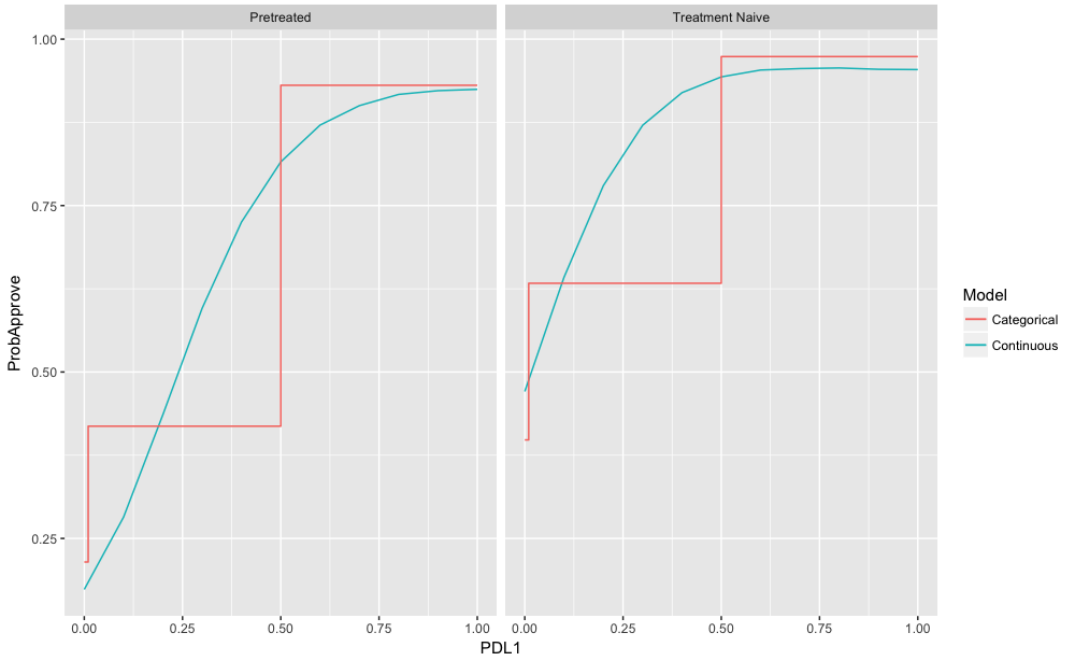



FIGURE D.7: Probability of approval in scenario 4 under the 411 models that analyse continuous and categorical PD-L1.

Performance degrades as the two covariates take values further from zero. In general, approval probability is always lower in PT (where PreTreat = 1) than TN patients (where PreTreat = 0) for matched PD-L1 and underlying event probabilities. Furthermore as PD-L1 increases in scenario 1c, the approval probabilities fall, despite the underlying event rates remaining constant. We revisit this below.

In scenario 2c, the model correctly disregards the treatment with high probability.

It is in scenario 4c that we see the real benefit of using continuous PD-L1. Figure D.7 shows the approval probability as a function of PD-L1 for models 411c and 411, the categorical model in Chapter 5. For model 411, the approval probabilities increase in steps, as reflected by the coarse categorisation scheme. For instance, as PD-L1 increases from 49% to 50%, the estimated probability of approval jumps by approximately 40% in PT patients under the categorical model. This is clearly undesirable because it fails to reflect the underlying biological reality. In contrast, the approval probability under the continuous model is a smooth increasing function of PD-L1, which mirrors that efficacy is a smooth function of PD-L1 score, as demonstrated in Figure D.2.

The PT cohort shows that the continuous model does not indiscriminately approve. Recall from Table D.6 that when PD-L1 is zero, the true efficacy probability is

