Bayesian Bandit Models for the Design of Clinical Trials

S. Faye Williamson, B.Sc. (Hons.), M.Res.

Centre for Doctoral Training in Statistics & Operational Research

Submitted for the degree of Doctor of Philosophy at

Lancaster University.

November 2019

Abstract

Development of treatments for rare diseases is challenging due to the limited number of patients available. Since a substantial proportion of all patients may be included in the trial, the goal is to treat those patients within the trial as effectively as possible. This motivates the use of response-adaptive designs which skew allocation towards the better performing treatment(s) but often reduce the statistical power. Consequently, this raises the question of how to allocate patients in order to attain a compromise between these conflicting objectives. This can be formalised as a multi-armed bandit problem with the dynamic programming and Gittins index solutions considered here.

Dynamic programming is utilised to propose a randomised design for a two-arm sequential trial with binary outcomes. This design maximises the total number of patient successes and penalises if a minimum number of patients are not allocated to each treatment so that sufficient power is achieved. Moreover, the treatment effect estimator exhibits a very small bias and mean squared error. This design is shown to be fairly robust to delays, with only a slight reduction in patient benefit. Solutions to ameliorate this loss are therefore proposed, for both the fixed and random delay settings.

A design based upon the Gittins index — which is randomised and orientated towards a patient benefit objective — is proposed for normal outcomes, illustrated in the multi-armed setting where patients are allocated in blocks. Patient benefit gains are observed when using this design with a continuous outcome instead of dichotomising it. These gains persist even when missing data is imputed.

Throughout, we compare the proposed designs to alternative designs via extensive simulations in a range of scenarios.

This thesis helps bridge the gap between theory and practice by addressing key issues that have prevented bandit models from being implemented in practice.

For my family;

past, present and future.

Acknowledgements

People usually put their parents last in these things but I will put them first because that is where they belong, above everybody else. I simply do not know how to thank my Mum and Dad, Eve and Kevin, enough. I know they will be relieved to see the back of this thesis, perhaps even more so than me, but I hope they are proud of me for battling on.

I am eternally grateful to Sue and Keith for not just providing me with a room to lodge in over the past eight years, but for so much more than that; the cups of tea, cake, chats, calmness and, above all, welcoming me into their family.

I would like to thank my supervisors, Prof. Thomas Jaki and Dr. Peter Jacko, for their guidance and patience throughout this process. I have been very fortunate to work with both of them and reap the benefits of their complementary skill set. I extend my thanks to Dr. Sofía Villar who I had the pleasure of working with during an internship at the Biostatistics Unit in Cambridge. Her encouragement has been a blessing. My sincere gratitude also goes to Prof. Jon Tawn for his wholehearted compassion and support, right the way through until submission.

I am most thankful to the Engineering and Physical Sciences Research Council

(EPSRC) funding provided by the STOR-i Centre for Doctoral Training which has helped broaden my mind in more than just an academic sense. It has given me the opportunity to travel to places I never imagined I would and meet some of the most fascinating people from all over the world.

I have so much to owe to Lancaster University and, in particular, the Mathematics and Statistics department for nurturing my growth during my undergraduate, masters and doctorate studies. I am proud to have been a part of such a thriving, yet down-to-earth, community.

Finally, to all of the special people that I have lost, and found, along the way; thank you for everything. ♡

Declaration

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

Chapter 3 has been published as Williamson, S. F., Jacko, P., Villar, S. S., and Jaki, T. (2017). A Bayesian adaptive design for clinical trials in rare diseases. *Computational Statistics & Data Analysis*, **113**, 136–153. doi: 10.1016/j.csda.2016.09.006.

Chapter 6 has been published as Williamson, S. F. and Villar, S. S. (2020). A response-adaptive randomization procedure for multi-armed clinical trials with normally distributed outcomes. *Biometrics*, **76**(1), 197–209. doi: 10.1111/biom.13119.

S. Faye Williamson

Contents

List of Figures

List of Tables

List of Recurring Abbreviations

CRDP	Constrained Randomised Dynamic Programming
DP	Dynamic Programming
(D)RPWR	(Delayed) Randomised Play the Winner Rule
FCRDP	Fixed (Delay) Constrained Randomised Dynamic Programming
FLGI	Foward-Looking Gittins Index
FR	Fixed Randomisation
GI	Gittins Index
MAB(P)	Multi-Armed Bandit (Problem)
MDP	Markov Decision Process
MSE	Mean Squared Error
RAR	Response-Adaptive Randomisation
RCRDP	Random (Delay) Constrained Randomised Dynamic Programming
RCT	Randomised Controlled Trial
TS	Thompson Sampling

Chapter 1

Introduction and Motivation

Before any new medical treatment is made available to the public, clinical trials must be undertaken in humans to ensure that the treatment is safe and efficacious (Pocock, 1983). Such trials are usually divided into the following phases. Phase I trials generally involve a small number of healthy volunteers, but in some circumstances, e.g. when testing treatments for a fatal disease, these may be patients who have exhausted other treatment options. The focus of phase I trials is on the study of the pharmacokinetics (i.e. the movement of the treatment through the body), pharmacodynamics (i.e. the treatment's effect on the body) and toxicity of a treatment, with the primary objective being to establish a tolerable dose range. Phase II trials (and onwards) are performed on patients that have the disease of interest. They are initial efficacy studies aimed at determining the dose, and frequency of dosing, required to successfully treat patients (Peace and Chen, 2010). If a treatment is indicated as effective during phase II, then it proceeds to phase III. These are large-scale, costly confirmatory trials which usually compare the experimental treatment to a control (standard treatment or placebo), with the primary objective of confirming the efficacy of the treatment. After the treatment has been approved, its long-term effects in the wider population are monitored during phase IV trials.

In March 2004, the United States Food and Drug Administration (FDA) released the landmark report, *Innovation/Stagnation: Challenge and Opportunity on the Critical Path to New Medical Products* (U.S. Food and Drug Administration, 2004), expressing concern over the slowdown, instead of the expected acceleration, in innovative treatments being submitted to the FDA for approval despite advances in biomedical science. It highlights "an urgent need for improvement in the efficiency and effectiveness of the clinical trial process, including improved trial design" and in particular, "much more attention and creativity need to be applied to disease-specific trial design". Consequently, the FDA released a follow-up document, the *Critical Path Opportunities Report* (U.S. Food and Drug Administration, 2006), indicating that *biomarker development* and *streamlining clinical trials* are the two most important areas for improving medical product development. Streamlining clinical trials includes advancing innovative trial designs, such as *adaptive* designs (Chow and Chang, 2012), which provides the fundamental motivation and underlying theme throughout this thesis.

Adaptive designs have gained increasing popularity amongst researchers, industry and regulatory bodies (Lipsky and Lewis, 2013). In particular, this has been demonstrated by the FDA's recent release of an updated draft guidance on *Adaptive Designs for Clinical Trials of Drugs and Biologics* (U.S. Food and Drug Administration, 2018) as part of their mission to "modernise clinical trials and advance the development of safe, effective drugs". Here, they define an adaptive design as "a clinical trial design that allows for prospectively planned modifications to one or more aspects of the design based on accumulating data from subjects in the trial". In contrast with the traditional approach adopted in clinical trials, the ability to use information dynamically as it accrues to improve efficiency makes adaptive designs a particularly attractive alternative (Pallmann *et al.*, 2018).

The current gold standard design used in clinical trials is the *randomised controlled*

trial (RCT), in which patients are randomised to either the control or experimental treatment(s) in a pre-fixed, and typically equal, proportion. Although this design detects a clinically meaningful treatment difference with a high probability, i.e. it maximises the statistical power of the design under the condition of equal variances (Atkinson and Biswas, 2014), which is of benefit to future patients outside of the trial, it lacks the flexibility to incorporate other desirable criteria, such as the participant's well-being. As such, a large number of patients within the trial are randomised to the inferior treatment, which raises ethical issues. This is particularly concerning in a clinical trial for a *rare* disease in which a substantial proportion of all patients with the disease may be included in the trial, and hence the priority should now be on treating those patients *within* the trial as effectively as possible (Palmer and Rosenberger, 1999). This highlights the inherent conflict present in a clinical trial between *individual ethics* (doing what is best for the patients within the trial) and *collective ethics* (doing what is best for the future target population as a whole), see e.g. Lellouch and Schwartz (1971). RCTs focus on gathering information, and thus place emphasis on the latter (Pullman and Wang, 2001).

The justification for a RCT is that there exists a state of equipoise throughout the trial. Freedman (1987) defines this as a state of genuine uncertainty about which treatment is superior. However, it may be argued that even if there exists a state of equipoise at the beginning of a trial, some idea of which treatment is superior is likely to be obtained as the trial progresses and the data accumulates (the fundamental concept of an adaptive design).

This motivates the use of a particular type of adaptive design, namely, *response-adaptive* designs which take advantage of the accumulating data on patient responses to skew the allocation probabilities towards the better performing treatments, thus reducing patient exposure to seemingly inferior treatments (Rosenberger and Lachin, 1993). The aim is to maximise the expected number of successful responses within

the trial, whilst still maintaining sufficient power. Consequently, this type of design is well-suited for rare disease trials and small population trials more generally, including paediatric trials and trials involving subgroups of common diseases, following the recent surge in personalised medicine (e.g. Lee and Wason, 2019), which "leads toward a fractioning of the target population for each drug" (Zhang *et al.*, 2019). Another pertinent application area that has been highlighted in the literature is in trials for highly contagious diseases, where it is hoped that the disease might be eradicated by the treatment being tested (Berger, 2015). Acute care research, including diseases with high mortality and no existing treatments (Meurer *et al.*, 2012), is a further example where response-adaptive designs may be particularly beneficial (McEvoy *et al.*, 2016). In contrast to RCTs, response-adaptive designs tend to favour individual ethics (Pullman and Wang, 2001).

This raises the question of how to design a clinical trial which provides a compromise between the collective and individual ethics. This is a perfect example of an *exploration versus exploitation trade-off*, which is prevalent in many decision-making problems, since the fundamental tension is between exploiting treatments that have performed well (individual ethics) and exploring new treatments in case they are even better (collective ethics). The formalisation of this problem subsequently became known as the *multi-armed bandit problem* (MABP) which seeks to balance this underlying exploration versus exploitation trade-off in order to provide an optimal allocation rule (Berry and Fristedt, 1985). *Dynamic programming* (Bellman, 1956) is one possible method that can be implemented to obtain the optimal solution of the MABP. However, this approach suffers from the "curse of dimensionality" which limits its practical applicability, particularly when the number of treatment arms is large. Remarkably, Gittins and Jones (1974) showed that an optimal solution exists by decomposing the MABP into smaller sub-problems, thus removing the prohibitive computational complexity. Moreover, it takes the form of an index policy based on

what has become widely known as the *Gittins index*. Both of these solution concepts will be described in further detail in Section 2.2.3, and will form the foundations of the methods proposed in Chapters 3–6.

Indeed, across the bandit literature, the use of bandit solutions to optimally design a clinical trial has been the primary motivation for their study. Gittins (1979) even states that their "chief practical significance is in the context of clinical trials". However, rather ironically, they have never actually been implemented in clinical practice for reasons which will be discussed in Chapter 2. Nevertheless, in recent years, there has been some evidence of progress, and response-adaptive randomised designs based on the MABP, although not optimal, have been implemented in clinical trials (e.g. Barker *et al.*, 2009). Despite these recent advances, bandit theory and clinical trial practice continue to remain relatively separate entities. Therefore, the overarching aim of this thesis is to overcome some of the existing practical barriers and thus bridge the gap between bandit theory and clinical trial practice. As a result, this will contribute to the streamlining of clinical trials in order to improve medical product development, as identified by the *Critical Path Opportunities Report* (U.S. Food and Drug Administration, 2006).

1.1 Outline of Thesis

Chapter 2 aims to introduce the general background information and key concepts which underpin the main ideas proposed in the subsequent chapters of this thesis. The relevant literature with regards to both clinical trials and bandit theory will be outlined in Sections 2.1 and 2.2, respectively, so that the reader is necessarily equipped for the material that follows.

Chapter 3 utilises the dynamic programming solution of the MABP to propose a response-adaptive treatment allocation rule in the context of a two-arm sequential

clinical trial with binary endpoints (i.e. successes or failures) which are assumed to be available immediately. Chapters 4 and 5 extend this modelling framework to encompass the practical issue of delayed patient responses which significantly increases the complexity of the problem. In Chapter 6, attention moves to the alternative solution concept of the MABP, namely, the use of Gittins indices to allocate patients. The focus is now on continuous endpoints, assumed to be normally distributed with unknown mean and variance, predominantly in the multi-armed setting where patients are allocated in blocks rather than sequentially. Additionally, the issue of artificially dichotomising a continuous endpoint and dealing with missing data is touched upon. Simulations in the context of real and hypothetical trials are used throughout to motivate and illustrate the proposed methodology. Chapters 3 and 6 form two self-contained papers which have been reproduced verbatim from the corresponding published versions and as such, there is necessarily some overlapping material.

Chapter 7 concludes this thesis by summarising the main contributions and suggesting avenues for further research.

Chapter 2

Background and Literature Review

2.1 Randomisation in Clinical Trials

The concept of randomisation[1] was popularised by Fisher (1926) in an agricultural study (see e.g. Hall, 2007) and was first considered for use in clinical research by Amberson *et al.* (1931) who randomised patients to treatments using the outcome of a coin toss. However, the first iconic RCT is widely recognised as the streptomycin trial designed by Sir Austin Bradford Hill and conducted by the Medical Research Council (1948) in which random numbers were used to allocate patients.

Since then, randomisation of patients to treatments has been considered paramount in comparative clinical trials in order to: (i) generate comparable groups that are similar in terms of extraneous factors, except for the intervention of the treatment; (ii) minimise several types of bias, e.g. treatment allocation bias[2], which will ultimately add validity to the subsequent statistical tests; and (iii) provide a probabilistic basis for frequentist inference (Rosenberger *et al.*, 2019).

The randomisation methods commonly used in clinical trials can be broadly cat-

[1] Randomisation in clinical trials may refer to either the random selection of patients from the population into the trial or the random allocation of patients to treatments within the trial. We use it exclusively to mean the latter throughout.

[2] Note that some authors, e.g. Chow and Chang (2012) and Rosenberger and Lachin (2016), use the term *selection bias* analogously.

egorised into two groups, namely, conventional (or fixed) randomisation (in which the treatment allocation probabilities remain *constant* throughout the trial, as in the RCT) and adaptive randomisation (in which the treatment allocation probabilities *vary* during the trial) (Chow and Chang, 2012). Examples of adaptive randomisation schemes include: treatment-adaptive (or restricted) randomisation which seeks to balance the sample sizes between treatment groups; covariate-adaptive randomisation which aims to balance covariates of interest between treatment groups by adapting the allocation probabilities according to patient prognostic imbalance; response-adaptive randomisation in which the randomisation probabilities change as patient responses are observed in order to favour the better performing treatments; and covariate-adjusted response-adaptive (CARA) randomisation which is similar to response-adaptive randomisation, but now the patient's covariate profile is also taken into consideration. In the latter case, since the randomisation probabilities depend on responses of patients with similar characteristics, such as certain types of biomarkers, this is an important step towards personalised medicine (see Hu, 2012).

The primary focus of this thesis will be on *response-adaptive randomisation* methods. For an overview of the other adaptive randomisation methods, the reader is referred to Chow and Chang (2012, Chapter 3) or Sverdlov (2015, Chapter 1). CARA designs are also discussed in Hu and Rosenberger (2006, Chapter 9), Antognini and Giovagnoli (2015, Chapter 6) and Rosenberger and Lachin (2016, Chapter 10).

2.1.1 Response-Adaptive Randomisation (RAR)

The exact definition of response-adaptive randomisation[3] (RAR) varies in the literature. Some authors, e.g. Rosenberger and Lachin (2016), use it explicitly to refer to response-adaptive designs that are fully *randomised* (i.e. non-deterministic) so that the allocation probabilities are strictly between 0 and 1. Others, e.g. Coad (2008),

[3]This is sometimes referred to as *outcome-*, or *data-*, *dependent randomisation* within the literature.

use it more generally to refer to any design (whether randomised or not) which uses patient responses to adapt the allocation probabilities towards the most promising treatment(s). Throughout this thesis, we adopt the former interpretation of RAR.

The ultimate aim of RAR is to allocate more patients to the treatment(s) performing better, thus reducing patient exposure to inefficacious treatment(s), without sacrificing randomisation. This is more ethically acceptable compared to conventional randomisation, particularly if a treatment failure represents an extreme, or fatal, outcome (Pullman and Wang, 2001). Although RAR does not fully eliminate the ethical problem of randomising patients to inferior treatment(s), it certainly mitigates it by reducing the probability of allocation to the inferior treatment(s) (Rosenberger and Lachin, 2016). This can be considered a "necessary evil" which ensures a valid comparison between the treatment groups can take place in order to maintain a sufficient level of power at the end of the trial, and hence provides a compromise between the collective and individual patient benefit. Consequently, RAR is subject to attack from both sides of the collective versus individual ethics debate (Tamura *et al.*, 1994) so remains a very controversial subject within statistical and clinical trial communities (see Korn and Freidlin (2011); Berry (2011); Lee *et al.* (2012); Thall *et al.* (2015); Hey and Kimmelman (2015) and corresponding commentaries; London (2018)).

Finally, the adaptive designs guideline by the U.S. Food and Drug Administration (2018) provides an additional pragmatic rationale advocating the use of RAR, namely, that patients may be more willing to enrol in the trial because RAR improves their chance of being allocated to the better treatment, therefore increasing speed and ease of recruitment. This has been demonstrated in various studies; see Tehranisa and Meurer (2014) and McEvoy *et al.* (2016), for example, who illustrate that effectively communicating the randomisation scheme to patients improves their understanding and leads to an even higher participation rate. Ultimately, implementing RAR could help alleviate recruitment problems which poses one of the most challenging aspects

in the conduct of RCTs (see Sully *et al.*, 2013, for example).

However, RAR may lead to *accrual bias* in which participants wait until later on in the trial to enrol since that way, they will have a higher probability of receiving the superior treatment (Rosenberger, 1996). Further, if there is heterogeneity in patient enrolment over time, such as the most severely ill patients enrolling as soon as possible[4], then a bias will be introduced which will affect the validity of the results (see Chappell and Karrison, 2006). One solution is to use CARA randomisation if the underlying covariates causing the heterogeneity are known in advance (Rosenberger *et al.*, 2012, Section 4.3). Examples of recent developments in this area include Villar and Rosenberger (2018) and Villar *et al.* (2018). Alternatively, one may consider using block RAR to reduce the bias caused by population drift (see Magirr, 2011; Korn and Freidlin, 2011, for example).

An important consideration when choosing between conventional randomisation and RAR is the *context* of a clinical trial and, more specifically, the patient horizon (i.e. the total number of patients with the disease of interest both inside and outside the trial). That is, does the trial include essentially every patient who will have the condition of interest during a particular time period? Or is there a large number of patients outside the trial who could gain from the results of the trial? Berry and Eick (1995) compare the performance of conventional equal randomisation, in which half of the patients are randomly allocated to each of the two treatment groups, to four response-adaptive designs. Their main conclusion is that a design employing equal randomisation is very nearly optimal when the condition is relatively common. However, if the condition being treated is rare, then response-adaptive designs can perform substantially better and might be a more suitable alternative. This is because, in the latter case, a substantial proportion of all patients exhibiting the condition are included in the trial. Therefore, *learning* about treatment effectiveness with a view

[4]This is often referred to as *patient (or population) drift*.

to treating patients in the "larger" outside population is much less important. Now, the primary concern is to treat the patients *within* the trial as effectively as possible.

A more recent study by Du *et al.* (2015) obtains similar conclusions when comparing equal randomisation with RAR under fixed patient horizons but varying trial sizes. In particular, they show that equal randomisation is preferred when the number of patients outside the trial is much larger than the number inside the trial, and RAR is favoured for large treatment differences or when the number of patients outside the trial is relatively small.

2.1.2 Examples of RAR Procedures

RAR procedures proposed in the literature generally belong to either one of two main families, namely, those that are: (i) design-driven, i.e. based on an intuitive rule which can be completely non-parametric, or (ii) target-driven, i.e. based on an optimal (or desired) allocation target which depends upon estimated parameters of the assumed response distribution (Rosenberger and Lachin, 2016, Section 10.3). An example of each is provided below.

(i) Randomised Play-the-Winner Rule (RPWR)

The most famous (non-randomised) response-adaptive design is the *play-the-winner rule* (PWR) which was first introduced in a clinical trial context by Zelen (1969). For a clinical trial comparing two treatments (A and B) with binary responses (success or failure) in which patients enter the trial sequentially, the PWR proceeds as follows: a success on a particular treatment causes the next patient to receive the same treatment, whereas a failure on a treatment causes the next patient to receive the alternative treatment. Suppose the first patient is randomly allocated to either treatment A or B with probability 0.5, then an example of a response sequence from a trial employing the PWR is displayed in Table 2.1.1.

Treatment A	Success	Success	Failure		Success	Failure		...
Treatment B				Failure			Success	...

Table 2.1.1: Play-the-winner allocation rule.

A fully randomised version, the *randomised* play-the-winner rule (RPWR), was proposed by Wei and Durham (1978) which has the advantage that it is no longer deterministic (so is less vulnerable to allocation bias, for example). The RPWR has been the most studied urn model in the RAR literature (Rosenberger and Lachin, 2016) and is easily implemented in two-arm trials with binary responses as follows:

1. Initially, an urn contains u balls of type A and u balls of type B. Therefore, clinical equipoise is assumed at the onset of the trial.

2. When a patient enters the trial, a ball is drawn randomly from the urn *with* replacement. If it is a type $i \in \{A, B\}$ ball, the patient receives treatment i.

3. When a patient's response is available, the urn is updated as below:

   (a) A success on treatment A, or a failure on treatment B, generates an additional β type A balls and α type B balls in the urn.

   (b) Similarly, a success on treatment B, or a failure on treatment A, generates an additional β type B balls and α type A balls in the urn, where $0 \leq \alpha \leq \beta$ are integers.

This rule is denoted by RPWR(u, α, β). Therefore, the urn accumulates more balls representing the more successful treatment, thus increasing the probability that a patient will be allocated to the current best treatment. Unlike the PWR, the allocation probability is now a function of all past allocations and responses (rather than depending only on that of the previous patient). In particular, it is proportional to the number of balls of each treatment in the urn.

Moreover, since sampling is with replacement, delayed responses can be accommodated by simply updating the urn composition when the response becomes available. Hardwick *et al.* (2006) refer to this as the *delayed* RPWR (DRPWR). Two different DRPWR models are discussed in the literature; these are summarised in Atkinson and Biswas (2014, Chapter 3). First, Wei (1988) extended the RPWR to incorporate delayed responses by including another set of indicator variables (in addition to the treatment allocation and response indicator variables of the RPWR) which determine whether or not a previous patient's response has been observed before allocation of the next patient. Tamura *et al.* (1994) employ this DRPWR model using response indicators for a surrogate endpoint instead which is observed sooner than the long-term endpoint. Bandyopadhyay and Biswas (1996) introduce a second model which has a slight modification that ensures the denominator of the conditional allocation probability is free of any random variables. Biswas (1999) compares these two models showing that they are asymptotically equivalent and there is no significant difference between their performances. Hence, we will consider the first version of the DRPWR as a comparator in Chapter 4.