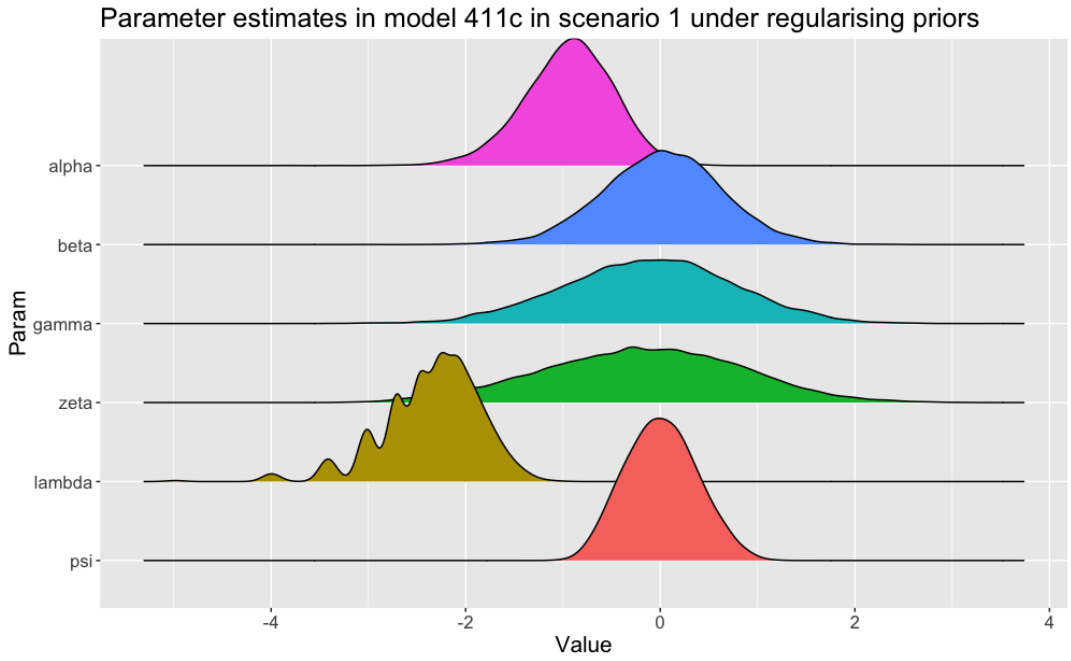



FIGURE D.8: Simulated posterior mean parameter estimates of model 411c in scenario 1c.

only 9.1% so it is incorrect to approve here. Model 411c is less likely to approve than model 411. When PD-L1 is zero in TN patients, the efficacy probability is 16.7% and it is acceptable to approve here.

There is notable negative bias in the estimations of efficacy at PD-L1 $\geq$ 75% in scenarios 4c and 5c. This is because the analysis model is less complex than the EGM, and the efficacy probability is underestimated at high PD-L1 values. Irrespective, the approval probabilities are not unduly impaired.

The continuous model performs very similarly in scenario 5c as model 411 performed in scenario 5, correctly rejecting with high probability.

Returning to scenario 1c, Figure D.8 illustrates why approval falls in PD-L1. It shows the distribution of the final parameter estimates over the simulated trial iterations. The expected values for $\beta, \gamma$ & $\zeta$ are zero, as required. However, the distributions for $\gamma$ & $\zeta$ are relatively wide. These are the two coefficients in (D.3) for terms that contribute to the gradient with respect to PD-L1 score. If they are estimated to be negative, which happens relatively frequently, the approval probability will decrease as PD-L1 increases. Nevertheless, the bias values are low throughout and the coverage values close to the theoretical 90%.

A natural temptation is to use a prior that take only positive values, like beta

or half-normal distributions. This would reflect our belief that efficacy is positively associated with PD-L1. However, when regularising half-normal priors are used, undesirable biases manifest elsewhere. In particular, material inflation occurs in the estimated efficacy rate and approval probability when PD-L1 is 0% and the underlying true efficacy rate is only 9.1%.

# Bibliography


[1] Chul Ahn, Seung-Ho Kang, and Yang Xie. "Optimal Biological Dose for Molecularly-Targeted Therapies". In: *Wiley StatsRef: Statistics Reference Online* (2014). URL: https://onlinelibrary.wiley.com/doi/10.1002/9781118445112.stat07078.

[2] Douglas G Altman and Patrick Royston. "Statistics Notes 52: The cost of dichotomising continuous variables". In: *British Medical Journal* 332.7549 (2006), p. 1080. ISSN: 1468-5833. DOI: 10.1136/bmj.332.7549.1080. URL: https://www.bmj.com/content/332/7549/1080.1/article-info.

[3] Revathi Ananthakrishnan et al. "Extensions of the mTPI and TEQR designs to include non-monotone efficacy in addition to toxicity for optimal dose determination for early phase immunotherapy oncology trials". In: *Contemporary Clinical Trials Communications* 10.January (2018), pp. 62–76. ISSN: 24518654. DOI: 10.1016/j.conctc.2018.01.006. URL: https://doi.org/10.1016/j.conctc.2018.01.006.

[4] Paolo Anderlini et al. "Cyclophosphamide conditioning in patients with severe aplastic anaemia given unrelated marrow transplantation: a phase 1–2 dose de-escalation study". In: *Lancet Haematology* 2.9 (2015), pp. 367–375. DOI: 10.1016/S2352-3026(15)00147-7.

[5] Miranta Antoniou, Andrea L Jorgensen, and Ruwanthi Kolamunnage-Dona. "Biomarker-Guided Adaptive Trial Designs in Phase II and Phase III: A Methodological Review". In: *Plos One* 11.2 (2016), e0149803. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0149803. URL: http://dx.plos.org/10.1371/journal.pone.0149803.

[6] Yung Jue Bang et al. "Trastuzumab in combination with chemotherapy versus chemotherapy alone for treatment of HER2-positive advanced gastric or gastro-oesophageal junction cancer (ToGA): A phase 3, open-label, randomised controlled trial". In: *The Lancet* 376.9742 (2010), pp. 687–697. ISSN: 01406736. DOI: 10.1016/S0140-6736(10)61121-X. URL: http://dx.doi.org/10.1016/S0140-6736(10)61121-X.