Simulation studies, such as those by Rosenberger *et al.* (2001b) and Stallard and Rosenberger (2002), have illustrated that the RPWR may exhibit high variability in the allocation proportions and a significant reduction in power for certain parameter values.In particular, in a two-arm trial with binary endpoints, if the sum of the success probabilities is greater than 3/2, then the asymptotic variance of the allocation proportion depends on the initial urn composition (Hu and Rosenberger, 2003) and a high variability with reduced power is observed (e.g. Rosenberger *et al.*, 2001b, Table 1). In contrast, when the sum of the success probabilities is strictly less than 3/2, the asymptotic variance is independent of the initial urn composition and as such, the RPWR has a smaller variability and larger power. This is illustrated in Coad and Rosenberger (1999, Table 1) in which the power values of the RPWR are

very similar to those attained by conventional randomisation, or Rosenberger and Hu (2004, Table 2). An alternative type of urn model which has the same limiting allocation proportion as the RPWR but exhibits far less variability is the *drop-the-loser rule* proposed by Ivanova (2003). Note that the theoretical relationship between the power and variability of a RAR procedure has been derived in Hu and Rosenberger (2003), confirming that the average power is a decreasing function of the variability (see also Hu and Rosenberger, 2006, Chapter 2).

The practical consequences of using an allocation rule that is too variable are demonstrated by the infamous extracorporeal membrane oxygenation (ECMO) trial (Bartlett *et al.*, 1985) in which the high variability of the RPWR led to an extreme imbalance in the allocation (Hu *et al.*, 2009a). In particular, the investigators chose to use the RPWR(1, 0, 1) design, that is, the urn contained one ball of each type initially and an ECMO (control) ball was added each time a patient survived on ECMO (control) or failed on the control therapy (ECMO). The first patient was randomly allocated to ECMO and survived. The second patient was randomly allocated to the control and died. Hence, this meant that the odds of the next patient being randomly allocated to ECMO were 3:1. All subsequent patients thereon received ECMO by chance and survived. The trial terminated after 12 patients; one control patient who had died and 11 ECMO patients, all of whom survived.

Although the study of Bartlett *et al.* (1985) provided encouraging evidence for the efficacy of ECMO, the results were not convincing due to the very limited comparative data and have since generated much controversy. Unfortunately, this has contributed considerably to the limited application of RAR methods in practice. However, the ECMO trial is atypical of adaptive designs in general and should not constitute a reason to neglect adaptive designs in future modern clinical trials (Rosenberger, 1999). Ultimately, the Bartlett *et al.* (1985) ECMO trial highlights the need for caution when replacing conventional randomisation with adaptive schemes.

(ii) Doubly Adaptive Biased Coin Design (DBCD)

The RPWR above possesses a purely *myopic* structure which means that each patient is allocated to the treatment that is currently performing the best. As we have seen from the ECMO trial, this can result in unfortunate randomisation sequences when applied in practice. The RPWR is not based on any formal optimality criterion and cannot target any pre-specified allocation proportion (Atkinson and Biswas, 2014), so we now turn attention to the second major family of RAR procedures; those that *can* target some desired, often optimal, allocation proportion. A general approach for deriving the optimal allocation proportion is based on the framework proposed in Jennison and Turnbull (2000, Chapter 17); this is discussed further within the RAR context in Hu and Rosenberger (2006, Chapter 2) and Atkinson and Biswas (2014, Chapter 8).

One example is the *doubly adaptive biased coin design* (DBCD) originally proposed for the two-treatment case by Eisele (1994) and further generalised to the multi-treatment case by Hu and Zhang (2004). This design is based on Efron's (1971) biased coin design and is "doubly adaptive" because it depends on both the current allocation proportion and the current estimate of the target allocation proportion (rather than just the former, as in Efron's biased coin design). The basic idea is to define an allocation function, g, from $[0,1]\times[0,1]$ to $[0,1]$ (satisfying certain regularity conditions) which, for every patient, maps the actual allocation proportion and estimated target proportion, so far, to the randomisation probability for the next patient. Since this function involves unknown parameters of the response distribution, which are sequentially updated using the incoming data, these designs require a pre-run of conventional randomisation in order to obtain the initial parameter estimates.

Hu and Zhang (2004) introduced the following allocation function for the two-

treatment case, which is commonly employed in the literature

$$g_\alpha(x,y) = \begin{cases} \frac{y(y/x)^\alpha}{y(y/x)^\alpha + (1-y)((1-y)/(1-x))^\alpha} & \text{if } 0 < x < 1, \\ 1-x & \text{if } x = 0, 1, \end{cases} \tag{2.1.1}$$

where $\alpha \geq 0$ is a tuning parameter that controls the variability of the allocation proportions (as it increases, the variability decreases), x denotes the current allocation proportion and y the current estimated target.

Suppose that the response distribution depends on the unknown parameter vector $\boldsymbol{\theta}$, and $\rho(\boldsymbol{\theta})$ is the target proportion of patients to be allocated to treatment A. Assuming that we have observed j patient responses, N_{Aj} of which are from treatment A, then the DBCD allocates patient $j+1$ to treatment A with probability $g\left(N_{Aj}/j, \rho(\hat{\boldsymbol{\theta}}_{\boldsymbol{j}})\right)$, where $\rho(\hat{\boldsymbol{\theta}}_{\boldsymbol{j}})$ is the estimated target allocation based on the first j patient responses. When $\alpha = 0$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{j}}$ is the maximum likelihood estimator of $\boldsymbol{\theta}$, this reduces to the *sequential maximum likelihood procedure* (Melfi and Page, 2000), i.e. allocating with probability equal to $\rho(\hat{\boldsymbol{\theta}}_{\boldsymbol{j}})$.

The DBCD has the advantage that it can be used to target *any* desired allocation proportion and can be applied to continuous, as well as binary, responses (Hu and Zhang, 2004). Moreover, relative to other RAR procedures (such as the RPWR), it exhibits a smaller variability of the allocation proportions (as shown in Hu and Rosenberger (2003), for example). Hu *et al.* (2009b) proposed an alternative to the DBCD — the *efficient randomised adaptive design* (ERADE) — which can also adapt to any desired allocation proportion but is *asymptotically best*[5] so has even lower variability.

An example of an optimal allocation target for a binary response trial which minimises the expected number of treatment failures for a fixed variance of the test

[5]Asymptotically best procedures attain the lower bound on the asymptotic variance of the allocation proportions (for a particular allocation target); see Hu *et al.* (2006) for details. The drop-the-loser rule (mentioned on p.14) is an example of an asymptotically best procedure.

statistic is provided in Rosenberger *et al.* (2001b). Zhang and Rosenberger (2006) propose a corresponding version for continuous responses which minimises the total expected response from all patients within the trial (when a smaller response is more desirable to patients). They compare several DBCD procedures theoretically and by simulation for trials with continuous outcomes and conclude that the DBCD targeting the optimal allocation is the best to use in practice. Biswas *et al.* (2007) illustrate that this target allocation proportion is not suitable for normally distributed outcomes in general since it cannot be calculated for negative means. A correction is provided by Biswas and Bhattacharya (2009) which is the version we will implement in Chapter 6 (described therein). A limitation of this optimal allocation proportion is that it is not easily extended to the multi-armed case. However, other target allocation proportions comparing multiple treatments have become increasingly prevalent in the literature; examples include Tymofyeyev *et al.* (2007), Zhu and Hu (2009) and Jeon and Hu (2010). Methods for finding allocation targets are also discussed in the book by Antognini and Giovagnoli (2015).

2.1.3 Bayesian Adaptive Randomisation (BAR)

So far, the RAR procedures discussed have fallen within the frequentist paradigm which is the standard statistical approach to designing and analysing clinical trials (Berry *et al.*, 2011). We now turn our attention to the alternative, and increasingly popular, Bayesian approach which is perfectly suited to online learning and thus lends itself naturally to the adaptive design framework. Under this approach, the unknown parameters of the response distribution are assumed to be random and follow some prior distribution. The incoming data is used to determine the corresponding posterior distribution, according to Bayes' Theorem, and hence the allocation probabilities (which are based on some function of the posterior distributions) are updated. For a thorough overview of Bayesian adaptive methods applied to the design and analysis

of clinical trials, the reader is referred to the book by Berry *et al.* (2011).

(i) Thompson Sampling (TS)

The idea of incorporating RAR within a Bayesian framework, commonly referred to as *Bayesian adaptive randomisation* (BAR), originates from Thompson (1933) who suggested randomising a patient to a treatment based on its posterior probability of being better than the alternative treatment. A sampling method using this concept later became known as Thompson Sampling (TS). Although seemingly attractive, this posterior probability is very variable (particularly earlier in the trial when not much data has been attained) and can lead to more patients being allocated to the inferior treatment. Moreover, using the posterior probability to allocate patients can result in extreme imbalance and hence low statistical power. Thall and Wathen (2007) therefore introduce a tuning parameter to stabilise the allocation probabilities; this is the version we will implement and describe in Chapter 6, where it will be used as a comparator in both the two-arm and multi-arm settings. Other ways to avoid extreme imbalance, e.g. by imposing bounds on the allocation probabilities so they do not converge to 0 or 1, are discussed in Du *et al.* (2015). In multi-arm trials, where there is a shared control group, the power of the trial can be preserved by protecting allocation to the control group (see e.g. Trippa *et al.*, 2012; Villar *et al.*, 2015a; Viele *et al.*, 2020). In this case, the adaptive randomisation scheme is applied amongst the experimental treatments but the allocation to the control is fixed and determined independently. An example now follows.

(ii) Trippa *et al.* Procedure (TP)

Trippa *et al.* (2012) proposed a BAR design which is similar to that of TS since sampling is again from the posterior distribution of the unknown parameters. However, instead of computing the posterior probabilities that arm k is the best (as in

TS), it allocates patients based on the posterior probabilities that each experimental arm is better than the control arm ($k = 0$), given the current observed data, i.e. $\mathbb{P}(\mu_k > \mu_0 \mid \text{data})$ for $k = 1, \ldots, K$. Assume that the trial is composed of j blocks (or stages) testing K experimental treatments against a shared control, where μ_k represents the unknown parameter of the response distribution (e.g. the population mean of treatment k if responses are normally distributed, or the success probability of treatment k if responses are binary) and $n_{k,j}$ patients have been allocated to treatment k by block j, then the probability of allocating treatment k to patients in block j is given by:

$$\pi_{k,j}^{TP} = \frac{\bar{\pi}_{k,j}}{\sum_{k=0}^{K} \bar{\pi}_{k,j}}, \tag{2.1.2}$$

where

$$\bar{\pi}_{k,j} = \begin{cases} \frac{\mathbb{P}(\mu_k > \mu_0 | \text{data})^{\gamma_j}}{\sum_{k=1}^{K} \mathbb{P}(\mu_k > \mu_0 | \text{data})^{\gamma_j}} & \text{if} \quad k = 1, \ldots, K, \\ \frac{1}{K} \left\{ \exp\left(\max(n_{k,j})_{k=1}^{K} - n_{0,j} \right) \right\}^{\eta_j} & \text{if} \quad k = 0, \end{cases} \tag{2.1.3}$$

with $\gamma_j = 3\left(\frac{n_j}{T}\right)^{1.75}$, $\eta_j = \frac{n_j}{4T}$ and $n_j = \sum_{k=0}^{K} n_{k,j}$. For information on the selection of these tuning parameters, refer to Wason and Trippa (2014) or the online Appendix of Trippa *et al.* (2012). The fundamental idea is that they tune the exploration versus exploitation trade-off inherent in the randomisation procedure. For example, when $\gamma_j = 0$, then the patients in block j will be randomly allocated to each of the experimental arms with identical probabilities, i.e. equal, fixed randomisation, which makes sense during the initial exploratory stage of the trial ($j = 1$) when no responses have yet been observed. However, during later stages of the trial, larger values of γ are preferable in order to exploit the information contained in the observed responses by giving rise to larger allocation probabilities to the better arms. At the extreme, as $\gamma \to \infty$, patients would be randomly allocated to either the best experimental treatment or the control arm only. Thus, the chosen value of γ needs to lie between

these two extremes.

The purpose of the expression for $\bar{\pi}_{0,j}$ is to protect allocation to the control arm (and thus preserve power) since if the difference between the number of patients on the control, $n_{0,j}$, and the number of patients on the most commonly used experimental arm, $\max(n_{k,j})_{k=1}^{K}$, is too large, then the allocation procedure will try to compensate for this by allocating a larger number of patients to the control arm to make the size of these two groups more comparable.

We will implement the TP with allocation probabilities as defined in (2.1.3) in Chapter 6 to be used as a comparator in the multi-armed setting.

BAR schemes, such as TS and TP, have become increasingly popular in practice (see Biswas *et al.*, 2009; Lee *et al.*, 2010) and have been implemented in several real-life trials, particularly cancer trials, to allocate more patients to treatments that have performed well for similar patients (Wason *et al.*, 2015). Notable examples include the I-SPY 2 (Barker *et al.*, 2009), BATTLE (Kim *et al.*, 2011) and BATTLE-2 (Papadimitrakopoulou *et al.*, 2016) trials. These trials utilise designs that match patients with the most appropriate treatment for them according to their biomarker profiles and are thus geared towards personalised medicine (Zhou *et al.*, 2008). The BAR design that is implemented in the BATTLE-2 trial is described in Gu *et al.* (2016). For brief reviews of the I-SPY 2 and BATTLE trials, see Berry *et al.* (2011, Chapter 4).

Another type of Bayesian adaptive design is a *bandit* allocation rule which utilises prior information on the unknown parameters in combination with the accruing patient observations to ascertain the optimal treatment allocation at each stage of the trial (see e.g. Hardwick and Stout, 1991; Zhang *et al.*, 2019). Bandit rules are central to the methods proposed in this thesis and therefore we provide the fundamental concepts in the following section.

2.2 Bandit Models

2.2.1 The Multi-Armed Bandit Problem (MABP)

The *multi-armed bandit problem* (MABP) owes its name to its resemblance to the situation faced by a gambler with a choice between several slot machines (or "one-armed bandits"). It is a sequential decision problem in which, at each time, a player must decide which bandit to operate next in order to maximise their total expected winnings (reward) over the whole time horizon. Do they operate one which has performed well in the past so has the largest posterior mean of winning, i.e. exploit, or one with a larger posterior variance which therefore has the potential to perform even better, i.e. explore? Considering only the former leads to a *myopic*, or *one-step-look-ahead*, policy which seeks solely to maximise the *immediate* reward and is not necessarily globally optimal (Berry and Fristedt, 1985). All of the rules discussed in Section 2.1.2 were of this form.

The MABP, however, provides a mathematical formulation of this inherent *exploitation versus exploration trade-off*[6] which aims to balance these competing goals and maximise the total reward in order to obtain an *optimal policy*. This policy accounts for the fact that gaining new information could potentially lead to greater rewards in the future. Consider any situation which requires a decision to be made, e.g. choosing which chocolate bar to purchase in a shop, and notice that this trade-off is prevalent in most real-life decision-making problems (see Cohen *et al.*, 2007), irrespective of the context, as reflected in Whittle's (1982) statement that the MABP "embodies in essential form a conflict evident in all human action". This therefore makes the MABP an extremely useful and important problem to solve, and explains why it has attracted so much attention from a wide range of disciplines; see Gittins *et al.* (2011, Chapter 9) and Lattimore and Szepesvári (2019, Section 1.2) for

[6]Depending on the context, other terminologies for this trade-off may instead be adopted, e.g. earn versus learn.

examples.

The application area that we will be focussing on throughout this thesis is the design of clinical trials and, more specifically, how to allocate patients to treatments[7] (or arms) in order to optimise some pre-determined performance criterion. Attention will be centred around a patient benefit criterion, such as maximising the number of successful responses[8] from patients within the trial or, equivalently, the proportion of patients allocated to the superior treatment (if it exists), but there are many other objectives that one may wish to optimise over. Interestingly, the problem of sequentially allocating patients within a clinical trial provided the initial impetus for the study of MABPs, first posed by Thompson (1933) and subsequently developed by Robbins (1952), in which the term "bandits" did not yet even appear.

The *classic* MABP formulation assumes that any arm which is not selected remains *passive*, that is, it does not change state or produce any reward. An important generalisation relaxes this assumption so that passive arms can also change state, and more than one arm can be activated at any decision time, if appropriate. This gives rise to so-called *restless bandits*, introduced by Whittle (1988), which substantially extends the modelling power of MABPs so that they can be applied to a much wider variety of practical problems. For example, the restless bandit framework can incorporate finite horizons (unlike the classic MABP which assumes an infinite horizon) and delayed feedback, both of which are particularly relevant to the clinical trial setting and hence will be considered in this thesis. Restless bandits are discussed in further detail in Gittins *et al.* (2011, Chapter 6).

[7] Note that the terms "treatments", "arms" and "bandits" may be used interchangeably.

[8] Within the Biostatistics literature, the term patient *response* is often used to imply a patient *success*. However, throughout this thesis we refer to it in its most general sense to mean either a success or a failure.

2.2.2 Markov Decision Processes (MDPs)

MABPs are typically modelled as *Markov decision processes* (MDPs), which are extensions of Markov processes to include a set of decisions (or actions) and associated rewards at each stage. Therefore, to formulate an MDP, the following quintuple must be defined: decision epochs, states, actions, transition probabilities and rewards. A detailed description of these are provided in Puterman (2014, Chapter 2) and summarised briefly below. Decision epochs t are simply the points in time at which decisions are made and will be referred to as "time t", where $t \in \mathcal{T} \equiv \{0, 1, \ldots, T\}$, $T \leq \infty$[9]. The states at time t, $\boldsymbol{z}_t$, contain all of the information required to be able to choose an action a from the set of available actions $\mathcal{A}$. In the clinical trial context, the state represents one's state of knowledge about the effectiveness of the corresponding treatment (which is updated once the patient's response has been observed), and an action corresponds to allocating a patient to a treatment. These actions can be deterministic (if they are selected with certainty) or randomised (in which case each action is selected with some probability, e.g. Cheng and Berry (2007)). In the clinical trial context, it is desirable for actions to be randomised, that is, patients should be randomly allocated to treatment arms, for the reasons outlined in Section 2.1. Deterministic actions would enable the treatment allocation sequence to be predicted, and therefore unmasked, if the state of the trial was known. In Chapter 3, we explore this issue further, showing how randomisation can be introduced and what effect this has on the behaviour of the proposed RAR design. As a result of the action taken at time t, a_t: (i) the system transitions to a new state at time $t+1$, $\boldsymbol{z}_{t+1}$, according to the transition probability $\mathbb{P}(\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_t, a_t)$, and (ii) some reward $\mathcal{R}^{a_t}(\boldsymbol{z}_t)$ accrues which provides the basis for evaluating the chosen action. The transition probabilities and rewards at each t depend only on the *current* state and action chosen in that state, thus giving rise to a Markovian ("memoryless") system.

[9]Note that although decisions are not made at decision epoch T, it is included here for completeness so that the final state of the system can be evaluated (Puterman, 2014, Section 2.1.1).

The time horizon, T, of an MDP can be finite or infinite. In the latter case, rewards are usually discounted by introducing a discount factor $d \in (0, 1)$ which ensures that the *total* reward obtained is finite. To complete the Markov decision problem formulation, an optimality criterion (or objective) needs to be specified which is partly determined by the time horizon. Assuming an *infinite horizon* and following Bellman (1956), the typical objective of the classic MABP is to maximise the *expected total discounted reward* over the infinite horizon, which is discussed in Gittins *et al.* (2011, Chapter 2) and Puterman (2014, Chapter 7). An alternative objective, however, is to consider the (long-run) *average expected reward* over the infinite horizon (Puterman, 2014, Chapter 8).

In *finite horizon* problems, which will be the focus of this thesis, interest is in the *expected (discounted) total reward* (Puterman, 2014, Chapter 4). Suppose that the system is in state $\boldsymbol{z}$ at time t and $\mathbb{E}^{\pi}$ represents the expectation under *policy*[10] $\pi \in \Pi$ (where Π is the set of *past-measurable*[11] policies), then the expected total discounted reward over the remainder of the time horizon $T - t$ is

$$\mathcal{V}_t^{\pi}(\boldsymbol{z}) = \mathbb{E}^{\pi}\left[\sum_{u=t}^{T} d^u \mathcal{R}^{a_u}(\boldsymbol{z}_u) \;\middle|\; \boldsymbol{z}_t = \boldsymbol{z}\right], \tag{2.2.1}$$

where a_u denotes the action that is chosen at time u ($u = t, \ldots, T$) under policy π and $\mathcal{R}^{a_u}(\boldsymbol{z}_u)$ is the reward received from *all* arms when action a_u is taken. In the classic MABP formulation, when actions are deterministic and rewards are immediate, this is simply the reward from the arm corresponding to the chosen (or active) action.

Note that in the finite horizon case, rewards are not necessarily discounted as in the infinite horizon case. The *undiscounted* finite horizon objective, which is equivalent to substituting $d = 1$ into (2.2.1) and sometimes referred to as *uniform discounting*

[10] A *policy* is any rule that determines which action to take given the information available in state $\boldsymbol{z}$ at time t, i.e. it is a mapping from states to actions.

[11] The action prescribed by a *past-measurable* policy at time t does not depend on what happens after t. This is also known as history-dependent (Puterman, 2014) or non-anticipating (Jacko, 2019b).

(e.g. Wang, 1991b; Hardwick and Stout, 1991), is the most pertinent in many applications (Jacko, 2019b). This includes the clinical trial setting, in which case there is: (i) a finite horizon since there is a pre-determined finite number of patients in the trial, and (ii) uniform discounting since each patient response carries the same weight (Hardwick, 1995). Therefore, throughout this thesis, attention is restricted to the finite horizon problem with the principal objective being to maximise the expected total reward in (2.2.1) which, as previously mentioned, in the clinical trial setting translates to maximising the expected total patient benefit. A thorough examination of the finite horizon bandit problem from a statistical and theoretical perspective, within the clinical trial setting, is provided in the book by Berry and Fristedt (1985) (which includes an extensive annotated bibliography).

Maximising the expected total reward over the specified time horizon T gives rise to the *optimal policy*[12]; Section 2.2.3 below discusses how this can be obtained. Note that throughout this thesis, it is assumed that T is the total number of patients *inside* the trial, n. However, one may also wish to incorporate what happens *after* the trial, in which case T would represent the total number of patients both inside *and* outside the trial, N, so that the optimal criterion is defined for the entire patient population instead. Such a criterion is considered in Berry and Eick (1995), Cheng and Berry (2007) and Zhang *et al.* (2019), for example, where it is assumed that the patients outside the trial will receive the treatment that performed best during the trial. Thus, the number of successes expected after the n patients in the trial have responded is taken to be the size of the remaining population, $N - n$, multiplied by the maximum current estimate of the treatment success rates. An example of this type of optimal response-adaptive allocation procedure, formulated as a two-arm bandit problem, is the *robust Bayes* (RB) procedure which is described and compared to other randomisation procedures in Berry and Eick (1995). A toy example of the

[12]The existence of an optimal policy for a finite horizon MDP is shown in Berry and Fristedt (1985, Chapter 2).

RB procedure, illustrating how the patients in a trial should be allocated in order to maximise the expected number of patient successes over N, is presented in Berry and Stangl (1996, pp. 25–29) when $n = 7$ and $N = 100$. More recently, Zhang *et al.* (2019) implemented the optimal design of Berry and Eick (1995) to investigate how the size of the patient horizon affects the power and patient benefit trade-off. This procedure hinges on the method of dynamic programming which is described in the following section.

2.2.3 Solution Methods to the MABP

Two possible solution methods to the MABP are now discussed. The first — dynamic programming — is an exact approach, giving rise to a Bayes-optimal solution (see Jacko, 2019b, Section 7.2) and will be used to implement the methods proposed in Chapters 3–5. The alternative index-based solution, however, yields a near-optimal approximation and will be utilised in Chapter 6.

In contrast with most RAR procedures in the literature, including those introduced in Section 2.1.2, bandit solutions have the advantage that they *look-ahead*, or are *forward-looking*, since they balance the myopic goal with future rewards (Hu and Rosenberger, 2003). In other words, they maximise not only the immediate reward but the cumulative reward, which takes account of all possible future rewards.