[7] T G Barrett and S E Bundey. "Wolfram (DIDMOAD) syndrome". In: *Journal of Medical Genetics* 34 (1997), pp. 838–841.

[8] T. G. Barrett, S. E. Bundey, and A. F. Macleod. "Neurodegeneration and diabetes: UK nationwide study of Wolfram (DIDMOAD) syndrome". In: *The Lancet* 346.8988 (1995), pp. 1458–1463. ISSN: 01406736. DOI: 10.1016/S0140-6736(95)92473-6.

[9] M. Suzette Blanchard and Jeffrey A. Longmate. "Toxicity equivalence range design (TEQR): A practical Phase I design". In: *Contemporary Clinical Trials* 32.1 (2011), pp. 114–121. ISSN: 15517144. DOI: 10.1016/j.cct.2010.09.011. URL: http://dx.doi.org/10.1016/j.cct.2010.09.011.

[10] H. Borghaei et al. "Nivolumab versus Docetaxel in Advanced Squamous-Cell Non-Small-Cell Lung Cancer." In: *The New England journal of medicine* (2015), pp. 1–13. ISSN: 1533-4406. DOI: 10.1056/NEJMoa1504627. URL: http://www.ncbi.nlm.nih.gov/pubmed/26028407.

[11] A. Bouckaert and M. Mouchart. "Sure outcomes of random events: A model for clinical trials". In: *Statistics in Medicine* 20.4 (2001), pp. 521–543. ISSN: 02776715. DOI: 10.1002/sim.659.

[12] Thomas M. Braun. "The bivariate continual reassessment method: Extending the CRM to phase I trials of two competing outcomes". In: *Controlled Clinical Trials* 23.3 (2002), pp. 240–256. ISSN: 01972456. DOI: 10.1016/S0197-2456(01)00205-7.

[13] Evangelos Briasoulis et al. "Dose selection trial of metronomic oral vinorelbine monotherapy in patients with metastatic cancer: a hellenic cooperative oncology group clinical translational study." In: *BMC Cancer* 13.1 (2013), p. 263. ISSN: 1471-2407. DOI: 10.1186/1471-2407-13-263. URL: http://

bmccancer.biomedcentral.com/articles/10.1186/1471-2407-13-263.

[14] Samuel L Brilleman et al. "Joint longitudinal and time-to-event models for multilevel hierarchical data". In: *Statistical Methods in Medical Research* (2018). DOI: https://doi.org/10.1177/0962280218808821. URL: https://journals.sagepub.com/doi/abs/10.1177/0962280218808821.

[15] Kristian Brock. *clintrials*. 2016. URL: https://github.com/brockk/clintrials.

[16] Kristian Brock. *trialr - clinical trial designs in R & Stan*. 2017. URL: https://cran.r-project.org/web/packages/trialr/index.html.

[17] Kristian Brock et al. "Implementing the EffTox dose-finding design in the Matchpoint trial". In: *BMC Medical Research Methodology* 17.1 (2017), p. 112. ISSN: 1471-2288. DOI: 10.1186/s12874-017-0381-x. URL: http://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-017-0381-x.

[18] P. Brutti, S. Gubbiotti, and V. Sambucini. "An extension of the single threshold design for monitoring efficacy and safety in phase II clinical trials". In: *Statistics in Medicine* 30.14 (2011), pp. 1648–1664. ISSN: 02776715. DOI: 10.1002/sim.4229.

[19] J Bryant and R Day. "Incorporating toxicity considerations into the design of two-stage phase II clinical trials." In: *Biometrics* 51.4 (1995), pp. 1372–1383. ISSN: 0006-341X. DOI: 10.2307/2533268.

[20] Marc Buyse et al. "Integrating biomarkers in clinical trials". In: *Expert Review of Molecular Diagnostics2* 11.2 (2011), pp. 171–182. ISSN: 1473-7159. DOI: 10.1586/ERM.10.120.

[21] Kunthel By and Bahjat F. Qaqish. *binarySimCLF*. URL: https://cran.r-project.org/src/contrib/Archive/binarySimCLF/.

[22] Bob Carpenter et al. "Stan: A Probabilistic Programming Language". In: *Journal of Statistical Software* 76.1 (2017). DOI: 10.18637/jss.v076.i01.

[23] Tara L. Chen et al. "Cyclosporine Modulation of Multidrug Resistance in Combination with Pravastatin, Mitoxantrone, and Etoposide for Adult Patients with Relapsed/Refractory Acute Myeloid Leukemia (AML): A Phase 1/2 Study". In: *Leuk Lymphoma* 54.11 (2013), pp. 2534–6. ISSN: 15378276. DOI: 10.3109/10428194.2013.777836. arXiv: NIHMS150003.

[24] Ying Kuen Cheung. *Dose Finding by the Continual Reassessment Method*. New York: Chapman & Hall / CRC Press, 2011.

[25] M R Conaway and G R Petroni. "Bivariate sequential designs for phase II trials." In: *Biometrics* 51.2 (1995), pp. 656–664. ISSN: 0006341X. DOI: 10.2307/2532952.