(i) Dynamic Programming (DP) Approach

Since MABPs can be formulated as MDPs, they are, in principle, amenable to solution by the standard dynamic programming (DP) technique which was developed by Bellman (1956) (and popularised in the classic book by Bellman (1957)). Informally, this approach involves breaking the problem down into a series of smaller sub-problems, each of which are solved (and the solution stored) to yield the complete solution to the original problem. Decomposing the problem in this way and storing the solutions

of the sub-problems so that they can be re-used reduces the computational burden considerably.

More formally, DP is based upon calculation of a *value function*, $\mathcal{F}_t$, which represents the best possible value of the objective in (2.2.1), i.e. the maximum expected total reward, over the set of all policies π for every possible state at time t. When starting in state $\boldsymbol{z}$ at time t and following policy π thereafter, the value function is defined as

$$\mathcal{F}_t(\boldsymbol{z}) = \max_{\pi \in \Pi} \mathcal{V}_t^{\pi}(\boldsymbol{z}) = \max_{\pi \in \Pi} \mathbb{E}^{\pi} \left[\sum_{u=t}^{T} d^u \mathcal{R}^{a_u}(\boldsymbol{z}_u) \;\middle|\; \boldsymbol{z}_t = \boldsymbol{z} \right].$$

A fundamental property of the value function is that it satisfies the following recursive relationship, which is commonly referred to as the *Bellman equation*[13]

$$\mathcal{F}_t(\boldsymbol{z}) = \max_{a \in \mathcal{A}} \left\{ \mathcal{R}^a(\boldsymbol{z}) + d \sum_{\boldsymbol{z}'} \mathbb{P}(\boldsymbol{z}' \mid \boldsymbol{z}, a) \mathcal{F}_{t+1}(\boldsymbol{z}') \right\} \quad \text{for} \;\; 0 \leq t \leq T-1, \qquad (2.2.2)$$

where $\mathbb{P}(\boldsymbol{z}' \mid \boldsymbol{z}, a)$ is the transition probability of moving from state $\boldsymbol{z}$ at time t to some new state $\boldsymbol{z}'$ at time $t+1$ under action a, and $\mathcal{A}$ is the action space containing all available actions. Intuitively, the Bellman equation expresses a relationship between the value of a state and the value of its successor states, which is the essence of DP. It is helpful to notice that equation (2.2.2) comprises of two parts: (i) the *immediate reward* $\mathcal{R}^a(\boldsymbol{z})$ received by choosing action a when in state $\boldsymbol{z}$, plus (ii) the expected (discounted) *future reward* earned from the successor states as a result of taking this action. The second term contains the product of the probability of being in state $\boldsymbol{z}'$ at time $t+1$ if action a is taken, and the expected total reward obtained if policy π is followed from time $t+1$ to T when the "new" starting state is $\boldsymbol{z}'$. The idea is that, in every state, the action which maximises the expected combination of immediate and

[13] This was originally termed the *functional equation* by Bellman (1957). Alternative names also include the *fundamental equation of dynamic programming* (Berry and Fristedt, 1985); the *dynamic programming equation* (Gittins *et al.*, 2011); the *optimality equation* (Puterman, 2014), etc.

future rewards is chosen. In finite horizon problems, no action is taken at time $t = T$ and so the final decision occurs at time $T - 1$. Therefore, the terminal reward at time $t = T$ is a function of the state only, that is, $\mathcal{F}_T(\boldsymbol{z}) = \mathcal{R}(\boldsymbol{z})$. This is sometimes referred to as the salvage (or scrap) value in the operational research literature (Puterman, 2014). Although the terminal reward is usually 0, there are instances when this is not the case, e.g. if an artificial terminal reward (i.e. penalty) is introduced to avoid certain states (as in Chapter 3) or the reward is not obtained immediately (as in Chapters 4 and 5).

The ultimate optimisation problem is to find the maximum expected total reward over the entire time horizon when $t = 0$ for a given initial state $\boldsymbol{z}_0 = \boldsymbol{z}$, that is, $\mathcal{F}_0(\boldsymbol{z})$. By calculating this value, the policy that gives rise to it, namely, the complete optimal policy π^*, is also found and can be expressed as $\pi^*(\boldsymbol{z}_0) \equiv \arg\max_{\pi \in \Pi} \mathcal{V}_0^{\pi}(\boldsymbol{z})$. Note that the initial state $\boldsymbol{z}_0$ is usually a very natural choice and has just one possibility, as in the clinical trial setting when there are no observations before the start of the trial at time $t = 0$. However, in some applications, the initial state may be less obvious and could take, for example, a set over some distribution.

When the horizon is finite, this can be solved exactly using *backward induction*, in which the value function is evaluated recursively by first determining the maximum expected reward (together with the corresponding actions) at the *final* time period[14] of the decision process for all possible states. Proceeding towards the penultimate period, the maximum expected reward for every possible state is again calculated, but this time incorporating the information just obtained for the subsequent period as well. Continuing backwards in time until the start of the problem at time $t = 0$ allows the optimal reward $\mathcal{F}_0(\boldsymbol{z}_0)$, along with the corresponding optimal policy (or policies), to be determined for every state at every point in time.

More formally, the backwards induction algorithm can be summarised as follows:

[14] A *period* (or stage) represents the time between two consecutive decision epochs.

1. Let $t = T$ and $\mathcal{F}_T(\boldsymbol{z}) = \mathcal{R}(\boldsymbol{z})$ for all $\boldsymbol{z} = \boldsymbol{z}_T$,

2. For $t = T-1,\ T-2 \ldots, 0$ and for each $\boldsymbol{z} = \boldsymbol{z}_t$, calculate:

   (a)
   $$\mathcal{F}_t(\boldsymbol{z}) = \max_{a \in \mathcal{A}} \left\{ \mathcal{R}^a(\boldsymbol{z}) + d \sum_{\boldsymbol{z}'} \mathbb{P}(\boldsymbol{z}' \mid \boldsymbol{z}, a) \mathcal{F}_{t+1}(\boldsymbol{z}') \right\},$$

   (b)
   $$\pi^*(\boldsymbol{z}) \equiv \arg\max_{a \in \mathcal{A}} \left\{ \mathcal{R}^a(\boldsymbol{z}) + d \sum_{\boldsymbol{z}'} \mathbb{P}(\boldsymbol{z}' \mid \boldsymbol{z}, a) \mathcal{F}_{t+1}(\boldsymbol{z}') \right\},$$

3. If $t = 0$, stop. Otherwise, repeat Step 2.

At the end of the algorithm, $\mathcal{F}_0(\boldsymbol{z})$ will contain the maximum expected discounted sum of rewards received by following policy $\pi^*(\boldsymbol{z})$ from state $\boldsymbol{z} = \boldsymbol{z}_0$.

Thus, when implementing this algorithm computationally, two multi-dimensional arrays, indexed by state, need to be created: (a) the value $\mathcal{F}$ containing the maximum expected total reward for the corresponding combination of states, and (b) the optimal policy π^* containing the actions which give rise to these values.

An illustrative example of the backward induction algorithm applied to the finite horizon two-armed bandit problem of optimally allocating patients in a clinical trial, i.e. when $T = n$, is provided in Appendix 3.6.1 of Chapter 3. See also the example provided in Berry and Stangl (1996, pp. 25–29) which includes accompanying figures to demonstrate how backwards induction is used to optimally allocate patients over the entire patient horizon, i.e. when $T = N$.

The DP approach requires a considerable amount of computational power and memory to calculate and store the solution, even for relatively small problems. For example, consider the problem of allocating patients in a clinical trial of size $T = n$ with two treatments (A and B) available and binary responses (success or failure). At each time t, there will be four possibilities (success on A, failure on A, success on B or failure on B) and hence, 4^n possible paths which need to be enumerated

in order to determine the optimal policy[15]. As the number of arms increase, the size of the problem grows exponentially; a phenomenon referred to as the *curse of dimensionality* (Bellman, 1961). See Villar *et al.* (2015a, Figure 1) which illustrates how the computational requirements of DP rapidly increase with T, even for a small number of arms. Furthermore, Zhang *et al.* (2019, Section 3.3) find that "with three arms and sample sizes of 100, it becomes infeasible". For this reason, DP is often thought to be of limited applicability in practice. However, with the advancement in modern day computers, DP methods can be used to solve MDPs with millions of states (Sutton and Barto, 2017) and the survey by Jacko (2019b) shows that DP solutions are tractable for much larger horizons than are commonly believed. For example, the author develops a package which computes the DP design in a few minutes for $T \approx 1000$, a few hours for $T \approx 2000$ or a few days for $T \approx 4000$ (refer to Jacko (2019a) for details of implementation).

When the dimension of the state space is too large to deem exact DP methods suitable, the value function can be approximated using a heuristic algorithm instead. A plethora of such algorithms (e.g. TS discussed in Section 2.1.3) prevail the bandit and operational research literature, and many fall under the umbrella term of *approximate dynamic programming* (ADP), but we do not go into details here since they are beyond the scope of this thesis. The interested reader is referred to Powell (2011) for an accessible introduction to ADP.

Some of the most popular bandit algorithms have been evaluated within a clinical trial context by Kuleshov and Precup (2000) to determine whether they constitute effective adaptive trial strategies. Simulation studies showed that they all performed similarly. In particular, they successfully treated at least 50% more patients, resulted in fewer adverse effects and greater patient retention compared to fixed randomisation, but had more difficulty identifying the superior treatment. Even though the bandit

[15]In the clinical trial context, the policy is synonymous to the allocation rule, or design, that specifies which treatment arm each patient receives.

algorithms received delayed feedback (since the response was observed 13 days after administration of the treatment), this had minimal impact on their effectiveness. In addition, it is worth noting that patient dropout was interpreted as a treatment failure which is a common assumption in trials with a binary response, and we discuss a possible way to deal with this for continuous responses in Chapter 6. See Kaibel and Biemann (2019) for another simulation study comparing a range of bandit algorithms with fixed randomisation, but this time for normally distributed outcomes.

A further impediment to the use of the DP design in clinical trial practice is that "it is difficult to describe and cumbersome to communicate" (Berry, 1978). Therefore, an alternative solution which both reduces the computational difficulties and is simpler to communicate is now discussed below.

(ii) Index-Based Approach

A key breakthrough for the *infinite horizon* MABP was provided by Gittins and Jones (1974), who showed that instead of solving the K-dimensional MDP (where K is the number of arms), an optimal solution can be found by decomposing this into K one-dimensional optimisation problems (where the computational cost now increases linearly with K rather than exponentially). Remarkably, it was shown that the optimal policy obtained by backward induction is equivalent to an *index policy*. That is, an index can be computed separately for each arm as a function only of its current state, such that the optimal policy is always to continue the arm with the largest current index. This was originally called the *dynamic allocation index* by Gittins (1979) and subsequently named the *Gittins index* by Whittle (1980), which is how we will refer to it hereafter (and how it widely appears in the literature). The

Gittins index for arm k ($k = 1, \ldots, K$) is defined by

$$\mathcal{G}(\boldsymbol{z}^k) = \sup_{\tau > 0} \frac{\mathbb{E}\left[\sum_{t=0}^{\tau-1} d^t \mathcal{R}_k(\boldsymbol{z}_t^k) \;\middle|\; \boldsymbol{z}_0^k = \boldsymbol{z}^k\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau-1} d^t \;\middle|\; \boldsymbol{z}_0^k = \boldsymbol{z}^k\right]}, \tag{2.2.3}$$

where τ is a stopping time and $\mathcal{R}_k(\boldsymbol{z}_t^k)$ is the immediate reward obtained from allocating arm k when in state $\boldsymbol{z}$ at time t. Here, the rewards are *geometrically* discounted[16] (which is the second most frequently considered discount sequence in the literature (Berry and Fristedt, 1985)). Note that the numerator in equation (2.2.3) represents the expected total discounted *reward* up to τ, whilst the denominator represents the expected total discounted *time* up to τ. Thus, the Gittins index is interpreted as the maximum expected reward per unit of discounted time when starting from the initial state.

For a given discount factor d, the method of calculating these indices is described in Gittins (1979) and Gittins *et al.* (2011, Chapter 7). In addition to Bernoulli endpoints, Gittins indices have been derived for a variety of others, including: normal (with known (Jones, 1970) and unknown variance (Jones, 1975)), multinomial (Glazebrook, 1978) and exponential (Amaral, 1985). Tables containing the calculated Gittins index values are provided in Gittins *et al.* (2011). Since the Gittins index is independent of K, the relevant table can be used for all possible trials, which reduces the computational requirements even further (Villar *et al.*, 2015a).

When using Gittins indices to solve the K-armed MABP in a clinical trial context, the Gittins index theorem no longer applies because the horizon is finite and hence, the solution obtained will not be optimal. However, although not optimal, it can still be used to *approximate* the optimal policy when applied with a truncated horizon $T < \infty$ instead, and, as Bather (1981) stated, "the principle can still be effective". We now provide some examples, both past and present, of treatment allocation rules

[16] Berry and Fristedt (1985, Theorem 6.2.1) show that for the classical MABP, Gittins indices are optimal only if the discount sequence is geometric.

that are based on Gittins indices, and outline what remains to be done in the future to increase their desirability of being applied in practice.

Robinson (1983) was the first to consider the relative merits of Gittins indices for a *finite* horizon. In particular, sequential allocation rules based on Gittins indices for a Bernoulli two-armed bandit problem using discount factors 0.99, 0.995 and 0.9999 were compared against three different adaptive rules (as well as equal allocation). The appropriate choice of d remained unclear to the author, although he did comment that "for medical applications, it should be near one". Berry and Fristedt (1985, pp. 249-250) queried the choice of discount factor and suggested that a more reasonable choice may be $1 - 1/T$, which was also suggested by Wang (1991b). As with most of the early literature on response-adaptive allocation rules for Bernoulli responses, these comparisons were based on two criteria: expected successes lost[17] (a measure of the patient benefit; the smaller, the better) and the error probability (the probability that the inferior treatment has the higher proportion of successes). Simulation results showed that the Gittins index rules had slightly larger error probabilities, but considerably smaller expected successes lost, than equal allocation, even when there was a substantial delay in observing the response. Moreover, the author commented that this rule is easy to use given the availability of tables for the indices, and performs well for a wide range of model parameters.

Hardwick and Stout (1991) and Hardwick (1995) also consider an allocation rule based on approximations to the Gittins indices using its lower bounds[18] (which are easier to compute) and compare it to the optimal rule based upon the DP solution to the finite horizon two-armed bandit problem. They show that their proposed rule satisfies both power and patient benefit criteria adequately, thus providing a good compromise. Therefore, as Hardwick (1995) points out, the Gittins lower bound (or

[17] Note that minimising the expected successes lost is equivalent to maximising the expected number of successes (which is the criterion we will consider).

[18] The general expression for the lower bound given an arbitrary prior distribution is provided in Berry and Fristedt (1985, Example 5.4.6).

modified bandit) allocation rule "has a special appeal for clinical trial applications since it can be viewed as offering an ethically equitable mechanism for balancing outcomes of present and future patients".

Wang (1991b) advocates the use of Gittins indices as "a promising choice to use in clinical trials" and consequently proposes an adaptive allocation rule using Gittins indices with discussion on the appropriate choice of discount factor. In particular, he suggests that although a reasonable choice of d is $1 - 1/T$, it is not the best way to use Gittins indices to approximate the optimal solution and ideally, the discount factor should get smaller as fewer patients remain in the trial. Therefore, he recommends choosing the discount factor dynamically, based on the number of patients remaining in the trial, which he refers to as the *dynamic* Gittins index (DGI) allocation. Simulation results comparing the DGI with the optimal solution obtained using DP for small trial sizes (up to $T = 20$) reveal that the difference between DGI and the optimal policy is negligible in terms of the expected total number of successes received. Moreover, he concludes that if the disease is rare, in which case the focus is on treating patients in the trial as effectively as possible, then Gittins indices should be used. Alternatively, if the disease is common, he suggests using the least failures rule, i.e. the limit case of the Gittins index rule as $d \to 1$ (Kelly, 1981).

In another paper by Wang (1991a), it is shown that introducing a constraint parameter into the Gittins index rules significantly reduces the error probability incurred when using Gittins indices to allocate patients. Wang (1991a) also generalises these rules to the case when the response distribution is unknown. This *constrained Gittins index* rule is further explored by Coad (1991b, 1995) in the normal response setting and implemented in Chapter 6 of this thesis as a comparator method (where a description of the method is provided). Coad (1992) also studied the effect of linear time trends on sequential allocation rules based on Gittins indices and considered analysing the data in blocks (as we do in Chapter 6) as a means of ameliorating the

effect of time trends.

An important extension of the Gittins index was introduced by Whittle (1988) who proposed a heuristic rule, known as the *Whittle index*, as a solution to the multi-armed *restless* bandit problem (see Section 2.2.1). This reduces to the Gittins index in the classic case when passive bandits remain frozen. Although the Whittle index policy is not optimal in general, Weber and Weiss (1990) proved that it is asymptotically optimal under certain conditions. Refer to Gittins *et al.* (2011, Chapter 6) for further details. In Chapter 3, the Whittle index policy is considered as a comparator instead of the Gittins index because the corresponding MABP is restless due to the following reasons: (i) the horizon is finite and so the number of patients in the trial remaining to be treated is included as a state variable (which changes for all arms at each t, regardless of the action taken), and (ii) actions are randomised meaning that more than one arm can change state during a time period, thus removing the one-to-one correspondence between the action and arm chosen that exists in the deterministic case. Moreover, the Gittins index theorem only applies when actions are deterministic.

The relative advantages and disadvantages of using Gittins and Whittle index policies for the classic and restless Bernoulli MABP, respectively, as potential patient allocation rules are discussed at length in the paper by Villar *et al.* (2015a). In the former case, the horizon is truncated and in the latter case, the finite horizon Bernoulli MABP is reformulated as an equivalent infinite horizon restless MABP. Simulation results, in both the two-armed and multi-armed settings, show that the index-based policies perform extremely well with respect to the patient benefit criteria, and start to skew patient allocation towards the superior arm earlier than the alternative adaptive designs considered. The increase in the expected number of patient successes relative to the other designs considered was most pronounced in the multi-armed case. However, they suffer from a severe reduction in power which severely hinders their use in practice. Therefore, the authors suggest a modified version of the Gittins index rule

— the *controlled Gittins* approach — which protects allocation to the control group in a similar vein to the aforementioned TP in Section 2.1.3(ii), and thus improves the power. A similar approach is considered in Chapter 6.

The performance of the Gittins and Whittle index rules, amongst others, is further explored in Villar (2018) through both exact and simulated calculations, with a particular focus on their application to rare disease trials. Although the Gittins and Whittle index policies behave similarly, the Whittle index is shown to always outperform the Gittins index in terms of the expected proportion of patients allocated to the best arm. Moreover, simulation results in the two-armed case show that both the Gittins and Whittle index rules are almost identical to the optimal rule obtained by DP, with the sub-optimality gap increasing slightly in the multi-armed case.

All of the aforementioned index-based allocation rules are deterministic which is a major barrier to their implementation in practice. Although semi-randomised index-based rules have been proposed in the literature, e.g. Glazebrook (1980); Bather (1980, 1981), in which random perturbations are added to the index value, they are not fully randomised since they are not expressed in terms of allocation probabilities. Villar *et al.* (2015b), however, present a fully randomised design using Gittins indices — the *forward-looking Gittins index* (FLGI) — which is applied to blocks of patients, rather than individuals. Simulation results show that the FLGI continues to increase the number of patient successes significantly compared to alternative adaptive randomised designs (including TS and TP described in Section 2.1.3), yet fails to meet the required power level. This design is discussed further in Chapter 6 where it forms the foundations of the method proposed. An extension of the FLGI to incorporate binary covariates has also been suggested by Villar and Rosenberger (2018) which is beyond the scope of this thesis but forms an ongoing area of research.

With the exception of the constrained Gittins index, the above index-based allocation rules focus only on the Bernoulli bandit problem and examples of applying them

to endpoints other than binary within the clinical trial setting are limited. A recent example, however, is provided by Smith and Villar (2018) who investigate Gittins index-based allocation rules for the case when the outcome is normally distributed with known variance. The results support that the Gittins index-based designs achieve the largest patient benefit relative to alternative designs used in clinical trial practice, especially in the multi-armed case, at the expense of a power reduction.

2.3 Summary

Despite index-based allocation rules being: computationally feasible for multi-armed trials and large T; easy to implement; appealing to use if patient benefit within the trial is a primary concern (as in rare diseases); conceptually simple to summarise to clinicians and patients etc. due to the intuitive nature of always allocating the arm with the largest index, which is paramount since "if a scheme is impracticable then, no matter what its theoretical advantages happen to be, it will not be used" (Upton and Lee, 1981), they are yet to be implemented in clinical trial practice.

Moreover, bandit rules in general have been proposed and studied in the literature for many years so their theoretical properties are very well understood and, as already mentioned, their initial motivation was in the design of clinical trials, which makes us question *why* they have never been applied in practice. As a result of reviewing the pertinent literature, several possible reasons for this have been identified, the main findings of which are now summarised and used to provide the impetus for the methods proposed in the subsequent chapters.

- Patient responses need to be available *immediately*, or at least before the next patient is allocated. This only applies to a small proportion of clinical trials, e.g. some rare disease or paediatric trials, or if the treatment is fast-acting. How to obtain a solution to the exploration versus exploitation trade-off in the presence

of delayed responses, however, is possibly the greatest challenge for both bandit literature and clinical trial practice. Consequently, Chapters 4 and 5 attempt to tackle this problem for the DP approach.

- The allocation of patients to treatments following a bandit rule that is optimal with respect to patient benefit is *deterministic* and most of the bandit literature deals with non-randomised procedures (Rosenberger and Hu, 2004; Rosenberger and Lachin, 2016). Chapter 3 concentrates on introducing randomisation into the DP solution of the two-armed Bernoulli bandit problem, whereas Chapter 6 randomises groups of patients based on probabilities determined by the Gittins index, resulting in a fully randomised Gittins index-based allocation rule.

- Bandit allocation rules result in *insufficient statistical power* to detect a significant treatment difference at the end of the trial. This is a severe limitation from a practical perspective, even if the issue is mitigated in the rare disease setting (since there are comparatively few patients outside the trial), which is where these designs are deemed to be most applicable[19]. Note, however, that this is not limited to bandit rules, it extends to all RAR procedures because it is not possible to maximise both patient benefit and power simultaneously (Hu and Rosenberger, 2003). Therefore, interest is in utilising ways which can improve the power of bandit-based solutions. In the two-arm case of Chapter 3, as well as randomising the allocation probabilities, a constraint is introduced into the value function to avoid extreme imbalance which leads to low power. In the multi-armed case of Chapter 6, we adopt the effective approach identified in the literature of applying the index rule only to the experimental arms whilst fixing allocation to the control arm (see e.g. Viele *et al.*, 2020).

- Bandit allocation rules typically exhibit other undesirable frequentist proper-

[19]For example, in the Discussion of Bather (1981), Prof. D. Berry comments that "the primary hope for future applications [of bandit rules] lies in small clinical trials involving rare diseases".

ties, such as a *lack of type I error control* and *biased estimates* of the treatment effects. However, it is again important to note that this is a problem with (response-) adaptive rules more generally due to the dependence structure induced in the resulting observations (Rosenberger and Lachin, 2016). Investigating methods to minimise the bias of the resulting estimates is beyond the scope of this thesis, but the reader is referred to the following literature: Coad and Ivanova (2001); Bowden and Glimm (2008); Carreras and Brannath (2013); Robertson (2016); Bowden and Trippa (2017); Robertson and Glimm (2019), and references therein. See also Robertson and Wason (2019) for a recently proposed procedure which guarantees strong control of the error rate for RAR trials.