[26] M R Conaway and G R Petroni. "Designs for phase II trials allowing for a trade-off between response and toxicity." In: *Biometrics* 52.4 (1996), pp. 1375–1386. ISSN: 0006-341X. DOI: 10.2307/2532851.

[27] John D Cook. *Efficacy-Toxicity trade-offs based on L-p norms: Technical Report UTMDABTR-003-06*. Tech. rep. 2006, pp. 1–9.

[28] R J Cook and V T Farewell. "Guidelines for monitoring efficacy and toxicity responses in clinical trials." In: *Biometrics* 50.4 (1994), pp. 1146–1152. ISSN: 0006-341X. DOI: 10.2307/2533451.

[29] J E Cortes et al. "A phase 2 trial of ponatinib in Philadelphia chromosome-positive leukemias." In: *The New England journal of medicine* 369.19 (2013), pp. 1783–96. ISSN: 1533-4406. DOI: 10.1056/NEJMoa1306494. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3886799{\&}tool=pmcentrez{\&}rendertype=abstract.

[30] D Curran et al. "Identifying the types of missingness in quality of life data from clinical trials". In: *Statistics in medicine* 17.5-7 (1998), pp. 739–756. ISSN: 0277-6715. DOI: 10.1002/(SICI)1097-0258(19980315/15)17:5/7<739::AID-SIM818>3.0.CO;2-M. URL: http://dx.doi.org/10.1002/(SICI)1097-0258(19980315/15)17:5/7{\%}3C739::AID-SIM818{\%}3E3.0.CO;2-M.

[31] Peter J. Diggle, Kung-Yee Liang, and Scott L. Zeger. *Analysis of Longitudinal Data*. Oxford University Press, 2000. ISBN: 0-19-852284-3.

[32] Hartmut Döhner et al. *Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel*. eng. 2017. DOI: 10.1182/blood-2016-08-733196.

[33] Brian J. Druker et al. "Activity of a specific inhibitor of the BCR-ABL tyrosine kinase in the blast crisis of Chronic Myeloid Leukemia and Acute Lymphoblastic Leukemia with the Philadelphia chromosome". In: *New England Journal of Medicine* 344.14 (2001), pp. 1038–1042.

[34] Brian J. Druker et al. "Five-Year Follow-up of Patients Receiving Imatinib for Chronic Myeloid Leukemia". In: *New England Journal of Medicine* 355.23 (2006), pp. 2408–17.

[35] E. a. Eisenhauer et al. "New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1)". In: *European Journal of Cancer* 45.2 (2009), pp. 228–247. ISSN: 09598049. DOI: 10.1016/j.ejca.2008.10.026. URL: http://dx.doi.org/10.1016/j.ejca.2008.10.026.

[36] Boris Freidlin, Lisa M. McShane, and Edward L. Korn. "Randomized clinical trials with biomarkers: Design issues". In: *Journal of the National Cancer Institute* 102.3 (2010), pp. 152–160. ISSN: 00278874. DOI: 10.1093/jnci/djp477.

[37] Edward B Garon et al. "Pembrolizumab for the treatment of non-small-cell lung cancer." In: *The New England journal of medicine* 372.21 (2015), pp. 2018–28. ISSN: 1533-4406. DOI: 10.1056/NEJMoa1501824. URL: http://www.ncbi.nlm.nih.gov/pubmed/25891174.

[38] Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel / Hierarchical Models*. Cambridge University Press, 2007. ISBN: 978-0-521-68689-1.

[39] Andrew Gelman, Daniel Simpson, and Michael Betancourt. "The prior can generally only be understood in the context of the likelihood". In: (2017). arXiv: arXiv:1708.07487v2.

[40] Aklilu Habteab Ghebretinsae et al. "Joint modeling of hierarchically clustered and overdispersed non-gaussian continuous outcomes for comet assay data". In: *Pharmaceutical Statistics* 11.6 (2012), pp. 449–455. ISSN: 15391604. DOI: 10.1002/pst.1533.

[41] F J Giles et al. "Nilotinib is effective in imatinib-resistant or -intolerant patients with chronic myeloid leukemia in blastic phase." In: *Leukemia* 26.5 (2012), pp. 959–962. ISSN: 1476-5551. DOI: 10.1038/leu.2011.355. URL: http://www.nature.com/doifinder/10.1038/leu.2011.355.

[42] Yi Guo et al. "Selecting a sample size for studies with repeated measures". In: *BMC Medical Research Methodology* 13.1 (2013). ISSN: 14712288. DOI: 10.1186/1471-2288-13-100. arXiv: arXiv:1011.1669v3.

[43] Ruediger Hehlmann. "How I treat CML blast crisis How I treat How I treat CML blast crisis". In: *Blood* 120.4 (2012), pp. 737–748. DOI: 10.1182/blood-2012-03-380147. URL: http://bloodjournal.hematologylibrary.org/content/120/4/737.full.html.

[44] Roy S. Herbst et al. "Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): A randomised controlled trial". In: *The Lancet* 387.10027 (2016), pp. 1540–1550. ISSN: 1474547X. DOI: 10.1016/S0140-6736(15)01281-7.

[45] R Herrick et al. *EffTox*. 2015. URL: https://biostatistics.mdanderson.org/softwaredownload/SingleSoftware.aspx?Software{_}Id=2.

[46] Anastasia Ivanova. "A New Dose-Finding Design for Bivariate Outcomes". In: *Biometrics* 59.4 (2003), pp. 1001–1007. ISSN: 0006341X. DOI: 10.1111/j.0006-341X.2003.00115.x.

[47] Elias Jabbour et al. "Early response with dasatinib or imatinib in chronic myeloid leukemia : 3-year follow-up from a randomized phase 3 trial ( DASISION )". In: *Blood* 123.4 (2014), pp. 494–501. DOI: 10.1182/blood-2013-06-511592.The.

[48] Nitin Jain and Koen Van Besien. "Chronic Myelogenous Leukemia: Role of Stem Cell Transplant in the Imatinib Era". In: *Hematology/Oncology Clinics of North America* 25.5 (2011), pp. 1025–1048. ISSN: 08898588. DOI: 10.1016/j.hoc.2011.09.003.

[49] Yuan Ji and Sue Jane Wang. "Modified toxicity probability interval design: a safer and more reliable method than the 3 + 3 design for practical phase I trials." In: *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 31.14 (2013), pp. 1785–1791. ISSN: 15277755. DOI: 10.1200/JCO.2012.45.7903.

[50] Hua Jin. "Alternative designs of phase II trials considering response and toxicity". In: *Contemporary Clinical Trials* 28.4 (2007), pp. 525–531. ISSN: 15517144. DOI: 10.1016/j.cct.2007.03.003.

[51] Edward S. Kim et al. "The BATTLE trial: Personalizing Therapy for Lung Cancer". In: *Cancer Discovery* 1.1 (2011), pp. 44–53. ISSN: 21598274. DOI: 10.1158/2159-8274.CD-10-0010. arXiv: 1112.3563.

[52] T. Klopstock et al. "Persistence of the treatment effect of idebenone in Leber's hereditary optic neuropathy". In: *Brain* 136.2 (2013). ISSN: 14602156. DOI: 10.1093/brain/aws279.

[53] Thomas Klopstock et al. "A randomized placebo-controlled trial of idebenone in Leber's hereditary optic neuropathy". In: *Brain* 134.9 (2011), pp. 2677–2686. ISSN: 14602156. DOI: 10.1093/brain/awr170.

[54] Marina Konopleva et al. "Phase I/II study of the hypoxia-activated prodrug PR104 in refractory/relapsed acute myeloid leukemia and acute lymphoblastic leukemia". In: *e* 100.7 (2015), pp. 927–934. ISSN: 15928721. DOI: 10.3324/haematol.2014.118455.

[55] Byron L Lam et al. "Trial end points and natural history in patients with G11778A Leber hereditary optic neuropathy : preparation for gene therapy clinical trial." In: *JAMA ophthalmology* 132.4 (2014), pp. 428–36. ISSN: 2168-6173. DOI: 10.1001/jamaophthalmol.2013.7971. arXiv: NIHMS150003. URL: http://www.scopus.com/inward/record.url?eid=2-s2.0-84898603855{\&}partnerID=tZOtx3y1.

[56] Stanley Lambertus et al. "Progression of Late-Onset Stargardt Disease". In: *Investigative Opthalmology & Visual Science* 57.13 (2016), p. 5186. ISSN: 1552-5783. DOI: 10.1167/iovs.16-19833. URL: http://www.ncbi.nlm.nih.gov/pubmed/27699414{\%}5Cnhttp://iovs.arvojournals.org/article.aspx?doi=10.1167/iovs.16-19833.

[57] C. Lange et al. "Resolving the clinical acuity categories "hand motion" and "counting fingers" using the Freiburg Visual Acuity Test (FrACT)". In: *Graefe's Archive for Clinical and Experimental Ophthalmology* 247.1 (2009), pp. 137–142. ISSN: 0721832X. DOI: 10.1007/s00417-008-0926-0.

[58] RA Larson et al. "Nilotinib vs imatinib in patients with newly diagnosed Philadelphia chromosome-positive chronic myeloid leukemia in chronic phase : ENESTnd 3-year follow-up This article has been corrected since Advance Online Publication and a corrigendum is also printed". In: *Leukemia* 26 (2012), pp. 2197–2203. DOI: 10.1038/leu.2012.134.

[59] Christophe Le Tourneau, J. Jack Lee, and Lillian L. Siu. "Dose escalation methods in phase i cancer clinical trials". In: *Journal of the National Cancer Institute* 101.10 (2009), pp. 708–720. ISSN: 00278874. DOI: 10.1093/jnci/djp079.

[60] Kaifeng Lu, Xiaohui Luo, and Pei Yun Chen. "Sample size estimation for repeated measures analysis in randomized clinical trials with missing data". In: *International Journal of Biostatistics* 4.1 (2008). ISSN: 15574679. DOI: 10.2202/1557-4679.1098.