- Most of the relevant research has focused on the simplest context of a *two-armed, sequential* trial with *Bernoulli* responses. This is somewhat restrictive in the clinical setting where other endpoints are also of interest, and increasingly more trials are including multiple arms to improve efficiency in response to the *Critical Path Opportunities Report* (U.S. Food and Drug Administration, 2006), see e.g. Wason and Jaki (2016, 2018). In Chapter 6, we propose and evaluate a Gittins index-based design for normal outcomes (with unknown variance) in multi-armed trials. Moreover, we move from the fully sequential setting to the group sequential setting in which patients are randomised in groups at a finite number of interim analyses.

Chapter 3

A Bayesian Adaptive Design for Clinical Trials in Rare Diseases

3.1 Introduction

Before any new medical treatment is made available to the public, clinical trials must be undertaken to ensure that the treatment is safe and efficacious. Development of treatments for rare diseases is particularly challenging due to the limited number of patients available for experimentation.

The current gold standard design is the randomised controlled trial, in which patients are randomised to either the experimental or control treatment in a pre-fixed proportion. Its main goal is to learn about treatment effectiveness with a view to prioritising future patients outside of the trial. Although this design can detect a significant treatment difference with a high probability, i.e. it maximises the statistical power, which is of benefit to future patients, it lacks the flexibility to incorporate other desirable criteria, such as the trial participant's well-being. As such, a large number of patients within the trial receive the inferior treatment. This is particularly concerning for rare disease trials in which a substantial proportion of all patients with the disease

may be included in the trial. Moreover, there will be fewer patients available outside of the trial to benefit from the learning. Therefore, in this case, the priority should be on treating those patients within the trial as effectively as possible.

This motivates the use of response-adaptive designs for clinical trials involving rare diseases in which the accruing data on patient responses are used to skew the allocation towards the superior treatments, thus reducing patient exposure to inferior treatments. Although it does not fully eliminate the ethical problem of randomising patients to the inferior treatment, it certainly mitigates it by reducing the probability of allocation to the inferior treatment, if it exists.

Berry and Eick (1995) compare the performance of the traditional design, in which half of the participants receive treatment A and the other half receive treatment B, to four response-adaptive designs. They conclude that if the condition being treated is rare, then response-adaptive methods can perform substantially better and might be a more suitable alternative.

Despite the long history in clinical trials methodology, very few response-adaptive designs have actually occurred in practice and applications thus far have been disappointing (Rosenberger, 1999). This is largely attributable to the extracorporeal membrane oxygenation (ECMO) trial by Bartlett *et al.* (1985) which employed the randomised play-the-winner rule, a response-adaptive design described briefly in Section 3.2[1].

The problem of designing a clinical trial which aims to identify the superior treatment (exploration or learning) whilst treating the trial participants as effectively as possible (exploitation or earning) is a natural application area for bandit models, a type of response-adaptive design. Bandit models seek to balance the exploration versus exploitation trade-off in order to obtain an optimal allocation policy which maximises the expected number of patient successes over a finite number of patients.

[1]See also Section 2.1.2.

As such, they present an appealing alternative to the traditional approach used in clinical trials. Across the bandit literature, the use of bandit models to optimally design a clinical trial is often referred to as the primary motivation for their study (Gittins, 1979). However, to the best of our knowledge, they have never been implemented in real clinical practice for reasons including lack of randomisation and biased treatment effect estimates. Moreover, in contrast to the traditional approach taken in clinical trials, bandit models exhibit very low power since it is not possible to maximise both power and patient successes simultaneously. For a discussion of the benefits and challenges of bandit models in clinical trial practice, see Villar *et al.* (2015a).

In this chapter, we propose a novel bandit-based design which provides a very appealing compromise between these two conflicting objectives and addresses some of the key issues that have prevented bandit models from being implemented in clinical trial practice. We modify the optimal design, which aims to maximise the expected number of patient successes, in such a way that we overcome its limitations without having a significant negative impact on the patient benefit.

The modifications involve incorporating randomisation into a currently deterministic design, which was considered by Cheng and Berry (2007), and adding a constraint which forces a minimum number of patients on each treatment. These are described in Sections 3.2.2 and 3.2.3, respectively, building on the standard dynamic programming approach presented in Section 3.2.1. In Section 3.4, we compare our design to alternative designs via extensive simulations in several scenarios in the context of a recently published phase II clinical trial of isotonic fluid resuscitation in children with severe malnutrition and hypovolaemia (Akech *et al.*, 2010). We evaluate each design's performance according to the measures set out in Section 3.3. We summarise the main conclusions in Section 3.5 and highlight areas for future research.

3.2 Methods

In this section, we introduce different methods for allocating patients to treatments in a clinical trial. For simplicity of exposition, we consider a two-armed clinical trial with a binary endpoint and a finite number of patients within the trial, n. Patients enter the trial sequentially over time, one-by-one, and each patient is allocated to either treatment A or B on arrival. We assume that n is fixed but that the sample sizes for treatment groups A and B, denoted by N_A and N_B respectively, are random, where $N_A + N_B = n$. Let X and Y denote the patient's response (either a success or failure) from treatments A and B respectively, which we model as independent Bernoulli random variables. That is,

$$X \sim \text{Bernoulli}(1,\ \theta_A) \ \text{ and } \ Y \sim \text{Bernoulli}(1,\ \theta_B), \ \text{ for } \ 0 \leq \theta_A,\ \theta_B \leq 1,$$

where θ_A and θ_B are the unknown success probabilities of treatments A and B respectively. Further, assume that each patient's response from the allocated treatment becomes immediately available[2].

The *fixed randomised* design randomises patients to either treatment A or B with an equal, fixed probability, i.e. 50% in a two-armed trial. This will act as a reference to which each of the response-adaptive designs described below will be compared against.

One of the most well-known response-adaptive designs is the *randomised play-the-winner* (RPW) rule, a type of urn model, proposed by Wei and Durham (1978). This design is very intuitive and applies specifically to clinical trials comparing two treatments with binary responses. Initially, an urn contains u balls of type A and u balls of type B. When a patient is recruited, a ball is drawn randomly from the urn with replacement; if it is a type A ball, the patient receives treatment A and if

[2]Note that we relax this assumption in Chapters 4 and 5.

it is a type B ball, the patient receives treatment B. After each patient's outcome is observed, a decision about the urn composition is made depending on the observed result. Thus, a success on treatment A, or a failure on treatment B, generates an additional β type A balls and α type B balls in the urn. Similarly, a success on treatment B, or a failure on treatment A, will generate an additional β type B balls and α type A balls in the urn, where $0 \leq \alpha \leq \beta$ are integers. In this way, the urn accumulates more balls representing the superior treatment, thus increasing the probability that a patient receives the current best treatment. Note that the RPW rule is myopic (as are most response-adaptive designs) in the sense that it uses all of the past observations to treat the next patient as if this were the last patient in the trial.

3.2.1 Optimal Design using Dynamic Programming (DP)

The RPW rule described above is not constructed based on any formal optimality criterion so we now turn our attention to an alternative approach which utilises dynamic programming. With this approach, prior information on the unknown parameters is used in conjunction with the incoming data (and the number of remaining patients in the trial) to determine the optimal treatment allocation for every patient of the trial.

Note that we use t to denote both time and the last patient treated in this model since they are analogous, that is, at time t we have treated t patients. The trial time is therefore bounded by $0 \leq t \leq n$.

Since the treatment effects take values between zero and one, it is sensible to assign the parameters independent Beta prior distributions

$$\theta_A \sim \text{Beta}(s_{A,0},\ f_{A,0}) \ \text{ and } \ \theta_B \sim \text{Beta}(s_{B,0},\ f_{B,0}) \ \text{ for } \ 0 \leq \theta_A,\ \theta_B \leq 1,$$

where $s_{A,0}$ ($f_{A,0}$) and $s_{B,0}$ ($f_{B,0}$) represent the prior number of successes (failures)

on treatments A and B, respectively. Since this is a conjugate prior with respect to the Bernoulli likelihood function, the posterior distribution follows another Beta distribution with parameters summarising the relevant information from the trial to date (that is, the combination of the initial prior plus the accumulated data). At time $t \geq 1$, after observing $s_{A,t}$ ($f_{A,t}$) successes (failures) on treatment A, and $s_{B,t}$ ($f_{B,t}$) successes (failures) on treatment B, the posterior distribution is expressed by

$$\theta_A \mid s_{A,t}, f_{A,t} \sim \text{Beta}(s_{A,0}+s_{A,t},\ f_{A,0}+f_{A,t}) \ \text{ and } \ \theta_B \mid s_{B,t}, f_{B,t} \sim \text{Beta}(s_{B,0}+s_{B,t},\ f_{B,0}+f_{B,t}),$$

where $s_{A,t} + f_{A,t} + s_{B,t} + f_{B,t} = t$ for $t \geq 1$. Therefore, it will only be necessary to update the parameters of these distributions as the trial progresses. For simplicity, let the prior information and data combined be denoted as

$$\widetilde{s}_{A,t} = s_{A,0} + s_{A,t},\ \widetilde{f}_{A,t} = f_{A,0} + f_{A,t},\ \widetilde{s}_{B,t} = s_{B,0} + s_{B,t} \text{ and } \widetilde{f}_{B,t} = f_{B,0} + f_{B,t}. \quad (3.2.1)$$

Therefore, $\frac{\widetilde{s}_{j,t}}{\widetilde{s}_{j,t}+\widetilde{f}_{j,t}}$ is the posterior probability (i.e. the *current belief*) of success for treatment j given the prior information and data up to patient t.

Let $\delta_{j,t}$, for $t = 0, \ldots, n-1$, be the binary indicator variable representing whether patient $t+1$ is allocated to treatment $j \in \{A, B\}$, where

$$\delta_{j,t} = \begin{cases} 1, & \text{if patient } t+1 \text{ is allocated to treatment } j, \\ 0, & \text{otherwise.} \end{cases} \quad (3.2.2)$$

Using the jargon of dynamic programming, $\delta_{j,t}$ is the reward for every successfully treated patient, and thus $\frac{\widetilde{s}_{j,t}}{\widetilde{s}_{j,t}+\widetilde{f}_{j,t}} \cdot \delta_{j,t}$ is the expected (one-period) reward, where expectation is taken in the Bayesian sense, i.e. according to the current belief.

Let Π be the family of admissible designs (i.e. allocation policies) π, which are those such that $\sum_j \delta_{j,t} = 1$ for all t since only one treatment is allocated per patient. Let $\mathcal{F}_t(s_A, f_A, s_B, f_B)$ be the value function representing the maximum expected total

reward, i.e. the maximum Bayes-expected number of successes, in the rest of the trial after t patients have been treated when the combined information is (s_A, f_A, s_B, f_B), that is,

$$\mathcal{F}_t(s_A, f_A, s_B, f_B) := \max_{\pi \in \Pi} \mathbb{E}^{\pi} \left[\sum_{u=t}^{n-1} \sum_{j \in \{A,B\}} \frac{\widetilde{s}_{j,u}}{\widetilde{s}_{j,u} + \widetilde{f}_{j,u}} \cdot \delta_{j,u} \,\middle|\, \widetilde{s}_{A,t} = s_A, \widetilde{f}_{A,t} = f_A, \widetilde{s}_{B,t} = s_B, \widetilde{f}_{B,t} = f_B \right].$$

Note that this depends on the total number of patients n even though we do not state it explicitly to simplify the notation.

The ultimate optimisation problem is to find an optimal design which maximises the expected total reward, i.e. the Bayes-expected number of successes, over the set of all policies in the whole trial for a given prior at time $t = 0$, namely,

$$\mathcal{F}_0(s_{A,0}, f_{A,0}, s_{B,0}, f_{B,0}). \tag{3.2.3}$$

The problem summarised in equation (3.2.3) is known as a *finite-horizon Bayesian Bernoulli two-armed bandit problem* which can be solved exactly using dynamic programming methods, giving rise to an optimal adaptive treatment allocation sequence. Specifically, one can implement a backward induction algorithm which starts with the last patient, patient n, and proceeds iteratively towards the first patient. Details of this algorithm can be found in the Appendix 3.6.1[3].

Suppose that $t < n$. If treatment A is allocated to the next patient, then the expected total reward, i.e. the Bayes-expected number of successes, for patients $t + 1$

[3]See also Section 2.2.3(i).

to n under an optimal policy is

$$\mathcal{F}_t^A(s_A, f_A, s_B, f_B) = \frac{s_A}{s_A + f_A} \cdot [1 + \mathcal{F}_{t+1}(s_A + 1, f_A, s_B, f_B)] + \frac{f_A}{s_A + f_A} \cdot \mathcal{F}_{t+1}(s_A, f_A + 1, s_B, f_B).$$

Alternatively, if treatment B is allocated to the next patient, then the expected total reward, i.e. the Bayes-expected number of successes, for patients $t+1$ to n under an optimal policy is

$$\mathcal{F}_t^B(s_A, f_A, s_B, f_B) = \frac{s_B}{s_B + f_B} \cdot [1 + \mathcal{F}_{t+1}(s_A, f_A, s_B + 1, f_B)] + \frac{f_B}{s_B + f_B} \cdot \mathcal{F}_{t+1}(s_A, f_A, s_B, f_B + 1).$$

Therefore, the value function satisfies the following recurrence known as the *principle of optimality*,

$$\mathcal{F}_t(s_A, f_A, s_B, f_B) = \max\left\{\mathcal{F}_t^A(s_A, f_A, s_B, f_B),\ \mathcal{F}_t^B(s_A, f_A, s_B, f_B)\right\},\ \text{for } 0 \le t \le n-1,$$
$$\mathcal{F}_n(s_A, f_A, s_B, f_B) = 0,\ \text{otherwise}. \tag{3.2.4}$$

Unlike most response-adaptive designs, this is not a myopic allocation rule. Instead, all possible sequences of treatment allocations and responses are enumerated, and the sequence that maximises the expected number of patient successes over the finite planning horizon is selected (Hu and Rosenberger, 2006). As such, this approach is computationally intensive and suffers from the curse of dimensionality (Bellman, 1961). However, we provide an efficient algorithm for the optimal DP design, implemented in the programming language `R`; the computational times are shown in Table 3.6.1 of the Appendix 3.6.1.

The computational complexity of the dynamic programming methods to solve this problem is the main motivation behind the implementation of simpler index-based

solutions which circumvent the aforementioned problem of dimensionality. One such solution, which we include as a comparator, is the *Whittle index* (WI) proposed by Whittle (1988). This can be applied when the horizon is finite, which is the case with a clinical trial since there are a finite number of patients in the trial. It is derived from a relaxation of problem (3.2.3), allowing the multi-armed problem to be decomposed into single-armed problems in which the states are augmented, adding the number of patients remaining to be treated as an additional state. Although the WI is a heuristic solution, it has been found to be near-optimal in several cases. See Villar *et al.* (2015a) for a detailed review of the WI as a potential patient allocation rule in a clinical trial.

It is shown in Villar *et al.* (2015a) and Villar *et al.* (2015b), and further illustrated by our results, that optimal designs which achieve the highest patient benefit suffer from very low power. Moreover, optimal designs are completely deterministic (Cheng and Berry, 2007) which means there is a risk of introducing various sources of bias into the trial, e.g. selection bias (Blackwell and Hodges, 1957). Both of these factors contribute to making the optimal design unsuitable to implement in clinical trial practice. Therefore, in the rest of this section we focus on modifications to the DP design which address these shortcomings, i.e. its determinism and low power, while improving over a fixed randomised design in terms of patient benefit measures, such as overall response.

3.2.2 Optimal Design using Randomised Dynamic Programming (RDP)

Randomisation is a critical component in the design of clinical trials, not least to minimise the bias and confounding in order to achieve the desired accuracy and reliability (Chow and Liu, 2014). Therefore, a natural first step is to modify the optimal design by forcing actions to be randomised; see Cheng and Berry (2007). This is achieved by

assigning a probability to the allocation rule at each stage. In particular, we define the following actions so that each treatment has a probability of at least $1-p$ of being allocated to each patient, where $0.5 \leq p \leq 1$ for two-armed trials and will be referred to as the degree of randomisation. Note that $p = 0.5$ and $p = 1$ correspond to fixed, equal randomisation and the DP design, respectively.

(i) Action 1 ($a = 1$): The next patient receives treatment A with probability p and treatment B with probability $1 - p$.

(ii) Action 2 ($a = 2$): The next patient receives treatment B with probability p and treatment A with probability $1 - p$.

The associated expected total reward under this new action definition changes, along with the corresponding value function. Specifically, the expected total reward, i.e. the Bayes-expected number of successes, for patients $t+1$ to n when $a = 1$ is now given by

$$\mathcal{F}_t^1(s_A, f_A, s_B, f_B) = p \cdot \mathcal{F}_t^A(s_A, f_A, s_B, f_B) + (1-p) \cdot \mathcal{F}_t^B(s_A, f_A, s_B, f_B),$$

and analogously when $a = 2$,

$$\mathcal{F}_t^2(s_A, f_A, s_B, f_B) = (1-p) \cdot \mathcal{F}_t^A(s_A, f_A, s_B, f_B) + p \cdot \mathcal{F}_t^B(s_A, f_A, s_B, f_B).$$

Thus, in contrast to that shown in (3.2.4), the value function satisfies

$$\mathcal{F}_t(s_A, f_A, s_B, f_B) = \max\left\{\mathcal{F}_t^1(s_A, f_A, s_B, f_B),\ \mathcal{F}_t^2(s_A, f_A, s_B, f_B)\right\},\ \text{for } 0 \leq t \leq n-1,$$
$$\mathcal{F}_n(s_A, f_A, s_B, f_B) = 0,\ \text{otherwise.}$$

We refer to this design as the *randomised dynamic programming* (RDP) design hereafter.

Preferably, we would like p to be as close to one as possible so that the action that allocates to the superior treatment with probability p is as effective as possible. However, this would entail that sometimes, by chance, the inferior treatment is sampled too few times or not at all. The possibility of this undesirable event occurring makes this design unsuitable to implement in practice as it results in low power and estimates with large biases.

3.2.3 Optimal Design using Constrained Randomised Dynamic Programming (CRDP)

In order to circumvent having few or no observations on a treatment, we modify the optimal design further by adding a constraint to ensure that we always obtain at least ℓ observations from each treatment arm, where ℓ is a fixed predefined value and will be referred to as the degree of constraining. To do this, we add a penalty to the reward function for every combination of the states that give rise to fewer than ℓ observations on a treatment arm at the end of the trial.

We formulate this model as a Markov decision process with the following elements:

(i) Let $\boldsymbol{z}_t = (\widetilde{s}_{A,t},\ \widetilde{f}_{A,t},\ \widetilde{s}_{B,t},\ \widetilde{f}_{B,t},\ \widetilde{n})$ be the vector of states representing all the information that is needed in order to choose an action for patient t, where $\widetilde{s}_{A,t},\ \widetilde{f}_{A,t},\ \widetilde{s}_{B,t},\ \widetilde{f}_{B,t}$ are as defined previously in (3.2.1), and $\widetilde{n} = n - t$ is the number of patients in the trial remaining to be treated.

(ii) The action set, $\mathcal{A} = \{1, 2\}$, is composed of Action 1 ($a = 1$) and Action 2 ($a = 2$) as defined in Section 3.2.2.

(iii) The expected (one-period) reward under action a is given by $\mathcal{R}^a(\widetilde{s}_{A,t},\ \widetilde{f}_{A,t},\ \widetilde{s}_{B,t},\ \widetilde{f}_{B,t},\ \widetilde{n})$.

If we are not at the end of the trial ($\widetilde{n} \geq 1$), then

$$\mathcal{R}^a(\widetilde{s}_{A,t},\ \widetilde{f}_{A,t},\ \widetilde{s}_{B,t},\ \widetilde{f}_{B,t},\ \widetilde{n} \geq 1) = \begin{cases} p \cdot \frac{\widetilde{s}_{A,t}}{\widetilde{s}_{A,t}+\widetilde{f}_{A,t}} + (1-p) \cdot \frac{\widetilde{s}_{B,t}}{\widetilde{s}_{B,t}+\widetilde{f}_{B,t}}, & \text{if } a = 1, \\ (1-p) \cdot \frac{\widetilde{s}_{A,t}}{\widetilde{s}_{A,t}+\widetilde{f}_{A,t}} + p \cdot \frac{\widetilde{s}_{B,t}}{\widetilde{s}_{B,t}+\widetilde{f}_{B,t}}, & \text{if } a = 2. \end{cases}$$

Otherwise, if we are at the end of the trial with no more patients left to treat ($\widetilde{n} = 0$), then

$$\mathcal{R}(\widetilde{s}_{A,t},\ \widetilde{f}_{A,t},\ \widetilde{s}_{B,t},\ \widetilde{f}_{B,t},\ \widetilde{n} = 0) = \begin{cases} -n, & \text{if } s_{A,t} + f_{A,t} < \ell \text{ or } s_{B,t} + f_{B,t} < \ell, \\ 0, & \text{otherwise,} \end{cases}$$

where $-n$ is the penalty chosen because it is a large negative value which will cause the algorithm to avoid the undesirable states.

(iv) The non-zero transition probabilities, $\mathbb{P}(\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_t, a)$, representing the evolution of the states from patient t to $t+1$ under $a = 1$ and $a = 2$ are given as follows (where w.p. means "with probability").

When $a = 1$:

$$\boldsymbol{z}_{t+1} = \begin{cases} (\widetilde{s}_{A,t}+1,\ \widetilde{f}_{A,t},\ \widetilde{s}_{B,t},\ \widetilde{f}_{B,t},\ \widetilde{n}-1) & \text{w.p.} \quad p \cdot \frac{\widetilde{s}_{A,t}}{\widetilde{s}_{A,t}+\widetilde{f}_{A,t}}, \\ (\widetilde{s}_{A,t},\ \widetilde{f}_{A,t}+1,\ \widetilde{s}_{B,t},\ \widetilde{f}_{B,t},\ \widetilde{n}-1) & \text{w.p.} \quad p \cdot \frac{\widetilde{f}_{A,t}}{\widetilde{s}_{A,t}+\widetilde{f}_{A,t}}, \\ (\widetilde{s}_{A,t},\ \widetilde{f}_{A,t},\ \widetilde{s}_{B,t}+1,\ \widetilde{f}_{B,t},\ \widetilde{n}-1) & \text{w.p.} \quad (1-p) \cdot \frac{\widetilde{s}_{B,t}}{\widetilde{s}_{B,t}+\widetilde{f}_{B,t}}, \\ (\widetilde{s}_{A,t},\ \widetilde{f}_{A,t},\ \widetilde{s}_{B,t},\ \widetilde{f}_{B,t}+1,\ \widetilde{n}-1) & \text{w.p.} \quad (1-p) \cdot \frac{\widetilde{f}_{B,t}}{\widetilde{s}_{B,t}+\widetilde{f}_{B,t}}. \end{cases}$$

When $a = 2$:

$$\boldsymbol{z}_{t+1} = \begin{cases} (\widetilde{s}_{A,t}+1,\ \widetilde{f}_{A,t},\ \widetilde{s}_{B,t},\ \widetilde{f}_{B,t},\ \widetilde{n}-1) & \text{w.p.} \quad (1-p)\cdot\frac{\widetilde{s}_{A,t}}{\widetilde{s}_{A,t}+\widetilde{f}_{A,t}}, \\ (\widetilde{s}_{A,t},\ \widetilde{f}_{A,t}+1,\ \widetilde{s}_{B,t},\ \widetilde{f}_{B,t},\ \widetilde{n}-1) & \text{w.p.} \quad (1-p)\cdot\frac{\widetilde{f}_{A,t}}{\widetilde{s}_{A,t}+\widetilde{f}_{A,t}}, \\ (\widetilde{s}_{A,t},\ \widetilde{f}_{A,t},\ \widetilde{s}_{B,t}+1,\ \widetilde{f}_{B,t},\ \widetilde{n}-1) & \text{w.p.} \quad p\cdot\frac{\widetilde{s}_{B,t}}{\widetilde{s}_{B,t}+\widetilde{f}_{B,t}}, \\ (\widetilde{s}_{A,t},\ \widetilde{f}_{A,t},\ \widetilde{s}_{B,t},\ \widetilde{f}_{B,t}+1,\ \widetilde{n}-1) & \text{w.p.} \quad p\cdot\frac{\widetilde{f}_{B,t}}{\widetilde{s}_{B,t}+\widetilde{f}_{B,t}}. \end{cases}$$

We refer to our proposed design as the *constrained randomised dynamic programming* (CRDP) design hereafter.