[61] David Machin et al. *Sample Size Tables for Clinical Studies*. 3rd. Wiley, 2008. ISBN: 9781405146500.

[62] Sumithra J. Mandrekar, Suzanne E. Dahlberg, and Richard Simon. "Improving Clinical Trial Efficiency: Thinking outside the Box." In: *American Society of Clinical Oncology educational book. American Society of Clinical Oncology. Meeting* 35 (2015), e141–7. ISSN: 1548-8756. DOI: 10.14694/EdBook_AM.2015.35.e141. URL: http://meetinglibrary.asco.org/content/11500141-156{\%}5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/25993165.

[63] Sumithra J. Mandrekar, Rui Qin, and Daniel J. Sargent. "Model-based phase I designs incorporating toxicity and efficacy for single and dual agent drug combinations: Methods and challenges". In: *Statistics in Medicine* 29.10 (2010), pp. 1077–1083. ISSN: 02776715. DOI: 10.1002/sim.3706.

[64] Richard McElreath. *Statistical Rethinking*. CRC Press, 2015, p. 487. ISBN: 1482253445.

[65] John Monahan and Alan Genz. "Spherical-Radial Integration Rules for Bayesian Computation". In: *Journal of the American Statistical Association* 92.438 (1997), pp. 664–674. URL: http://www.jstor.org/stable/2965714.

[66] Tim P Morris, Ian R White, and Michael J Crowther. "Using simulation studies to evaluate statistical methods". In: *Statistics in Medicine* (2019). DOI: https://doi.org/10.1002/sim.8086. URL: https://onlinelibrary.wiley.com/doi/full/10.1002/sim.8086.

[67] Z Nagy, M M Esiri, and A D Smith. "The cell division cycle and the pathophysiology of Alzheimer's disease." In: *Neuroscience* 87.4 (1998), pp. 731–9. ISSN: 0306-4522. DOI: S0306452298002930[pii]. URL: http://www.ncbi.nlm.nih.gov/pubmed/9759963.

[68] J O'Quigley, M D Hughes, and T Fenton. "Dose-finding designs for HIV studies." In: *Biometrics* 57.4 (2001), pp. 1018–1029. ISSN: 0006341X. DOI: 10.1111/j.0006-341X.2001.01018.x.

[69] J O'Quigley, M Pepe, and L Fisher. "Continual reassessment method: a practical design for phase 1 clinical trials in cancer." In: *Biometrics* 46.1 (1990), pp. 33–48. ISSN: 0006-341X. DOI: 10.2307/2531628.

[70] J O'Quigley and L Z Shen. "Continual reassessment method: a likelihood approach". In: *Biometrics* 52.2 (1996), pp. 673–684. ISSN: 0006-{341X}. DOI: 10.2307/2532905.

[71] Francesca Palandri et al. "Chronic myeloid leukemia in blast crisis treated with imatinib 600 mg: Outcome of the patients alive after a 6-year follow-up". In: *Haematologica* 93.12 (2008), pp. 1792–1796. ISSN: 03906078. DOI: 10.3324/haematol.13068.

[72] Mahesh K B Parmar, Matthew R Sydes, and Tim P Morris. "How do you design randomised trials for smaller populations? A framework". In: *BMC Medicine* 14.183 (2016). DOI: 10.1186/s12916-016-0722-3.

[73] Jose Pinheiro and Douglas Bates. *Mixed-Effects Models in S and S-PLUS*. 1st. Springer, 2000, p. 548. ISBN: 0387989579.

[74] Jose Pinheiro et al. *nlme: Linear and Nonlinear Mixed Effects Models*. 2016. URL: http://cran.r-project.org/package=nlme.

[75] Bahjat F. Qaqish. "A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations". In: *Biometrika* 90.2 (2003), pp. 455–463. ISSN: 00063444. DOI: 10.1093/biomet/90.2.455.

[76] R Core Team. *R: A Language and Enviroment for Statistical Computing*. Vienna, Austria, 2017. URL: https://www.r-project.org/.

[77] Patrick Royston, Mahesh K.B. Parmar, and Wendi Qian. "Novel designs for multi-arm clinical trials with survival outcomes with an application in ovarian cancer". In: *Statistics in Medicine* 22.14 (2003), pp. 2239–2256. ISSN: 02776715. DOI: 10.1002/sim.1430.

[78] Giuseppe Saglio et al. "Dasatinib in imatinib-resistant or imatinib-intolerant chronic myeloid leukemia in blast phase after 2 years of follow-up in a phase 3 study: Efficacy and tolerability of 140 milligrams once daily and 70 milligrams twice daily". In: *Cancer* 116.16 (2010), pp. 3852–3861. ISSN: 0008543X. DOI: 10.1002/cncr.25123.

[79] Joan H Schiller et al. "Comparison of four chemotherapy regimens for advanced non-small-cell lung cancer." In: *The New England journal of medicine* 346.2 (2002), pp. 92–8. ISSN: 1533-4406. DOI: 10.1056/NEJMoa011954. URL: http://www.ncbi.nlm.nih.gov/pubmed/11784875.

[80] Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. "Sample size considerations for the evaluation of prognostic factors in survival analysis". In: *Statistics in medicine* 19 (2000), pp. 441–452.

[81] Nina Shah et al. "Phase I/II trial of lenalidomide and high dose melphalan with autologous stem cell transplantation for relapsed myeloma". In: *Leukemia* 29.9 (2015), pp. 1945–48. DOI: 10.1038/leu.2015.54.

[82] Fumiya Shimamura et al. "Two-stage approach based on zone and dose findings for two-agent combination Phase I/II trials". In: *Journal of Biopharmaceutical Statistics* 00.00 (2018), pp. 1–13. ISSN: 1054-3406. DOI: 10.1080/10543406.2018.1434190. URL: https://www.tandfonline.com/doi/full/10.1080/10543406.2018.1434190.

[83] Richard Simon and Aboubakar Maitournam. "Evaluating the Efficiency of Targeted Designs for Randomized Clinical Trials". In: *Clinical Cancer Research* 10.12 (2004), pp. 6759–6763. DOI: 10.1158/1078-0432.CCR-04-0496. URL: https://brb.nci.nih.gov/techreport/rct.pdf.

[84] Karline Soetaert. *rootSolve: Non-linear root finding, equilibrium and steady-state analysis of ordinary differential equations*. 2009.

[85] Simona Soverini et al. "Contribution of ABL kinase domain mutations to imatinib resistance in different subsets of Philadelphia-positive patients: By the GIMEMA working party on chronic myeloid leukemia". In: *Clinical Cancer Research* 12.24 (2006), pp. 7374–7379. ISSN: 10780432. DOI: 10.1158/1078-0432.CCR-06-1516.

[86] Michael Sweeting, Adrian Mander, and Tony Sabin. "bcrm : Bayesian Continual Reassessment Method Designs for Phase I Dose-Finding Trials". In: *Journal of Statistical Software* 54.13 (2013), pp. 1–26. ISSN: 1548-7660. DOI: 10.18637/jss.v054.i13. URL: http://www.jstatsoft.org/v54/i13/.

[87] P F Thall, R M Simon, and E H Estey. "New statistical strategy for monitoring safety and efficacy in single- arm clinical trials". In: *J.Clin.Oncol.* 14.0732-183X SB - M SB - X (1996), pp. 296–303. ISSN: 0732-183X.

[88] Peter F Thall. "Bayesian Models and Decision Algorithms for Complex Early Phase Clinical Trials". In: *Statistical Science* 25.2 (2010), pp. 227–244. DOI: 10.1214/09-STS315.Bayesian.

[89] Peter F. Thall, Hoang Q. Nguyen, and Elihu H. Estey. "Patient-specific dose finding based on bivariate outcomes and covariates". In: *Biometrics* 64.4 (2008), pp. 1126–1136. ISSN: 0006341X. DOI: 10.1111/j.1541-0420.2008.01009.x.

[90] Peter F Thall, Richard M Simon, and Elihu H Estey. "Single-Arm Clinical Trials With Multiple Outcomes". In: *Statistics in Medicine* 14.October 1993 (1995), pp. 357–379.

[91] Peter F. Thall and Hsi Guang Sung. "Some extensions and applications of a Bayesian strategy for monitoring multiple outcomes in clinical trials". In: *Statistics in Medicine* 17.14 (1998), pp. 1563–1580. ISSN: 02776715. DOI: 10.1002/(SICI)1097-0258(19980730)17:14<1563::AID-SIM873>3.0.CO;2-L.

[92] PF Thall and JD Cook. "Dose-Finding Based on Efficacy-Toxicity Trade-Offs". In: *Biometrics* 60.3 (2004), pp. 684–693.

[93] PF Thall, JD Cook, and EH Estey. "Adaptive dose selection using efficacy-toxicity trade-offs: illustrations and practical considerations." In: *Journal of biopharmaceutical statistics* 16.5 (2006), pp. 623–638. ISSN: 1054-3406. DOI: 10.1080/10543400600860394.

[94] PF Thall et al. "Effective sample size for computing prior hyperparameters in Bayesian phase I-II dose-finding". In: *Clinical Trials* 11.6 (2014), pp. 657–666. ISSN: 1740-7745. DOI: 10.1177/1740774514547397. URL: http://ctj.sagepub.com/cgi/doi/10.1177/1740774514547397.

[95] Nick Thatcher et al. "Gefitinib plus best supportive care in previously treated patients with refractory advanced non-small-cell lung cancer: Results from a randomised, placebo-controlled, multicentre study (Iressa Survival Evaluation in Lung Cancer)". In: *Lancet* 366.9496 (2005), pp. 1527–1537. ISSN: 01406736. DOI: 10.1016/S0140-6736(05)67625-8.