3.3 Simulation Set-Up

We implement all of the above designs in several two-arm trial scenarios via simulations which will now be discussed, along with the performance measures that we use to compare and evaluate each design. The scenarios created are motivated by a recently published trial, as reported by Akech *et al.* (2010), which evaluated the effect of two different resuscitation treatments for children aged over six months with severe malnutrition and shock. The aim of the trial was to recruit 90 eligible patients, where 45 would be randomly assigned to group 0 (low dose hypotonic fluid: HSD/5D) and 45 to group 1 (Ringer's Lactate: RL). The original trial allocated patients between the two arms with a fixed and equal randomisation probability of 0.5. The primary response outcomes were binary and available at eight and 24 hours after randomisation (resolution of shock by 8/24 hours). For this trial, 61 children were recruited, 26 received arm 0 and 29 received arm 1. At the end of the trial, the success rates observed in groups 0 and 1 at eight hours were 32% and 44%, respectively, and at 24 hours were 22% and 44%, respectively. Although these differences were not statistically significant, the relatively quickly observed primary endpoint, the life-threatening nature of the disease, and the fact that patient recruitment is challenging, makes this

trial an ideal motivating scenario for testing our proposed design.

Assuming that we begin the trial (at $t = 0$) in a state of equipoise, that is, a state of genuine uncertainty about which treatment is superior, we let $s_{A,0} = f_{A,0} = s_{B,0} = f_{B,0} = 1$, reducing this to a uniform prior.

We consider the following hypothesis

$$H_0 : \theta_A = \theta_B \text{ versus } H_1 : \theta_A \neq \theta_B,$$

which will be tested using Fisher's exact test (Routledge, 2005) for comparing the success probabilities of two binomial distributions. Fisher's exact test is probably the most common choice for binary outcomes and a small sample size. This test is a conditional test (conditioning on the marginals), which increases the discreteness and thus the conservatism of the test (Kateri, 2014). This means that the observed rejection rate is often far below the nominal significance level. Therefore, we set the nominal significance level to 0.1 throughout so that the observed type I error value will be closer to 0.05.

Alternatively, we could have followed a Bayesian inference procedure. However, in a clinical trial context a traditional hypothesis test is expected (due to both this being a common practice and because of regulatory requirements). Also, since all the simulations included in this chapter use an uninformative prior, the impact of using a Bayesian estimator instead of the sample proportion for point estimation and decision making would be negligible.

In order to create a comprehensive picture of our proposed design, we run our simulations for a range of combinations of the success probability parameters θ_A and θ_B. Specifically, we consider $\theta_A = 0.2$ against $\theta_B = (0.1,\ 0.2,\ \ldots,\ 0.9)$, and similarly for $\theta_A = 0.5$ and 0.8. In the following, we focus on the scenario where θ_A is fixed at 0.5 for all $\theta_B \in (0.1,\ 0.9)$ since the patterns observed for the other cases are very similar.

Furthermore, we repeat the simulations for different total sample sizes. The results for $n = 75$ are reported throughout because this shows a good range of power values across all scenarios and clearly highlights the differences between each design, thus enabling us to make better comparisons. The results for $n = 25$, 50 and 100 are shown in Figures 3.6.3–3.6.5 of the Appendix 3.6.7.

We evaluate the performance of these designs by simulating 10,000 replications of each trial and taking the average values over these runs.

3.3.1 Performance Measures

In addition to the operating characteristics, such as the power and type I error rate, we also consider the ethical performance of each design since this is one of the major advantages of response-adaptive designs over traditional fixed designs. Specifically, the criteria we focus on to assess the performance of each design are:

1. **Power**. The proportion of times we *correctly* reject H_0 in the 10,000 trial replicates, i.e. the probability of making the correct decision at the end of the trial, so we want this to be high. This provides an informative measure of how well a test performs. This is calculated when $\theta_A \neq \theta_B$.

2. **Type I error rate**. The proportion of times we *incorrectly* reject H_0, i.e. the probability of making the incorrect decision at the end of the trial, so we want this to be low. This is calculated when $\theta_A = \theta_B$.

3. **Percentage of patients allocated to the superior treatment arm**. This measures the ethical performance of each design, which we wish to maximise.

4. **Average bias of the estimator**. This provides a measure of the bias exhibited by the treatment effect estimator, where we define treatment effect as the treatment difference, $\hat{\Delta} = \hat{\theta}_A - \hat{\theta}_B$. The estimator of θ_A and θ_B is simply the sample proportion $\hat{\theta}_A = s_{A,n}/N_A$ and $\hat{\theta}_B = s_{B,n}/N_B$, respectively. This is

the observed proportion of successes in either treatment group by the end of the trial (at time $t = n$). The average bias of this estimator is defined to be the difference between the estimated success probability difference and the true success probability difference, that is,

$$\text{Bias}(\hat{\Delta}) = \mathbb{E}(\hat{\Delta} - \Delta) = \mathbb{E}(\hat{\theta}_A - \hat{\theta}_B) - (\theta_A - \theta_B)\,. \tag{3.3.1}$$

5. **Mean squared error (MSE) of the estimator**. The MSE provides a measure of the quality and variability of the estimator, $\hat{\Delta}$, and is defined by

$$\text{MSE}(\hat{\Delta}) = \mathbb{E}\left[(\hat{\Delta} - \Delta)^2\right],$$

which can be expressed in terms of the bias and variance of the estimator as,

$$\text{MSE}(\hat{\Delta}) = \text{Bias}(\hat{\Delta})^2 + \text{Var}(\hat{\Delta}). \tag{3.3.2}$$

3.4 Simulation Results and Design Comparison

We compare our proposed design to the alternative designs outlined in Section 3.2 based upon the performance measures highlighted in Section 3.3.1. We set $p = 0.9$ as the degree of randomisation and $\ell = 0.15n$ as the degree of constraining in our proposed CRDP design, which we believe yields robust design characteristics for many scenarios of interest and could be used as a quick rule of thumb. Alternatively, ℓ could be heuristically determined as the minimum sample size per arm required to attain a power of $(1 - \gamma)$ in a fixed randomised design, where $(1 - \gamma) \leq (1 - \beta)$ and $(1 - \beta)$ is the power level obtained by a fixed randomised trial of size n. In the following two paragraphs, we describe a more formal heuristic approach to determine p and ℓ when higher precision is needed to trade-off power and patient benefit.

We tried a range of values for $\ell \in (0.05n, 0.50n)$ (where $0.50n$ corresponds to fixed equal randomisation) and found that as ℓ increases, the power of the design increases hyperbolically, while the percentage of patients allocated to the superior treatment decreases linearly. This is illustrated in Figure 3.6.1 of the Appendix 3.6.3. We recommend choosing $\ell \in (0.10n, 0.15n)$ because for values of $\ell < 0.10n$, the power is insufficient, and for values of $\ell > 0.15n$, the very small gains in power do not outweigh the considerable reduction in the percentage of patients allocated to the superior treatment.

Similarly, we tried a range of values for $p \in (0.5, 1)$ (where $p = 0.5$ and $p = 1$ correspond to fixed equal randomisation and the DP design, respectively) and observed that there is a decrease in power, but a large increase in the percentage of patients allocated to the superior treatment as p increases from 0.5 to 0.9; see Tables 3.6.2–3.6.5 in the Appendix 3.6.4 which illustrate this for the RDP design (i.e. without the constraint). We take $p = 0.9$ since this produces a good balance between the power and patient benefit across a wide range of scenarios and sample sizes.

3.4.1 Power and Type I Error

Figure 3.4.1 illustrates the changes in statistical power, and type I error rate, for each design across a range of scenarios in a study with 75 observations (where the result for $\theta_A = \theta_B$ corresponds to the type I error rate). It can be seen that fixed randomisation attains the highest power for all scenarios, whereas that of the DP and WI designs is drastically reduced, even for large treatment differences. This is what we would expect since it is not possible to maximise both power and patient successes simultaneously and, unlike the fixed design, the DP design aims to maximise the expected number of successes within the trial. Therefore, although the DP and WI designs are able to identify the superior treatment arm, they are unable to do so with sufficient statistical significance. We can see that the power of these designs lies

below 0.3 for all $\theta_B \in (0.1, 0.9)$, confirming that they are severely underpowered. As a result, they are clearly unsuitable to implement in practice.

Figure 3.4.1 also shows that once randomisation is incorporated into the DP design to form RDP, there is a substantial improvement in power compared to the DP and WI designs, which even exceeds the 0.8 level (illustrated by the upper dashed line) for some scenarios. Our proposed CRDP achieves even better power, with its power values lying much closer to those for the fixed design than the other bandit designs.

The obvious patterns, such as the power increasing with the size of the treatment difference for each design, are apparent in Figure 3.4.1. Furthermore, additional evaluations for other sample sizes show similar patterns and can be seen in Figure 3.6.3 of the Appendix 3.6.7.

Turning our attention to the type I error rates, we see that the type I error rate of both the DP and WI designs lies markedly below the nominal significance level at 0.1 (illustrated by the lower dashed line on Figure 3.4.1) and is therefore greatly deflated for both designs. However, all of the other designs attain similar, higher observed type I error rates which are much closer to the nominal significance level and thus have better controlled type I error rates.

3.4.2 Patient Benefit

Figure 3.4.2 shows the percentage of patients (out of a total of 75) that receive the superior treatment within the trial. Note that when $\theta_A = \theta_B$, we define treatment A as the superior treatment for illustrative purposes and all designs show that approximately 50% of patients receive the superior treatment in this case, as expected.

The DP and WI designs perform the best, resulting in the highest percentage of patients receiving the superior treatment. This is not at all surprising considering they are designed to maximise the expected total reward (patient successes) within the trial in order to satisfy the patient benefit criterion.

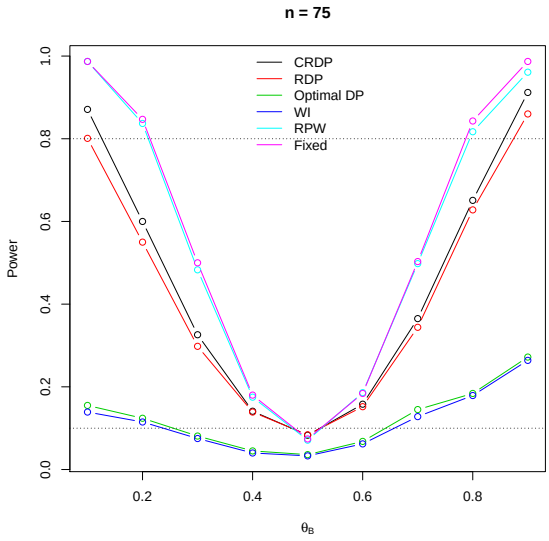


Figure 3.4.1: The changes in power and type I error for each design when $n = 75$, $\theta_A = 0.5$ and $\theta_B \in (0.1, 0.9)$. The upper dashed line at 0.8 represents the desired power level, and the lower dashed line at 0.1 represents the nominal significance level.

At the other extreme, by design, the fixed randomised design allocates only 50% of the patients to the superior treatment in every scenario. Although the RPW rule does outperform the fixed design in terms of the patient benefit, the percentage of patients that are on the superior treatment is still much lower compared to all of the other designs. It is useful to note that the limiting allocation proportion of patients on treatment A for the RPW rule is given by $(1-\theta_B)/(2-\theta_A-\theta_B)$ (Wei and Durham, 1978).

Figure 3.4.2 shows that the RDP and CRDP designs perform very well and the percentage of patients receiving the superior treatment is still sufficiently high, with the CRDP line lying slightly below the RDP line due to the addition of the constraint. The largest difference between CRDP and DP is approximately 10%, which occurs at either end of the plot when the size of the treatment difference is at its largest. Moreover, our proposed CRDP design allocates a maximum of approximately 21% and 35% more patients to the superior treatment than the RPW rule and fixed design, respectively, which occurs when $\theta_B = 0.1$.

For all designs (excluding the fixed), Figure 3.4.2 shows that the percentage of patients allocated to the superior treatment increases with the magnitude of the treatment difference, with the higher values occurring at the tails of the graph which correspond to the larger treatment differences. Furthermore, similar patterns are observed for other sample sizes; see Figure 3.6.4 in Appendix 3.6.7.

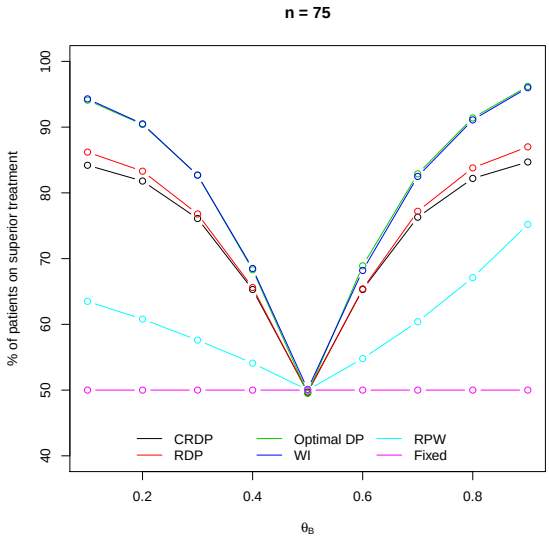


Figure 3.4.2: The percentage of patients on the superior treatment arm for each design when $n = 75$, $\theta_A = 0.5$ and $\theta_B \in (0.1, 0.9)$.

3.4.3 Bias

Figure 3.4.3 shows the average bias of the difference in the sample proportions as an estimator for the treatment effect, as defined by (3.3.1), in a study with 75 observations. We see that the fixed randomised design produces the best result in terms of the bias, with its associated estimator attaining zero bias for all scenarios, as it should.

At the other extreme, the DP and WI designs exhibit the largest statistical bias with a maximum absolute value of 0.2 occurring when $\theta_B = 0.9$. Therefore, the corresponding estimates following such bandit designs will be biased due to the underlying dependence structure induced in the resulting observations. This is reflected

in Table 3.4.1 which directly reports the raw estimates of the success probabilities, $\hat{\theta}_A$ and $\hat{\theta}_B$. Table 3.4.1 shows that in the DP design, the estimate of the success probability for the inferior arm is substantially underestimated. The estimate for the superior arm is also underestimated, but less than for the inferior arm, particularly when the treatment difference is relatively small. This implies that the estimate of the treatment difference, $\hat{\Delta}$, is generally overestimated. Since bandit designs allocate fewer patients to the inferior treatment, this may partially explain why the estimate corresponding to this arm is worse than that of the superior arm because there are fewer observations to base the inference on.

Once randomisation is incorporated into the DP design, we see from Figure 3.4.3 that the bias is drastically reduced across all scenarios, with a maximum absolute value of 0.027 which is 85% smaller than the worst-case bias of the other bandit designs. Moreover, our proposed CRDP design performs even better than the RDP and further reduces the bias of the treatment effect estimator. In fact, the bias values for our proposed CRDP are very close to zero for all scenarios with a maximum bias value of only 0.014 which is 93% smaller than the worst-case bias for the DP design. As such, the bias following our proposed CRDP is negligible compared to the very large bias exhibited by the other bandit designs and hence, the treatment effect estimator following our proposed CRDP design is essentially mean-unbiased. Again, this is reflected in Table 3.4.1 which shows that in our proposed CRDP design, $\hat{\theta}_A$ and $\hat{\theta}_B$ are now much closer to their true values. Moreover, there is a large improvement in the estimate of the success probability for the inferior arm compared to the DP design since it is now only slightly underestimated.

Note that we can clearly see from Figure 3.4.3 that all designs correctly attain a bias of zero when $\theta_A = \theta_B$. Similar results for different n are provided in Figure 3.6.5 of Appendix 3.6.7.

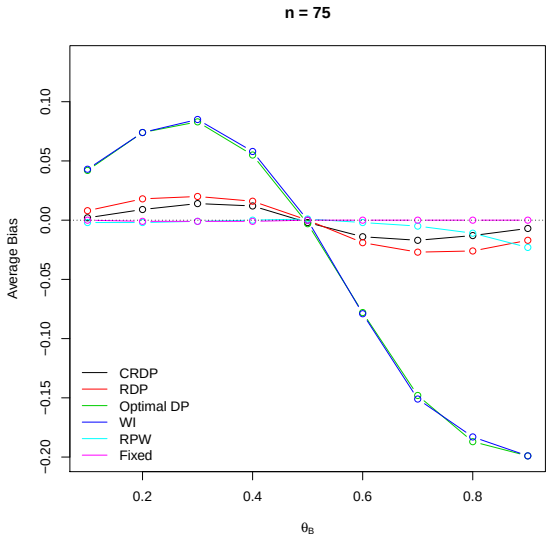


Figure 3.4.3: The average bias of the treatment effect estimator when $n = 75$, $\theta_A = 0.5$ and $\theta_B \in (0.1, 0.9)$.

True		Fixed		DP		CRDP	
θ_A	θ_B	$\hat{\theta}_A$ (s.e.)	$\hat{\theta}_B$ (s.e.)	$\hat{\theta}_A$ (s.e.)	$\hat{\theta}_B$ (s.e.)	$\hat{\theta}_A$ (s.e.)	$\hat{\theta}_B$ (s.e.)
0.500	0.100	0.500 (0.083)	0.100 (0.050)	0.498 (0.062)	0.057 (0.096)	0.499 (0.064)	0.097 (0.085)
0.500	0.200	0.500 (0.083)	0.201 (0.065)	0.493 (0.080)	0.119 (0.132)	0.496 (0.070)	0.187 (0.105)
0.500	0.300	0.500 (0.083)	0.301 (0.075)	0.474 (0.118)	0.191 (0.156)	0.489 (0.084)	0.275 (0.109)
0.500	0.400	0.500 (0.083)	0.401 (0.080)	0.434 (0.162)	0.279 (0.176)	0.475 (0.098)	0.364 (0.107)
0.500	0.500	0.500 (0.083)	0.500 (0.082)	0.386 (0.192)	0.389 (0.192)	0.462 (0.105)	0.464 (0.106)
0.500	0.600	0.500 (0.083)	0.600 (0.080)	0.340 (0.216)	0.518 (0.193)	0.461 (0.111)	0.575 (0.099)
0.500	0.700	0.500 (0.083)	0.699 (0.075)	0.303 (0.240)	0.652 (0.172)	0.472 (0.123)	0.689 (0.080)
0.500	0.800	0.500 (0.083)	0.800 (0.065)	0.290 (0.266)	0.780 (0.129)	0.484 (0.136)	0.797 (0.058)
0.500	0.900	0.500 (0.083)	0.900 (0.049)	0.291 (0.290)	0.895 (0.074)	0.493 (0.147)	0.900 (0.039)

Table 3.4.1: The estimates of success probabilities, $\hat{\theta}_A$ and $\hat{\theta}_B$, and corresponding standard errors (s.e.) for the success probabilities of treatments A and B, respectively, compared to their true values θ_A and θ_B. These results correspond to the scenario in which $n = 75$, $\theta_A = 0.5$ and $\theta_B \in (0.1, 0.9)$.

3.4.4 Mean Squared Error

Figure 3.4.4 shows the mean squared error (MSE) of the treatment effect estimator, as defined by (3.3.2), for a study with 75 observations. The fixed randomised design

results in the smallest MSE, with values fairly constant and close to zero for all scenarios.

The DP and WI designs exhibit the largest MSE values, with the MSE of the WI design exceeding those of the DP design for all scenarios. This is a direct consequence of the large bias observed in Figure 3.4.3. Moreover, these designs experience the largest increase in MSE as θ_B increases from 0.1 to 0.7, after which point they remain fairly constant. Specifically, as θ_B increases from 0.1 to 0.7, the MSE jumps from 0.016 to 0.141 for the WI design, and from 0.015 to 0.133 for the DP design. We also notice from Figure 3.4.4 that the associated MSE plots for the DP and WI designs are not symmetric about $\theta_B = 0.5$. This is a result of the variance of the estimator increasing markedly as θ_B increases from 0.1 to 0.6, in addition to the bias for the DP and WI being much larger for larger values of θ_B.

Once randomisation is incorporated into the DP, the MSE is reduced for all scenarios, from a worst-case value of 0.141 in the WI design to a worst-case value of 0.032 in the RDP design which is a 77.3% improvement. Moreover, our proposed CRDP design improves the MSE values even further, with a lower and upper bound of 0.011 and 0.026, respectively. The majority of the MSE values lie around 0.030 for the RDP design and 0.020 for our proposed CRDP design. In contrast to the steep curves of the DP and WI designs, the MSE values associated with the RDP and CRDP designs remain fairly constant (as with the fixed and RPW designs), thus giving rise to the relatively flat curves visible in Figure 3.4.4. Furthermore, we see that the curve corresponding to our proposed CRDP lies fairly close to the curve for the fixed design. Thus, the MSE values of the treatment effect estimator following our proposed CRDP design are comparable to those of the fixed design, staying close to zero for all scenarios, and are a huge improvement on those exhibited by the DP and WI bandit designs.

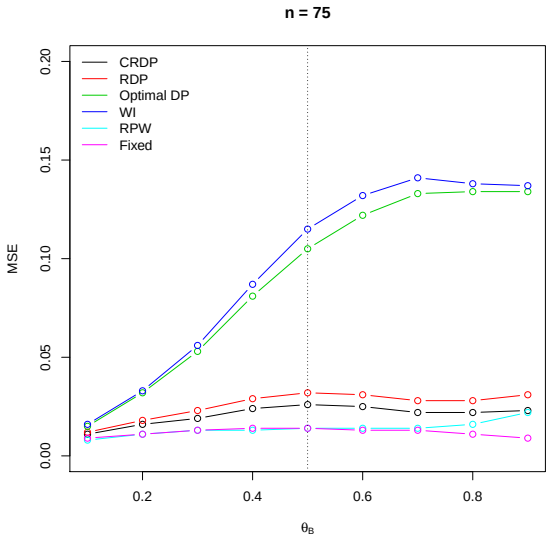


Figure 3.4.4: The mean squared error (MSE) of the treatment effect estimator when $n = 75$, $\theta_A = 0.5$ and $\theta_B \in (0.1, 0.9)$.

3.4.5 Overall Performance

Figure 3.4.5 shows a star plot for each design against power, patient benefit, average bias and MSE of the treatment effect estimator in a trial with 75 patients when $\theta_A = 0.5$ and $\theta_B = 0.2$. The most desirable values lie towards the outer edge of the star plot with the least favourable values towards the centre. Figure 3.4.5 summarises the key features of each design showing that the fixed design performs very well with respect to power, average bias and MSE but poorly with respect to patient benefit, whilst in contrast the DP design performs poorly with respect to power, average bias and MSE but very well with respect to patient benefit. Our proposed CRDP design, on the other hand, has values lying near to the outer edge of the star plot for power, average bias, MSE and patient benefit, thus showing that it performs well with respect to all of the performance measures. Table 3.6.6 in Appendix 3.6.6 reports additional combined measures that complement Figure 3.4.5 to compare the designs.

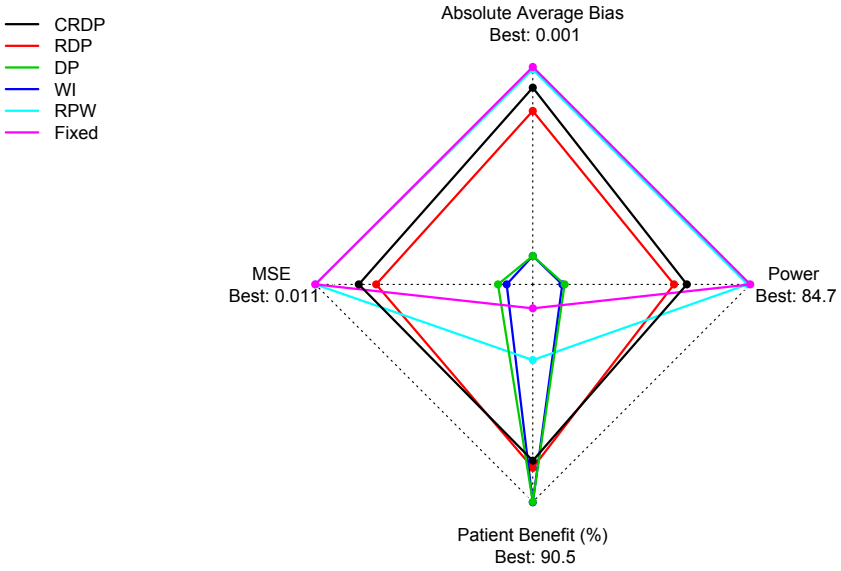


Figure 3.4.5: Star plot showing the performance of each design with respect to power, patient benefit, absolute average bias and MSE of the treatment effect estimator when $n = 75$, $\theta_A = 0.5$ and $\theta_B = 0.2$. The best value achieved for each performance measure is depicted at the outer edge. (Note that the absolute average bias and MSE axes have been inverted so that the smaller (favourable) values are towards the outer edge, unlike the power and patient benefit axes which have their larger values towards the outer edge.)

3.4.6 CRDP Patient Allocation

Figure 3.4.6 shows the average allocation probability to the superior treatment B under the CRDP design for every patient t in a trial with 75 patients when $\theta_A = 0.5$ and $\theta_B = 0.7$. This figure illustrates how the CRDP design adaptively allocates patients between the two treatments over time. The average allocation probability to a superior arm grows steadily through the trial towards the degree of randomisation selected ($p = 0.9$), but without reaching it in this scenario. As the trial approaches the treatment decisions for its final 15 patients, this probability markedly oscillates in order to satisfy the degree of constraining. This indicates that an important number of allocations to the inferior arm under the CRDP design tend to occur by the end of the trial rather than at the beginning of it[4]. Figure 3.4.7 also illustrates this point

[4]Note that this could circumvent the problem of accrual bias mentioned in Section 2.1.1.

by plotting the observed patient allocations during five different trial realisations.

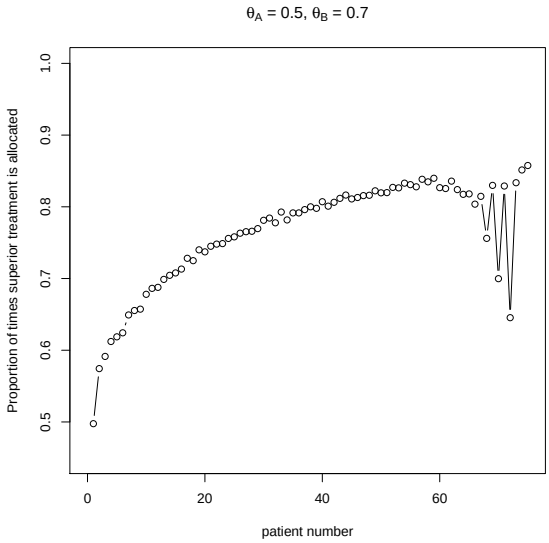


Figure 3.4.6: Probability of allocating a patient to the superior treatment B for CRDP when $\theta_A = 0.5$ and $\theta_B = 0.7$ in a trial of size $n = 75$.

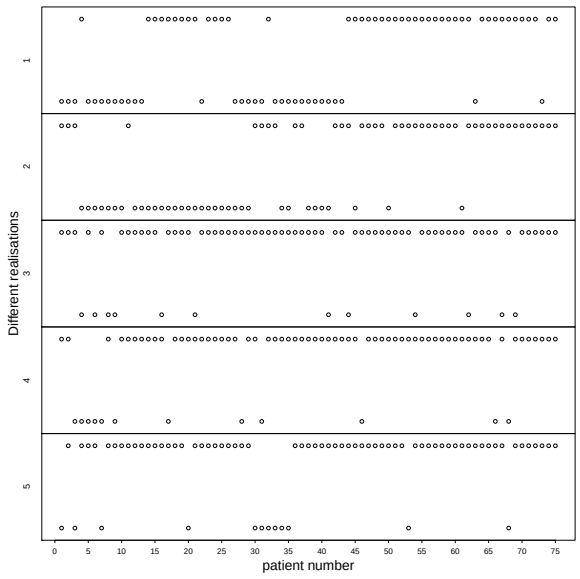


Figure 3.4.7: Patient allocations for CRDP when $\theta_A = 0.5$ and $\theta_B = 0.7$ in a trial of size $n = 75$ for five different trial realisations. Upper dots represent allocations to the superior treatment B while lower dots represent allocations to the inferior treatment A.

3.5 Discussion

In this chapter, we evaluate different methods for allocating patients to treatments. The DP design performs very well when considering patient benefit compared to traditional fixed randomisation. However, this method suffers from an extremely low power to detect a significant treatment difference, biased estimates of the treatment effect and a large MSE. Moreover, it is completely deterministic and thus at risk of many possible sources of bias.

At the other extreme, fixed randomisation performs very well in terms of the statistical criteria, exhibiting high power, unbiased estimates of the treatment effect and small MSE. However, it allocates a large proportion of patients to the inferior treatment arm. This is particularly detrimental for rare, and fatal, diseases in which a substantial proportion of patients exhibiting the disease may be included in the trial and therefore the priority should be to treat these patients as effectively as possible.

We propose modifications to the DP design which overcome its current limitations and offer patient benefit advantages over a fixed randomised design by randomising in an optimal way and forcing a minimum number of patients on each arm. Our formal, mathematical approach grounded in decision theory creates a continuum of designs, with DP and fixed randomisation at the extremes, which offers freedom in choosing the most appropriate balance by fixing a degree of randomisation and a degree of constraining. This greatly increases the prospects of a bandit-based design being implemented in real clinical trial practice, particularly for trials involving rare diseases and small populations where the fixed randomisation approach is no longer the most appropriate design to use and is often not feasible due to the small sample sizes involved.

Our proposed CRDP design, with suggested degree of randomisation $p = 0.9$ and degree of constraining $\ell = 0.15n$, seems to perform robustly in a range of simulated scenarios (not all of which are reported here). The power is only slightly lower than

with fixed randomisation, while almost as many patients are randomised to the superior treatment as in the DP design. Hence, this design strikes a very good balance in terms of the patient benefit and power trade-off, providing both power and ethical advantages, which acknowledges that clinical trials are multiple objective experiments.

The average bias and MSE of the treatment effect estimator following our proposed CRDP design are very low. It is well known that selection results in biased estimators (see e.g. Bauer *et al.*, 2010). This is also true for group-sequential trials which are, however, routinely used in practice nowadays because the benefit from these designs can outweigh the bias incurred, particularly in the case of rare diseases. In order to make this assessment, it is important to determine the magnitude of the bias (as well as the benefits of the design) and hence the evaluations provided are essential for these novel methods to be applied in a real-life trial. In cases where the magnitude of the bias could be considered excessive, there exists a bias-corrected estimator that can be used (which comes at the price of a notably increased variability); see Bowden and Trippa (2017). Coad and Ivanova (2001) also study the bias of the maximum likelihood estimators of the success probabilities following several response-adaptive designs in the two-arm, binary response case.

In this chapter, we consider a two-armed trial with binary endpoints for simplicity, yet the principles used easily extend to multi-arm trials. An area of further work is to generalise the proposed design so it can be applied to other endpoints. In addition, a natural extension of this work is to modify the heuristic WI policy in a similar way as we have with the optimal DP design since index policies are conceptually more intuitive (we allocate the patient to the treatment with the highest index), and hence easier to communicate and be understood by clinicians. Moreover, the WI is potentially very important for the extension to more than two treatment arms since the DP quickly becomes computationally intractable while the WI is still feasible (Villar *et al.*, 2015a).

In our proposed design, each patient's response is used to inform the subsequent allocation decision. This relies on the assumption that patient responses become available before the next patient receives treatment (which would be the case if patient responses were quickly observed, for example). In many clinical trial settings, this is unrealistic because often a treatment takes a substantial length of time to induce a response and so it is very likely that the accrual rate will exceed the response rate. However, in a rare disease setting, the accrual rate is likely to be relatively slow with some patients being recruited over several years, and hence this assumption would be reasonable. Further research is required to address the problem of incorporating delayed responses into bandit-based designs which would increase the generalisability of our proposed design[5].

Moreover, our proposed design can only be applied to relatively small-scale trials since the underlying backwards induction algorithm suffers from the curse of dimensionality (Bellman, 1961) and currently attains its practical limit at $n = 200$. Again, this is not an issue for a rare disease setting in which the number of patients available for participation in the trial is limited, or clinical trials involving children, for example, in which recruitment is challenging (Hampson *et al.*, 2014). In fact, many phase II trials have no more than 200 patients, even in common diseases.

Additional extensions of this work include considering the effect of changing the prior distribution assigned to the unknown success probabilities. For example, a Beta prior with carefully chosen parameters could alternatively be used if the investigator wishes to reflect a greater amount of knowledge or a bias in favour of a particular treatment, without increasing the complexity of the problem. See Hampson *et al.* (2014) in which the unknown model parameters of the prior distribution are determined by eliciting expert opinion and incorporating historical data from a related trial.

[5]See Chapters 4 and 5.

3.6 Appendix

3.6.1 Backward Induction Algorithm

- If $t = n$, there is nothing to do because all n patients have already been treated and their outcomes observed. Thus, $\mathcal{F}_n(\widetilde{s}_{A,n}, \widetilde{f}_{A,n}, \widetilde{s}_{B,n}, \widetilde{f}_{B,n}) = 0 \ \forall$ $\widetilde{s}_{A,n},\ \widetilde{f}_{A,n},\ \widetilde{s}_{B,n},\ \widetilde{f}_{B,n}$.

- If $t = n-1$, there is only one patient left to treat and interest is in determining which treatment to allocate to this patient $\forall\ \widetilde{s}_{A,n-1},\ \widetilde{f}_{A,n-1},\ \widetilde{s}_{B,n-1},\ \widetilde{f}_{B,n-1}$ that sum to $n-1$. There are two possibilities:

  - If treatment A is allocated to the remaining patient, then we compute the expectation

    $$\mathcal{F}^A_{n-1}(\widetilde{s}_{A,n-1}, \widetilde{f}_{A,n-1}, \widetilde{s}_{B,n-1}, \widetilde{f}_{B,n-1}) = \frac{\widetilde{s}_{A,n-1}}{\widetilde{s}_{A,n-1} + \widetilde{f}_{A,n-1}} \cdot 1 + \frac{\widetilde{f}_{A,n-1}}{\widetilde{s}_{A,n-1} + \widetilde{f}_{A,n-1}} \cdot 0,$$

    where $\frac{\widetilde{s}_{A,n-1}}{\widetilde{s}_{A,n-1}+\widetilde{f}_{A,n-1}}$ is the expectation of θ_A with respect to a Beta$(\widetilde{s}_{A,n-1}, \widetilde{f}_{A,n-1})$ distribution, and $\frac{\widetilde{f}_{A,n-1}}{\widetilde{s}_{A,n-1}+\widetilde{f}_{A,n-1}}$ is the probability of a failure if treatment A is allocated.

  - Alternatively, if treatment B is allocated to the remaining patient, then we compute the expectation

    $$\mathcal{F}^B_{n-1}(\widetilde{s}_{A,n-1}, \widetilde{f}_{A,n-1}, \widetilde{s}_{B,n-1}, \widetilde{f}_{B,n-1}) = \frac{\widetilde{s}_{B,n-1}}{\widetilde{s}_{B,n-1} + \widetilde{f}_{B,n-1}} \cdot 1 + \frac{\widetilde{f}_{B,n-1}}{\widetilde{s}_{B,n-1} + \widetilde{f}_{B,n-1}} \cdot 0,$$

    where $\frac{\widetilde{s}_{B,n-1}}{\widetilde{s}_{B,n-1}+\widetilde{f}_{B,n-1}}$ is the expectation of θ_B with respect to a Beta$(\widetilde{s}_{B,n-1}, \widetilde{f}_{B,n-1})$ distribution, and $\frac{\widetilde{f}_{B,n-1}}{\widetilde{s}_{B,n-1}+\widetilde{f}_{B,n-1}}$ is the probability of a failure if treatment B is allocated.

Interest is in choosing the optimal allocation such that

$$\mathcal{F}_{n-1}(\widetilde{s}_{A,n-1}, \widetilde{f}_{A,n-1}, \widetilde{s}_{B,n-1}, \widetilde{f}_{B,n-1}) =$$
$$\max\{\mathcal{F}^A_{n-1}(\widetilde{s}_{A,n-1}, \widetilde{f}_{A,n-1}, \widetilde{s}_{B,n-1}, \widetilde{f}_{B,n-1}),\ \mathcal{F}^B_{n-1}(\widetilde{s}_{A,n-1}, \widetilde{f}_{A,n-1}, \widetilde{s}_{B,n-1}, \widetilde{f}_{B,n-1})\}.$$

Thus, if $\mathcal{F}^A_{n-1}(\widetilde{s}_{A,n-1}, \widetilde{f}_{A,n-1}, \widetilde{s}_{B,n-1}, \widetilde{f}_{B,n-1}) > \mathcal{F}^B_{n-1}(\widetilde{s}_{A,n-1}, \widetilde{f}_{A,n-1}, \widetilde{s}_{B,n-1}, \widetilde{f}_{B,n-1})$, then it is optimal to allocate the remaining patient to treatment A, and vice versa. If they are equal, then both treatments are optimal choices.

- The next step is if $t = n - 2$, i.e. when there are two remaining patients to be allocated. To determine which treatment to allocate to patient $n - 1$, there are two possibilities:

  - If treatment A is allocated to patient $n - 1$, then we compute the expectation

$$\mathcal{F}^A_{n-2}(\widetilde{s}_{A,n-2}, \widetilde{f}_{A,n-2}, \widetilde{s}_{B,n-2}, \widetilde{f}_{B,n-2}) =$$
$$\frac{\widetilde{s}_{A,n-2}}{\widetilde{s}_{A,n-2} + \widetilde{f}_{A,n-2}} \cdot \left(1 + \mathcal{F}_{n-1}(\widetilde{s}_{A,n-2} + 1, \widetilde{f}_{A,n-2}, \widetilde{s}_{B,n-2}, \widetilde{f}_{B,n-2})\right) +$$
$$\frac{\widetilde{f}_{A,n-2}}{\widetilde{s}_{A,n-2} + \widetilde{f}_{A,n-2}} \cdot \left(0 + \mathcal{F}_{n-1}(\widetilde{s}_{A,n-2}, \widetilde{f}_{A,n-2} + 1, \widetilde{s}_{B,n-2}, \widetilde{f}_{B,n-2})\right).$$

  - Similarly, if treatment B is allocated, then we compute the expectation

$$\mathcal{F}^B_{n-2}(\widetilde{s}_{A,n-2}, \widetilde{f}_{A,n-2}, \widetilde{s}_{B,n-2}, \widetilde{f}_{B,n-2}) =$$
$$\frac{\widetilde{s}_{B,n-2}}{\widetilde{s}_{B,n-2} + \widetilde{f}_{B,n-2}} \cdot \left(1 + \mathcal{F}_{n-1}(\widetilde{s}_{A,n-2}, \widetilde{f}_{A,n-2}, \widetilde{s}_{B,n-2} + 1, \widetilde{f}_{B,n-2})\right) +$$
$$\frac{\widetilde{f}_{B,n-2}}{\widetilde{s}_{B,n-2} + \widetilde{f}_{B,n-2}} \cdot \left(0 + \mathcal{F}_{n-1}(\widetilde{s}_{A,n-2}, \widetilde{f}_{A,n-2}, \widetilde{s}_{B,n-2}, \widetilde{f}_{B,n-2} + 1)\right).$$

- *et cetera.*

These steps are just iterations, and can be expressed more succinctly in the general

form as follows.

If treatment A is allocated to the next patient, then the expected number of successes for patients $t+1$ through n under an optimal policy is

$$\begin{aligned}\mathcal{F}_t^A(\widetilde{s}_{A,t}, \widetilde{f}_{A,t}, \widetilde{s}_{B,t}, \widetilde{f}_{B,t}) =& \frac{\widetilde{s}_{A,t}}{\widetilde{s}_{A,t}+\widetilde{f}_{A,t}} \cdot \left(1 + \mathcal{F}_{t+1}(\widetilde{s}_{A,t}+1, \widetilde{f}_{A,t}, \widetilde{s}_{B,t}, \widetilde{f}_{B,t})\right) + \\ & \frac{\widetilde{f}_{A,t}}{\widetilde{s}_{A,t}+\widetilde{f}_{A,t}} \cdot \mathcal{F}_{t+1}(\widetilde{s}_{A,t}, \widetilde{f}_{A,t}+1, \widetilde{s}_{B,t}, \widetilde{f}_{Bt}).\end{aligned}$$

On the other hand, if treatment B is allocated to the next patient, then the expected total reward under an optimal policy is

$$\begin{aligned}\mathcal{F}_t^B(\widetilde{s}_{A,t}, \widetilde{f}_{A,t}, \widetilde{s}_{B,t}, \widetilde{f}_{B,t}) =& \frac{\widetilde{s}_{B,t}}{\widetilde{s}_{B,t}+\widetilde{f}_{B,t}} \cdot \left(1 + \mathcal{F}_{t+1}(\widetilde{s}_{A,t}, \widetilde{f}_{A,t}, \widetilde{s}_{B,t}+1, \widetilde{f}_{B,t})\right) + \\ & \frac{\widetilde{f}_{B,t}}{\widetilde{s}_{B,t}+\widetilde{f}_{B,t}} \cdot \mathcal{F}_{t+1}(\widetilde{s}_{A,t}, \widetilde{f}_{A,t}, \widetilde{s}_{B,t}, \widetilde{f}_{B,t}+1).\end{aligned}$$

Therefore, $\mathcal{F}$ satisfies the recurrence

$$\mathcal{F}_t(\widetilde{s}_{A,t}, \widetilde{f}_{A,t}, \widetilde{s}_{B,t}, \widetilde{f}_{B,t}) = \max\left\{\mathcal{F}_t^A(\widetilde{s}_{A,t}, \widetilde{f}_{A,t}, \widetilde{s}_{B,t}, \widetilde{f}_{B,t}),\ \mathcal{F}_t^B(\widetilde{s}_{A,t}, \widetilde{f}_{A,t}, \widetilde{s}_{B,t}, \widetilde{f}_{B,t})\right\}.$$

3.6.2 Computational Speed

Table 3.6.1 illustrates the computational speed of the backwards induction algorithm to compute the allocation policy of the DP design on a standard laptop with 16 GB of RAM. The maximum trial size that can be computed on a standard laptop using `R` is 215. Although trials of sizes larger than 215 are very unlikely to occur in a rare disease context, computations of the DP design are feasible on a standard performance workstation (1 TB of RAM) for $215 < n < 600$. Trials of a size up to 3500 patients would be feasible with today's number one supercomputer (with 1.3 PB of RAM).

n	EPS	Run Time	RAM
10	0.60218	0.01s	0.1 MB
30	0.63066	1s	6.2 MB
50	0.63993	6s	47.7 MB
70	0.64485	24s	183.2 MB
90	0.64799	1m:04s	0.56 GB
110	0.65020	2m:22s	1.1 GB
130	0.65186	4m:37s	2.1 GB
150	0.65316	8m:03s	3.86 GB
200	0.65547	25m:20s	11.9 GB

Table 3.6.1: Expected proportion of successes (EPS) when $s_{A,0} = f_{A,0} = s_{B,0} = f_{B,0} = 1$, i.e. EPS $= \mathcal{F}_0(1,1,1,1)/n$, run time in minutes (m) and seconds (s) and RAM memory requirements of the DP design on a standard laptop.

3.6.3 Choosing the Degree of Constraining, ℓ

Figure 3.6.1 illustrates the non-linearity of the power, based on which we recommend $\ell = 0.15n$ in our proposed CRDP design.

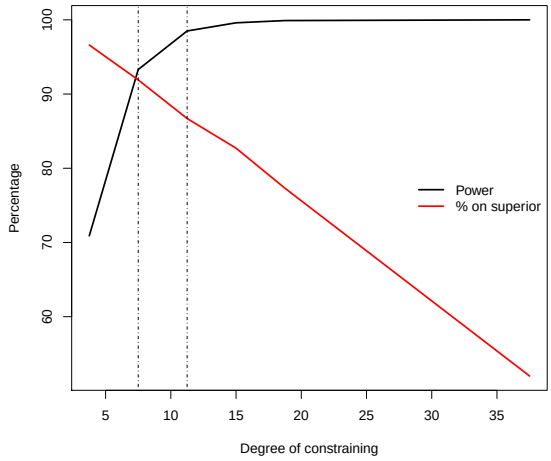


Figure 3.6.1: The effect of changing the degree of constraining, ℓ, on the power and percentage of patients on the superior treatment when $\theta_A = 0.2$ and $\theta_B = 0.8$ for the constrained DP design (without randomisation). The left and right dashed vertical lines correspond to $\ell = 0.10n$ and $\ell = 0.15n$ respectively, where $n = 75$ in this case.

3.6.4 Choosing the Degree of Randomisation, p

Tables 3.6.2–3.6.5 illustrate the effect of varying degrees of randomisation for a range of different scenarios, based on which we recommend $p = 0.9$ in our proposed CRDP design.

p	Bias	MSE	Type I Error	EPS	% on superior
0.5	0.000	0.004	0.035	0.200	50.0
0.6	-0.002	0.004	0.034	0.200	50.1
0.7	-0.001	0.005	0.027	0.200	50.2
0.8	0.000	0.005	0.022	0.200	50.0
0.9	0.000	0.006	0.008	0.200	50.2
1.0	0.001	0.008	0.000	0.200	49.7

Table 3.6.2: The effect of changing the degree of randomisation, p, on the performance measures when $n = 75$ and $\theta_A = \theta_B = 0.2$ for the RDP design (without the constraint).

p	Bias	MSE	Power	EPS	% on superior
0.5	-0.001	0.004	0.428	0.300	50.0
0.6	-0.002	0.005	0.406	0.315	57.3
0.7	-0.003	0.006	0.355	0.329	64.5
0.8	-0.007	0.007	0.289	0.344	71.4
0.9	-0.018	0.010	0.183	0.356	77.9
1.0	-0.058	0.017	0.021	0.368	83.6

Table 3.6.3: The effect of changing the degree of randomisation, p, on the performance measures when $n = 75$, $\theta_A = 0.2$ and $\theta_B = 0.4$ for the RDP design (without the constraint).

p	Bias	MSE	Power	EPS	% on superior
0.5	-0.001	0.004	0.938	0.400	50.0
0.6	-0.002	0.005	0.935	0.437	59.1
0.7	-0.002	0.007	0.910	0.473	68.2
0.8	-0.005	0.009	0.830	0.509	77.3
0.9	-0.015	0.015	0.636	0.544	86.0
1.0	-0.089	0.03	0.070	0.577	94.2

Table 3.6.4: The effect of changing the degree of randomisation, p, on the performance measures when $n = 75$, $\theta_A = 0.2$ and $\theta_B = 0.6$ for the RDP design (without the constraint).

p	Bias	MSE	Power	EPS	% on superior
0.5	-0.001	0.004	1.000	0.500	50.0
0.6	-0.001	0.005	1.000	0.557	59.6
0.7	-0.001	0.007	0.999	0.615	69.2
0.8	-0.004	0.010	0.995	0.672	78.8
0.9	-0.009	0.019	0.937	0.730	88.3
1.0	-0.100	0.043	0.118	0.786	97.6

Table 3.6.5: The effect of changing the degree of randomisation, p, on the performance measures when $n = 75$, $\theta_A = 0.2$ and $\theta_B = 0.8$ for the RDP design (without the constraint).