[96] G Virgili et al. *Reading aids for adults with low vision*. Tech. rep. 10. Cochran Library, 2013. DOI: 10.1002/14651858.CD003303.pub3.www.cochranelibrary.com. URL: http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD003303.pub3/epdf/standard.

[97] Jyoti Wadhwa et al. "Factors affecting duration of survival after onset of blastic transformation of chronic myeloid leukemia Factors affecting duration of survival after onset of blastic transformation of chronic myeloid leukemia". In: 99.7 (2012), pp. 2304–2309. DOI: 10.1182/blood.V99.7.2304.

[98] Nolan A. Wages and Christopher Tait. "Seamless Phase I/II Adaptive Design for Oncology Trials of Molecularly Targeted Agents." In: *Journal of biopharmaceutical statistics* 25.5 (2015), pp. 903–20. ISSN: 1520-5711. DOI: 10.1080/10543406.2014.920873. URL: http://www.tandfonline.com/doi/abs/10.1080/10543406.2014.920873{\#}.Vdjqwnh3910.

[99] Meihua Wang and Roger Day. "Adaptive Bayesian design for phase I dose-finding trials using a joint model of response and toxicity." In: *Journal of biopharmaceutical statistics* 20.1 (2010), pp. 125–144. ISSN: 1054-3406. DOI: 10.1080/10543400903280613.

[100] J. M. Wason and a. P. Mander. "The choice of test in phase II cancer trials assessing continuous tumour shrinkage when complete responses are expected". In: *Statistical Methods in Medical Research* (2011). ISSN: 0962-2802. DOI: 10.1177/0962280211432192.

[101] James M S Wason, Adrian P. Mander, and Tim G. Eisen. "Reducing sample sizes in two-stage phase II cancer trials by using continuous tumour shrinkage end-points". In: *European Journal of Cancer* 47.7 (2011), pp. 983–989. ISSN: 09598049. DOI: 10.1016/j.ejca.2010.12.007. URL: http://dx.doi.org/10.1016/j.ejca.2010.12.007.

[102] James M S Wason and Shaun R. Seaman. "Using continuous data on tumour measurements to improve inference in phase II cancer studies". In: *Statistics in Medicine* 32.26 (2013), pp. 4639–4650. ISSN: 02776715. DOI: 10.1002/sim.5867.

[103] J. Kyle Wathen et al. "Accounting for patient heterogeneity in phase II clinical trials J." In: *Statistics in medicine* 27 (2008), pp. 2802–2815. ISSN: 02776715. DOI: 10.1002/sim.

[104] Ian White. "Strategies for handling missing data in randomised trials". In: *Trials* 12.October (2011), A59. ISSN: 1745-6215. URL: http://www.trialsjournal.com/content/12/S1/A59.

[105] Ian White et al. "Strategy for intention to treat analysis in randomised trials with missing outcome Data Strategy for intention to treat analysis in randomised". In: *British Medical Journal* 342.d40 (2011). DOI: 10.1136/bmj.d40.

[106] Henry E. Wiley et al. "A crossover design for comparative efficacy: A 36-week randomized trial of bevacizumab and ranibizumab for diabetic macular edema". In: *Ophthalmology* 123.4 (2016), pp. 841–849. ISSN: 15494713. DOI: 10.1016/j.ophtha.2015.11.021.

[107] DJ Wolfram and HP Wagener. *Diabetes mellitus and simple optic atrophy among siblings: report of four cases*. Tech. rep. MAYO CLINIC PROCEEDINGS, 1938.

[108] Christina Yap et al. "Dose transition pathways: The missing link between complex dose-finding designs and simple decision-making". In: *Clinical Cancer Research* 23.24 (2017), pp. 7440–7447. ISSN: 15573265. DOI: 10.1158/1078-0432.CCR-17-0582.

[109] Mei Ling Yeh, Hsing Hsia Chen, and Yu Chu Chung. "One year study on the integrative intervention of acupressure and interactive multimedia for visual health in school children". In: *Complementary Therapies in Medicine* 20.6 (2012), pp. 385–392. ISSN: 09652299. DOI: 10.1016/j.ctim.2012.09.001. URL: http://dx.doi.org/10.1016/j.ctim.2012.09.001.

[110] Guosheng Yin, Yisheng Li, and Yuan Ji. "Bayesian dose-finding in phase I/II clinical trials using toxicity and efficacy odds ratios". In: *Biometrics* 62.3 (2006), pp. 777–787. ISSN: 0006341X. DOI: 10.1111/j.1541-0420.2006.00534.x.

[111] Ying Yuan, Hoang Q. Nguyen, and Peter F. Thall. *Bayesian Designs for Phase I–II Clinical Trials*. 1st. C: CRC Press, 2016. ISBN: 9781498709552.

[112] Wei Zhang, Daniel J. Sargent, and Sumithra Mandrekar. "An adaptive dose-finding design incorporating both toxicity and efficacy". In: *Statistics in Medicine* 25.14 (2006), pp. 2365–2383. ISSN: 02776715. DOI: 10.1002/sim.2325.