3.6.5 CRDP Patient Allocation: Other Scenarios

Figure 3.6.2 complements Figure 3.4.6 to show average allocation probabilities of our proposed CRDP design in other scenarios.

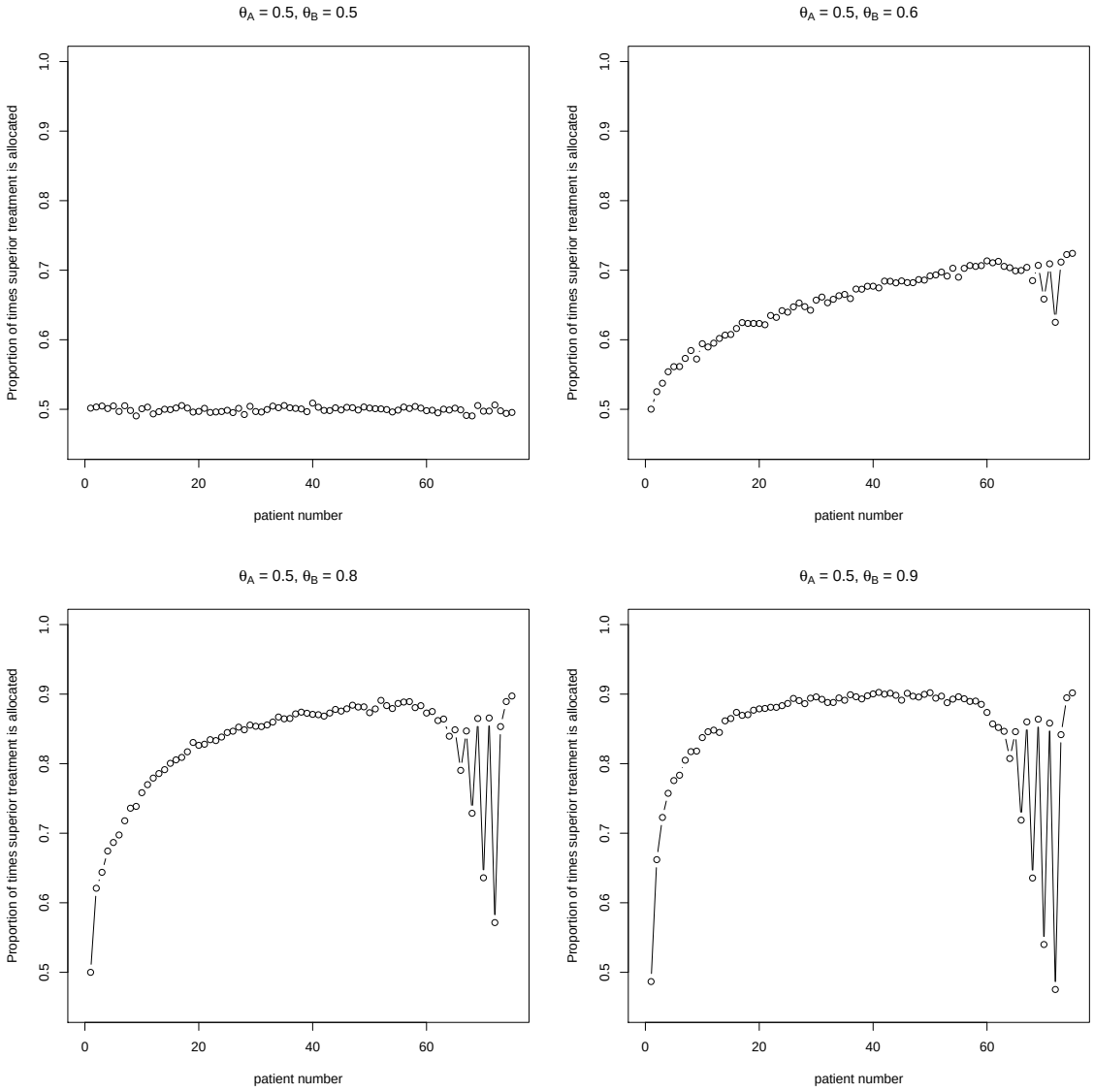


Figure 3.6.2: Probability of allocating a patient to treatment B for CRDP when $\theta_A = 0.5$ and $\theta_B = \{0.5, 0.6, 0.8, 0.9\}$ in a trial of size $n = 75$ estimated over 10,000 simulations.

3.6.6 Combined Performance Measures

Table 3.6.6 summarises the performance of the four key features (power, average bias, MSE and patient benefit) per design by showing the following measures: (i) sum of the distance of each key feature from the best achievable value (SDis), (ii) the maximum difference among each of the four key features from the best achievable value (MD), (iii) sum of the deviations of each key feature from the fixed randomisation design (SDev).

Design	SDis	MD	SDev
CRDP	32.925	24.7	53.513
RDP	36.936	29.7	63.009
DP	74.439	72.3	95.494
WI	73.307	73.2	113.695
RPW	30.714	29.7	11.801
Fixed	40.512	50.0	0

Table 3.6.6: The summary measures of performance in terms of the four key features. SDis: sum of the distance of each key feature from the best achievable value; MD: maximum difference among each of the key features from the best achievable value; SDev: sum of the deviations of each key feature from the fixed randomisation design. Note that these should be treated with some caution since the key features are measured on different scales.

3.6.7 Results for Other Sample Sizes

Figures 3.6.3–3.6.5 complement Figures 3.4.1–3.4.3, respectively, to compare the performance of our proposed CRDP design with alternative designs for different sample sizes.

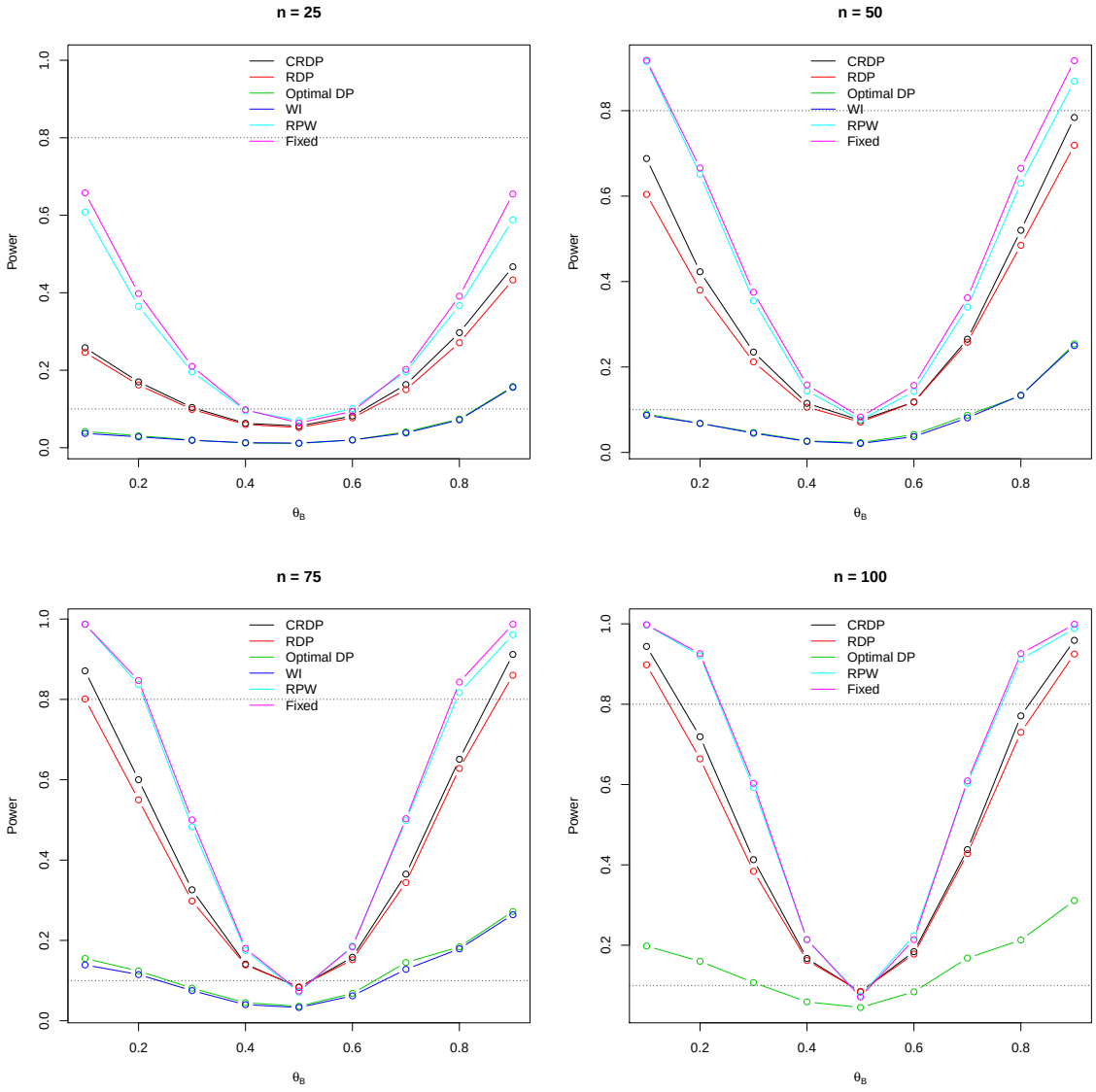


Figure 3.6.3: The changes in power and type I error for each design when $\theta_A = 0.5$ and $\theta_B \in (0.1, 0.9)$ for varying sample sizes. The upper dashed line at 0.8 represents the desired power level, and the lower dashed line at 0.1 represents the nominal significance level.

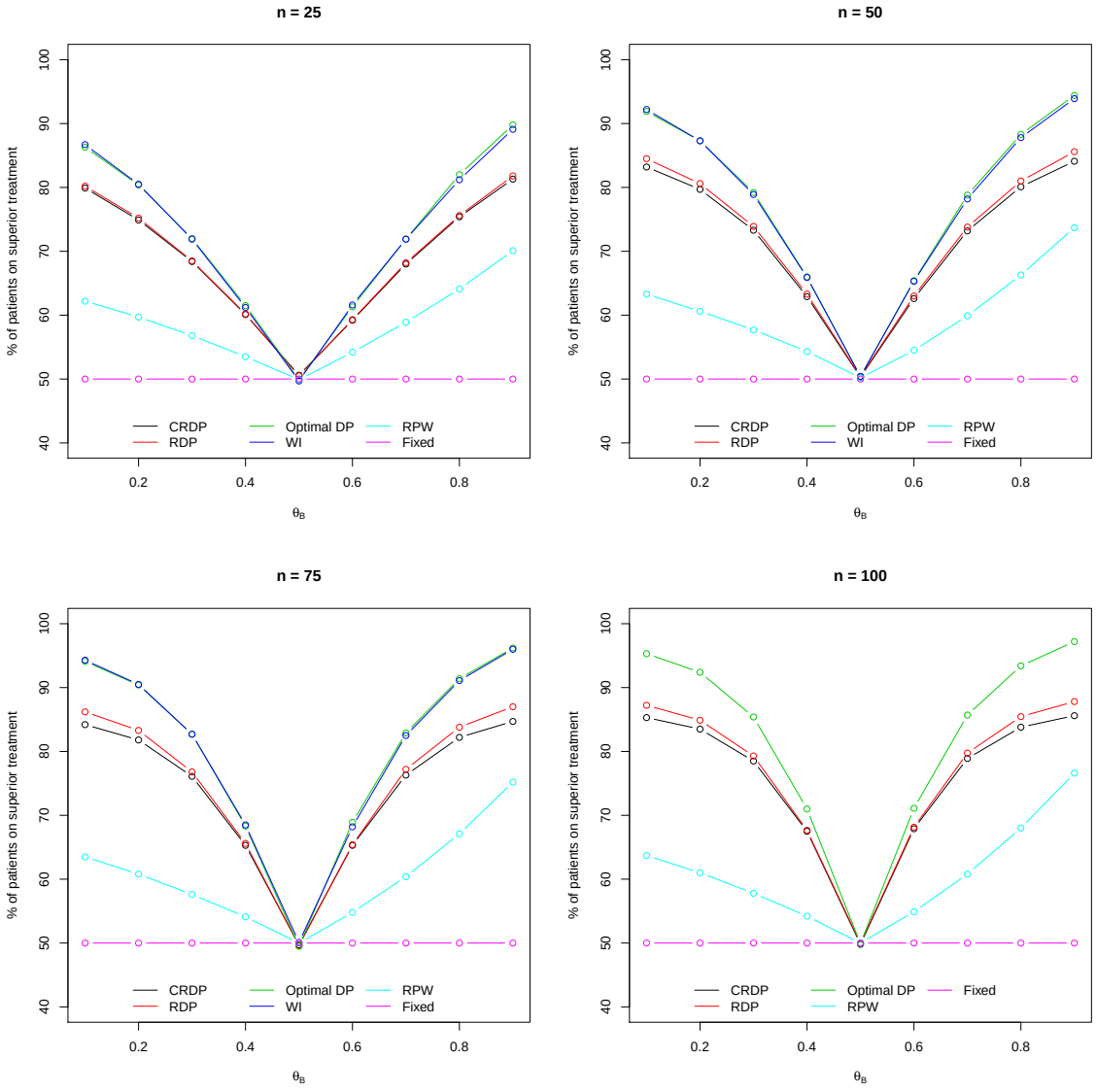


Figure 3.6.4: The percentage of patients on the superior treatment for each design when $\theta_A = 0.5$ and $\theta_B \in (0.1, 0.9)$ for varying sample sizes.

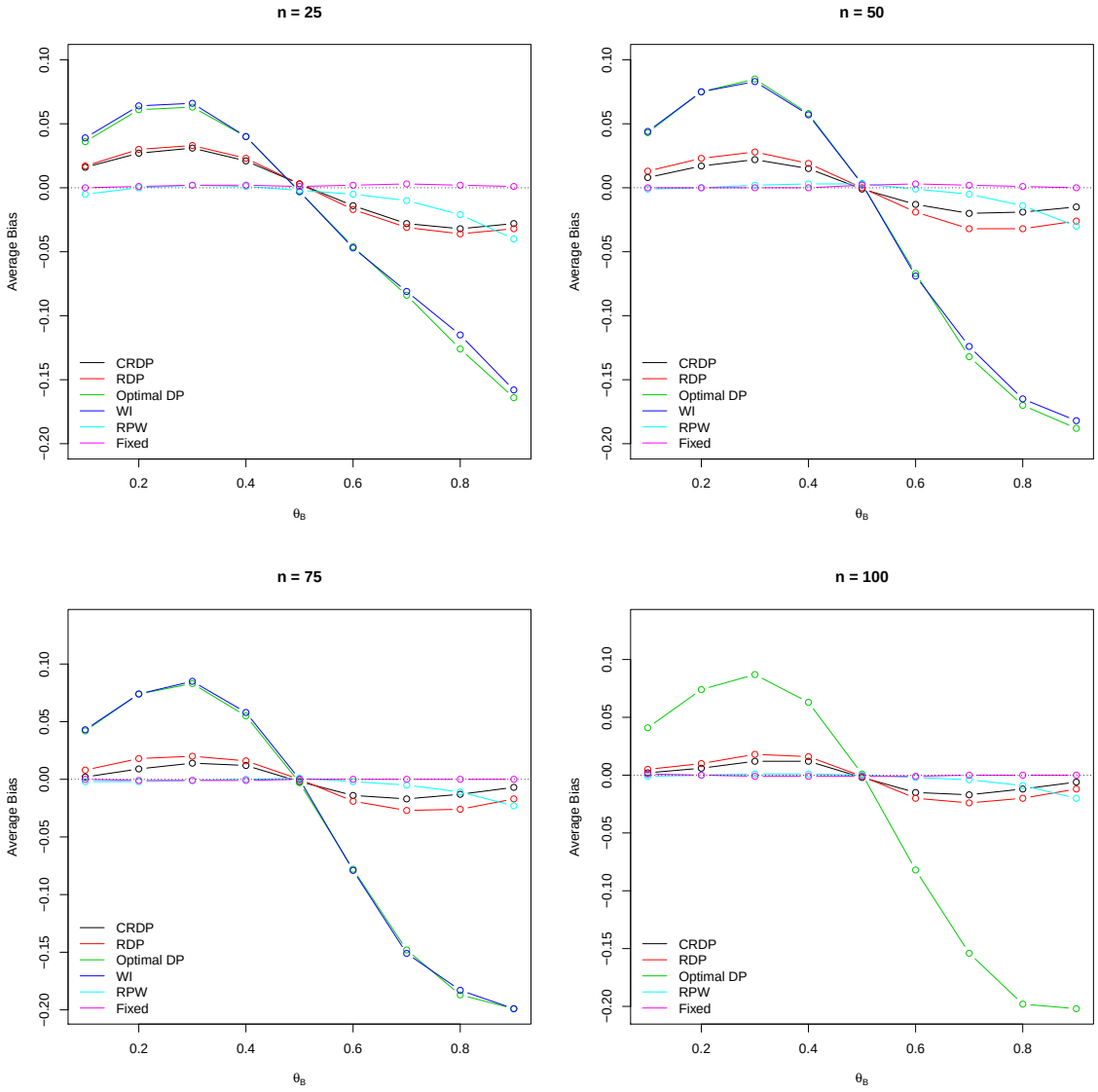


Figure 3.6.5: The average bias of the treatment effect estimator when $\theta_A = 0.5$ and $\theta_B \in (0.1, 0.9)$ for varying sample sizes.

Chapter 4

Extension to Delayed Responses

4.1 Introduction

In Chapter 3, each patient's response is used to inform the subsequent allocation decision which relies on the assumption of immediate patient responses or, more specifically, that a patient's response is available before the next patient enters the trial. Most response-adaptive designs in the literature are typically formulated under this assumption (Cheung *et al.*, 2006; Biswas *et al.*, 2008, Section 3.7). Although this may be appropriate for some clinical contexts, such as trials of surgical interventions (Rosenberger and Lachin, 2016, Chapter 12), trials for diseases with a slow recruitment rate (e.g. rare diseases) or rapidly observed endpoint (e.g. acute diseases), it is unrealistic in many clinical trial settings (e.g. oncology trials). This is because, not only may a treatment take a substantial length of time to induce a response, but there may also be an administrative delay in obtaining the response (Pocock, 1983) or implementing the adaptation to the allocation probabilities which, as Wason *et al.* (2019) discusses, "will reduce the efficiency advantage of an adaptive approach in exactly the same way as using an outcome that takes longer to observe". As a result, responses from all of the previously allocated patients may not be available before allocation of

the next patient; we will refer to these responses, or patients, as being in the *pipeline*[1]. One simple approach is to base the allocation decision only on the currently observed data and ignore the pipeline data. However, this can lead to biased parameter estimates and incorrect allocation decisions (Xu and Yin, 2014). Another possibility is to wait until all of the treated patients have responded before allocating the next patient(s), but this is impractical since the trial will take much longer (Wason *et al.*, 2019). Moreover, it is unethical to withhold treatment from trial participants (Yin, 2012, Chapter 7). Even though it has long been highlighted that "useful methods must have treatment assignment rules that do not require instantaneous observations and reporting" (Simon, 1977), the problem of *how* (response-)adaptive designs can be adjusted to incorporate delayed responses remains an important research question.

Several authors have illustrated the *effect* of delayed responses on particular response-adaptive designs (predominantly urn models), either by simulation, e.g. Robinson (1983); Rosenberger and Seshaiyer (1997); Rosenberger (1999); Ivanova and Rosenberger (2000); Kuleshov and Precup (2000); Karrison *et al.* (2003); Zhang and Rosenberger (2006, 2007); Wason *et al.* (2019), or theoretically, e.g. Bai *et al.* (2002) for urn models; Hu *et al.* (2008) for the doubly adaptive biased coin design. However, few have provided potential solutions to accommodate for the delay (which has been pointed out in many papers, e.g. Hardwick *et al.* (2006); Caro and Yoo (2010); Chick *et al.* (2017)). The inability of most response-adaptive designs to account for the delay has long been cited as one of the greatest limitations and barriers to their implementation in practice (see Simon, 1977; Armitage, 1985; Zhang and Rosenberger, 2006; Villar *et al.*, 2015a,b; Smith and Villar, 2018; Ahuja and Birge, 2019). Sverdlov *et al.* (2012), for example, describe it as "a major stumbling block in implementing adaptive designs", and Rosenberger *et al.* (2012, Section 4) list it as one of the main

[1]The pipeline includes those patients who have been allocated to a treatment but have not yet responded. This is consistent with the terminology used in related literature, e.g. Hampson and Jennison (2013); Chick *et al.* (2017).

criticisms of RAR. As such, there is a strong interest amongst the statistical and clinical trial community in whether RAR methods can be extended to accommodate delayed responses (see e.g. the comment made by D. S. Coad in the Discussion of the paper by Hampson and Jennison (2013)).

Consequently, the main objective of this chapter is to take a step towards addressing this problem by not only exploring the impact of delayed responses on the designs introduced in Chapter 3, but also proposing a method which can utilise the pipeline information in the adaptations. First, we outline some of the few existing designs that do adjust for delays.

Eick (1988b) introduces a model for a bandit problem with delayed responses in the context of a two-armed clinical trial where the distribution of one arm is known, thus referred to as a one-armed bandit, and the other has a geometric response time. The DP solution is presented and compared to the optimal solution (designed to maximise expected total patient lifetime) for the immediate response case. These results have been further extended by Wang (2000) and Wang (2002). Another paper by Eick (1988a) proves the existence and optimality of the Gittins index for the one-armed delayed response bandit when the discount factor is less than 1/2.

Biswas and Coad (2005) comment that "most of the available literature on adaptive designs overlooks possible delays in responses" and suggest how delays can be incorporated into their proposed multi-armed adaptive design for continuous multivariate responses.

Hardwick *et al.* (2006) consider a two-armed trial with Bernoulli responses in which patients arrive via a Poisson process and their response times are assumed to follow independent exponential distributions. The objective function of interest is to maximise the expected number of patient successes during the trial. Therefore, they model this problem as a two-armed bandit with delay which, in theory, is amenable to solution by DP to yield the optimal design. However, as discussed in Chapter

2, this approach is already computationally intensive and the additional complexity caused by the delay increases this even further. They discuss two alternatives based on the recursive DP equations: the first was computationally infeasible at the time of publication, therefore they present a second solution which reduces the computational requirements. In this chapter, we take a similar approach to Hardwick *et al.* (2006) by proposing, and implementing, a design for the delayed response bandit based on the DP equations, yet there are some important differences which will be highlighted in Chapter 5.

Xu and Yin (2014) propose a two-stage non-parametric fractional scheme based on RAR to address the issue of delayed response by treating unobserved outcomes as censored and calculating their fractional contribution to the response probability. See also Chick *et al.* (2017) for another, more recent, example of a design for a two-armed, sequential trial adjusted for delayed responses based on a Bayesian decision theoretic model.

This chapter is organised as follows. In Section 4.2, we begin by exploring the impact of fixed and random delays in responses on both the optimal DP response-adaptive design and the constrained randomised variant (CRDP) introduced in Chapter 3, which we will jointly abbreviate as (CR)DP from hereon for convenience. We compare it to the delayed randomised play-the-winner rule (DRPWR), described in Section 2.1.2, which is well-studied in the literature and is the rule most often suggested for delayed response settings (Hardwick *et al.*, 2006).

Similarly to Chick *et al.* (2017), the remainder of this chapter will focus on the case where there is a *fixed* number of patients in the pipeline at each stage and we will suggest two approaches to account for this delay in Section 4.3, along with corresponding simulation results. The first is an intuitive approach based on altering the time horizon used (see Section 4.3.1), and the second extends the MDP model defined in Chapter 3 by introducing another state variable (see Section 4.3.2). Finally,

the main conclusions are concisely summarised in Section 4.4.

A fixed number of pipeline patients at each stage will be imposed when the time between two consecutive allocations (i.e. the time period) is constant and patients are followed up at a fixed time after treatment (e.g. Facey, 1992; Whitehead, 1993). Although a patient response may occur at any time, in binary response trials (considered in this chapter), interest is only in if it has occurred by the specified follow-up time. Other examples leading to a fixed delay in response are due to administrative delays, such as staff availability, resource limitations, time taken to obtain the results (e.g. patients may require a blood test to determine whether the treatment has been successful, the results of which may only be available one week later), time taken to update and implement the adaptations, etc. (Pocock, 1983; Wason *et al.*, 2019). In the literature, other authors (including Langenberg and Srinivasan, 1981, 1982; Chick *et al.*, 2017) have also formulated delayed response models based upon the assumptions of a constant time period and fixed time until response. Furthermore, focusing on the fixed delay model is a natural first step which will aid in the development of a solution for more complex delay structures, such as when the number of patients in the pipeline is instead random (as in Chapter 5).

4.2 The Effect of Delayed Responses on (CR)DP

4.2.1 Trials with a Fixed Delay

In this section, we focus on a *deterministic* delayed response model which assumes that there is a constant time between allocations and a fixed delay of length $d > 0$ between allocating a patient to a treatment and observing their outcome. As a result, we will know exactly how many patients are in the pipeline at each stage in the trial which, for $t \in \{d+1, \ldots, n\}$, will remain of fixed length equal to d.

In order to explore the impact of delayed responses when applying the (CR)DP

designs, we use simulation to evaluate its performance in a range of scenarios for different delay lengths. By first understanding the impact of a delayed response, we can then take steps to modify the design accordingly. Moreover, as Wason *et al.* (2019) point out, "it is important that theoretical work that proposes and promotes adaptive designs clearly lays out any reduction in their reported efficiency benefits when there is substantial delay in outcome evaluation". We pay particular attention to the scenario in which $\theta_A = 0.5 \; \forall \; \theta_B \in (0.1, 0.9)$ and $n = 75$ so results are consistent with, and comparable to, those reported in Chapter 3. Since more interest is in what happens for shorter delays, as this is where the most marked changes in performance of these designs occur, the results are illustrated for $d = 0, 5, 15, 25, 50$ and 75. Furthermore, from a practical viewpoint, "adaptive allocation has no benefit when there are long delays" (Berry and Stangl, 1996, Chapter 4) because there is little, or no, chance to adapt the allocation, thus it would be inappropriate to employ an adaptive design in this setting. The reason for including the results for no delay is so we can clearly evaluate how the delayed responses are affecting the performance measures relative to the base case. Further, recall that $d = 75$ corresponds to fixed, equal randomisation. The results illustrated in Figure 4.2.1 correspond to changes in the performance of the CRDP design, and analogous results for the DP design are displayed in Figure 4.5.1 of the Appendix 4.5. We include results for the DP design to show how the delay affects the design in the absence of the randomisation and constraining.

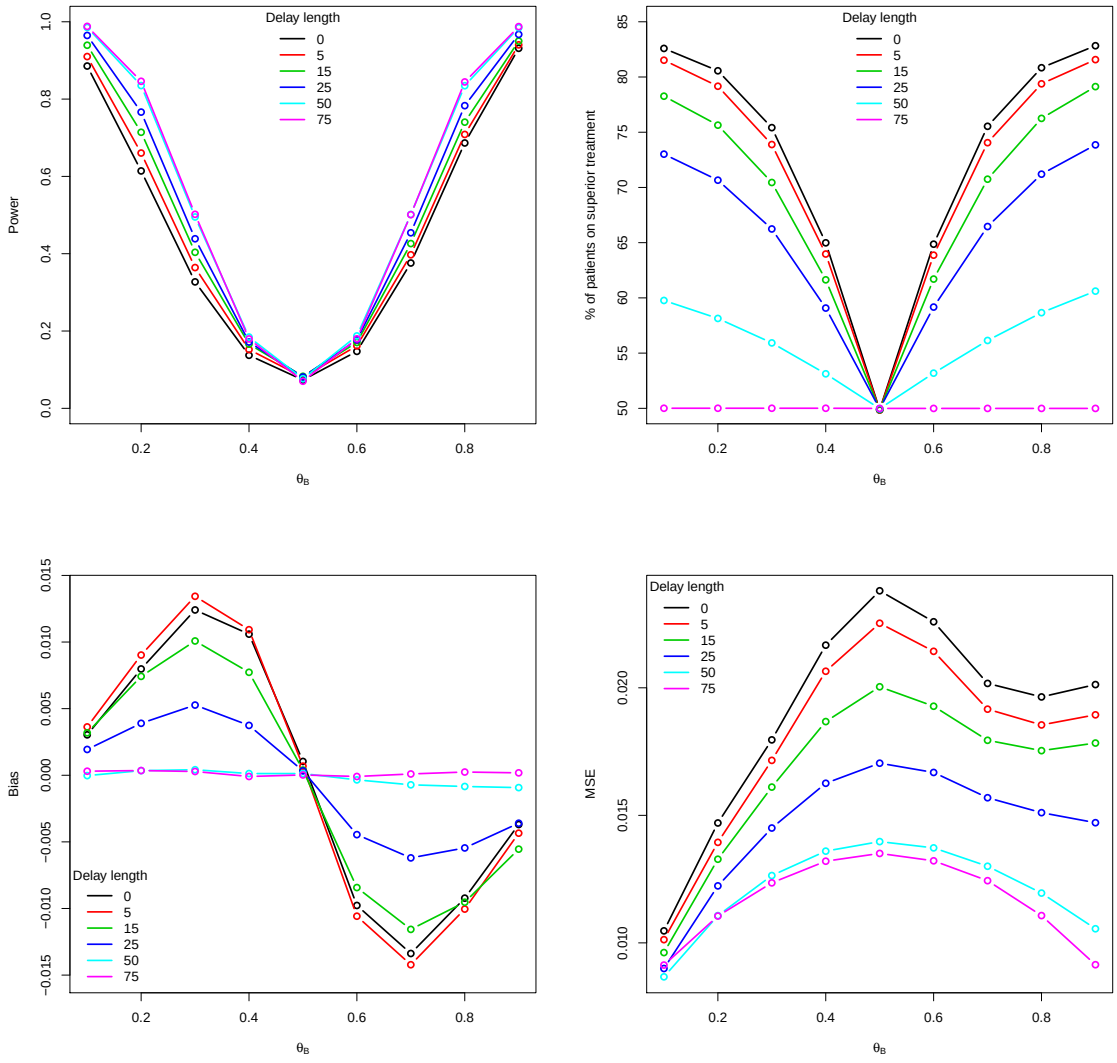


Figure 4.2.1: The changes in power (and type I error), % of patients on the superior treatment, the average bias and MSE of the treatment effect estimator for the CRDP design when $n = 75$, $\theta_A = 0.5$ and $\theta_B \in (0.1, 0.9)$ for different fixed delay lengths (estimated over $100,000$ simulations).

Power. The top left plot in Figure 4.2.1 illustrates the changes in statistical power for CRDP, with the results for $\theta_A = \theta_B$ corresponding to the type I error rates. The most notable observation is that the power *increases* with delay length. This is what we expect because as the length of the delay increases, the adaptation is slowed and the closer the design is to pure randomisation meaning there is less imbalance

between the treatment arms (and hence a greater power). However, the observed changes in power are only small, even for a considerable increase in the delay length. For example, when $\theta_B = 0.1$, the change in power from the "worst" case (i.e. when $d = 0$) to the "best" case (i.e. when $d = 75$) is approximately 10%, on average.

As mentioned previously, we see that the larger changes in power happen for shorter delays, with negligible changes as the delay length increases from 50 to 75, for example. The obvious patterns, such as the power increasing with the size of the treatment difference, are evident for all delay lengths. In terms of the type I error rates, we see that they are seemingly well controlled since they lie close to the nominal significance level at 0.1.

Patient benefit. The top right plot in Figure 4.2.1 illustrates the changes in the percentage of patients allocated to the superior treatment, i.e. the patient benefit, for CRDP. When $\theta_A = \theta_B$, the design allocates approximately 50% of patients to the superior treatment whatever the delay length, as expected. In general, we observe that the number of patients in the trial receiving the superior treatment *decreases* as the delay length increases. Again, this is not surprising since a longer delay means that there are fewer responses available to update the allocation probabilities, and the longer the equal randomisation phase at the start of the trial. This also confirms what has been noted in the literature; for example, in a Technical Report on two-armed bandit strategies by Berry (1976, Section 6), it was mentioned that "there is a decrease in the maximal expected proportion of success when there is response delay".

Consider the scenario in which $\theta_A = 0.5$ and $\theta_B = 0.1$. For the case of no delay, approximately 83% of patients in the trial are allocated to the superior treatment and for a delay of length 25, approximately 73% of patients are allocated to the superior treatment. Thus, we only lose roughly 10% of the patient benefit in this case (which is not a huge price to pay for a delay which causes one third of the information to be excluded). Furthermore, compared to standard randomisation (illustrated by the pink

line in Figure 4.2.1), the gain in patient benefit is still considerably higher. Even for a delay length of 50 (two thirds of the trial size), there are still worthwhile gains, relative to equal randomisation, with approximately 10% more patients being allocated to the superior treatment.

Bias. The bottom left plot of Figure 4.2.1 shows the changes in the average bias of the treatment effect estimator $\hat{\Delta} = \hat{\theta}_A - \hat{\theta}_B$ (where $\hat{\theta}_A = s_{A,n}/N_A$ and $\hat{\theta}_B = s_{B,n}/N_B$ are the observed proportions of successes on treatment A and B, respectively, by the end of the trial). We observe that, in general, the bias *decreases* as the delay length increases (with some slight discrepancy for delay lengths of 0 and 5). This pattern can be seen more clearly in the corresponding plot for the DP design (Figure 4.5.1 of the Appendix 4.5). The decrease in bias seems sensible since the values of $s_{A,n}$, $s_{B,n}$, N_A and N_B will be varying with delay length. As an example, consider the scenario in which $\theta_A = 0.5$ and $\theta_B = 0.1$. For shorter delays, there will be fewer patients allocated to the inferior treatment (arm B) so that $N_B < N_A$. As a result, $\hat{\theta}_B$ will be underestimated, which is shown in Chapter 3, so the treatment effect estimator, $\hat{\Delta}$, will be larger, leading to a larger bias. Alternatively, as $d \to 75$, then $N_B \to N_A$ until eventually $N_B \approx N_A$ when $d = 75$, i.e. as the delay increases, there will be less imbalance between the two treatment arms. Therefore, $\hat{\theta}_A$ and $\hat{\theta}_B$ will be closer to their true values, hence giving rise to a smaller bias. Note that it will be useful to look at the values of the raw estimates of θ_A and θ_B in Table 4.2.1 to support this.

It is also worth noting that the bias values for a delay of 50 are very close to those for equal randomisation, i.e. they are lying close to zero across all scenarios.

True		CRDP with delay 5				DRPWR with delay 5			
θ_A	θ_B	$\hat{\theta}_A$	$\hat{\theta}_B$	$\hat{\theta}_A - \hat{\theta}_B$	Bias	$\hat{\theta}_A$	$\hat{\theta}_B$	$\hat{\theta}_A - \hat{\theta}_B$	Bias
0.500	0.100	0.499853	0.096223	0.403630	0.003630	0.496576	0.097737	0.398840	-0.001160
0.500	0.200	0.497806	0.188783	0.309024	0.009024	0.496257	0.196250	0.300007	0.000007
0.500	0.300	0.491774	0.278339	0.213435	0.013435	0.495755	0.295181	0.200574	0.000574
0.500	0.400	0.480684	0.369758	0.110926	0.010926	0.495331	0.394947	0.100384	0.000384
0.500	0.500	0.470749	0.470066	0.000684	0.000684	0.494164	0.494433	-0.000269	-0.000269
0.500	0.600	0.469279	0.579858	-0.110578	-0.010578	0.492965	0.594547	-0.101582	-0.001582
0.500	0.700	0.477296	0.691518	-0.214222	-0.014222	0.490896	0.695328	-0.204432	-0.004432
0.500	0.800	0.487732	0.797777	-0.310045	-0.010045	0.487573	0.796879	-0.309306	-0.009306
0.500	0.900	0.495412	0.899759	-0.404347	-0.004347	0.480832	0.898330	-0.417498	-0.017498

True		CRDP with delay 25				DRPWR with delay 25			
θ_A	θ_B	$\hat{\theta}_A$	$\hat{\theta}_B$	$\hat{\theta}_A - \hat{\theta}_B$	Bias	$\hat{\theta}_A$	$\hat{\theta}_B$	$\hat{\theta}_A - \hat{\theta}_B$	Bias
0.500	0.100	0.499554	0.097617	0.401938	0.001938	0.497334	0.097964	0.399370	-0.000630
0.500	0.200	0.497649	0.193748	0.303900	0.003900	0.497083	0.196814	0.300270	0.000270
0.500	0.300	0.493642	0.288373	0.205269	0.005269	0.496828	0.296256	0.200572	0.000572
0.500	0.400	0.488466	0.384723	0.103742	0.003742	0.496502	0.396201	0.100300	0.000300
0.500	0.500	0.484371	0.484043	0.000329	0.000329	0.496184	0.496505	-0.000321	-0.000321
0.500	0.600	0.483978	0.588435	-0.104456	-0.004456	0.495625	0.596536	-0.100910	-0.000910
0.500	0.700	0.487662	0.693864	-0.206202	-0.006202	0.494818	0.697139	-0.202321	-0.002321
0.500	0.800	0.492379	0.797834	-0.305455	-0.005455	0.494015	0.798157	-0.304142	-0.004142
0.500	0.900	0.496040	0.899640	-0.403599	-0.003599	0.492782	0.898929	-0.406147	-0.006147

Table 4.2.1: The success probability estimates, $\hat{\theta}_A$ and $\hat{\theta}_B$, for treatments A and B, respectively, compared to their true values, θ_A and θ_B, following CRDP and DRPWR with a fixed delay. These results correspond to the scenarios in which $n = 75$, $\theta_A = 0.5$ and $\theta_B \in (0.1, 0.9)$ for a fixed delay of 5 (upper table) and 25 (lower table).

Mean squared error. The bottom right plot in Figure 4.2.1 shows that the mean squared error (MSE) of the treatment effect estimator *decreases* as the delay length increases across all scenarios. Since the MSE depends on the bias, this pattern could simply be due to the bias values decreasing with delay, but after plotting only the variances of the treatment effect estimator (not included here), which follow exactly the same pattern as the MSE plots, this confirms that the variability of the estimator does indeed decrease with delay.

A further interesting observation is that for longer delays, the MSE (and variance) plots become more symmetric around $\theta_B = 0.5$, yet for shorter delays, the MSE/variance of the estimator begins to slightly increase again after $\theta_B = 0.8$ (see

delay lengths of $0, 5$ and 15, for example, on the MSE plot in Figure 4.2.1).

Similar patterns of results are observed for the DP design in Figure 4.5.1 of the Appendix 4.5, but the changes with delay length are much more pronounced due to the lack of randomisation and constraining in the design meaning a greater level of imbalance can occur. Consider the first scenario in which $\theta_B = 0.1$. When the delay is 25 (one third of the trial size), there is a loss of approximately 15% in patient benefit relative to the optimal value attained in the no delay case. However, the percentage of patients on the superior treatment is still approximately 30% larger than equal randomisation. In terms of the power, a delay of 25 increases it by nearly as much as 80% relative to when there is no delay. Moreover, it lies very close to the power obtained by equal randomisation. For a delay of 15, the percentage of patients on the superior treatment is reduced by only around 8%, but the power is increased by approximately 66%. Therefore, by introducing a delay in response, although the DP design is no longer optimal with respect to patient benefit, it still allocates a considerably large percentage of patients to the superior treatment whilst achieving a substantially improved power.

Note that the scale of the bias and MSE plots for the DP is much larger than that used for the corresponding plots for the CRDP.

Comparison to DRPWR

In this section, we explore how the most commonly proposed rule for such problems, the DRPWR (see Section 2.1.2), compares to the (CR)DP designs for a range of delay lengths. We focus on two scenarios, with and without a treatment difference, to illustrate the differences between the performance measures of these designs. Note that plots corresponding to Figure 4.2.1 for just the DRPWR are provided in Figure 4.5.3 of the Appendix 4.5.

(i) **Scenario 1**: $\theta_A = 0.5$ and $\theta_B = 0.1$

Power. The first plot in Figure 4.2.2 illustrates the changes in power as the delay length, d, increases. We have already identified that the power of the (CR)DP design increases with delay, however this plot gives a much clearer visualisation of the rate of this increase. In particular, we see that it increases hyperbolically with the largest changes occurring for shorter delay lengths and practically no change occurring as d increases from 40 to 75.

In contrast, the length of the delay does not seem to affect the power of the DRPWR, which remains fairly constant for all delay lengths. The power of the RPWR is already high when there is no delay, because it does not create enough imbalance between the two treatments, and thus there is little room for improvement.

Comparing designs for this particular case, we see that although the DRPWR attains the highest power for delays up to around 45 (at which point all of the designs converge), the CRDP design still performs very well, whereas the power of the DP design is insufficient and lies below 80% for delays up to length 15. For example, when the delay is 5, the power of DRPWR and CRDP is above 90% but for DP, it is close to 50%.

Patient benefit. The second plot in Figure 4.2.2 shows how the percentage of patients allocated to the superior treatment varies as d increases. Similarly to the (CR)DP, as the delay length increases, the DRPWR allocates fewer patients to the superior arm. Again, this plot allows us to visualise the rate of this decrease much more clearly. For DP, we observe that the percentage of patients allocated to the superior treatment decreases linearly at a relatively constant rate compared to the CRDP which decreases at a slower rate, and the DRPWR which decreases at a much slower rate than both (CR)DP designs. Further, (CR)DP allocates substantially more patients to the superior treatment than the DRPWR, most markedly for shorter delay lengths. For example, Figure 4.2.2 shows that when $d = 5$, DP and CRDP allocate

approximately 30% and 20% more patients, respectively, to the superior arm than DRPWR.

Bias. The third plot in Figure 4.2.2 illustrates the changes in the average bias of the treatment effect estimator as the delay length varies. We have already identified that, generally, the bias of the (CR)DP design decreases with delay, and this plot shows that this happens for shorter delays up to around $d = 30$, after which point the bias fluctuates very closely around 0. Moreover, the bias values decrease at a much quicker rate for the DP design. In contrast, the bias values following the DRPWR appear to be fairly robust to changes in delay, remaining close to 0 for all delay lengths, with a very slight decrease evident as d increases.

Note that the scale of this plot is very small and although the DRPWR appears to perform slightly better with respect to bias for shorter delay lengths, the differences are only negligible (to three or four decimal places).

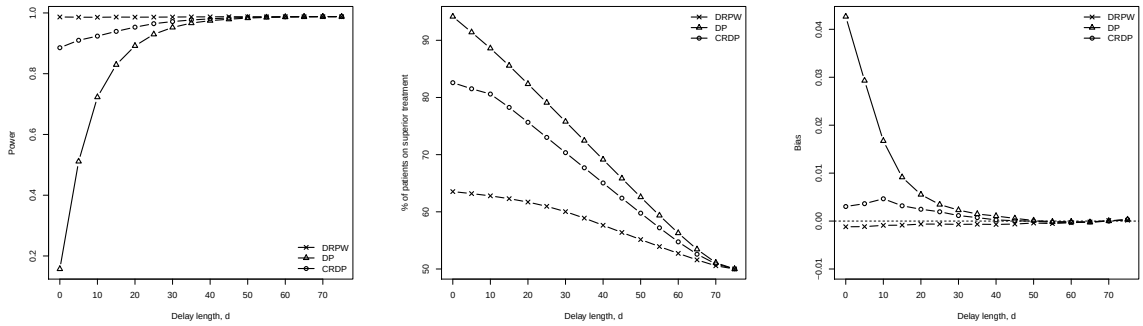


Figure 4.2.2: The changes in power, % of patients on the superior treatment and the average bias of the treatment effect estimator for (CR)DP and DRPWR as the length of the fixed delay increases, when $n = 75$, $\theta_A = 0.5$ and $\theta_B = 0.1$ (estimated over $100,000$ simulations).

(ii) **Scenario 2**: $\theta_A = \theta_B = 0.5$

Corresponding plots when there is no treatment difference are shown in Figure 4.2.3. Here, we notice that the differences in the performance measures of the (CR)DP and DRPWR are much less pronounced, as expected.

The first plot illustrates the changes in type I error rates for the (CR)DP and

DRPWR as the delay increases. We observe that the overall trend of the type I error rate appears to decrease with d for all designs, although more so for the (CR)DP design than the DRPWR. Further, the type I error rates for the DRPWR are consistently smaller, albeit very slightly, than those for (CR)DP (with delay) until around $d = 60$, after which they perform similarly.

Since the treatments have the same success rates, the percentage of patients allocated to either treatment behaves accordingly, that is, close to 50% irrespective of the design or delay length. Similarly, the bias values lie very close to 0 for all delay lengths regardless of the design since there is no imbalance between the two treatment arms.

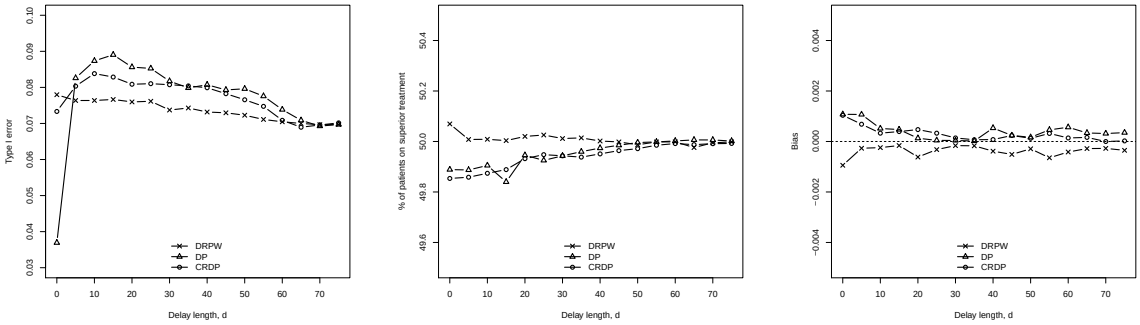


Figure 4.2.3: The changes in type I error, % of patients on the superior treatment and the average bias of the treatment effect estimator for (CR)DP and DRPWR as the length of the fixed delay increases, when $n = 75$, $\theta_A = \theta_B = 0.5$ (estimated over $100,000$ simulations).

4.2.2 Trials with a Random Delay

The assumption of a fixed delay in Section 4.2.1 is not very realistic in a clinical trial context where the length of the delay may vary from patient to patient. Thus, in this section, we relax the assumption of a fixed delay, and consider a simple stochastic delayed response model in which patients still arrive sequentially, at a constant rate,

but the length of time to observe a response is now random[2]. As a result, in contrast to Section 4.2.1, the number of patients in the pipeline at any stage of the trial is also random.

We use a Bernoulli random variable with probability r to determine which patients in the pipeline have responded at each stage in the trial. If a patient has responded, we record their observation, update the states accordingly and remove this patient from the pipeline. Otherwise, if the patient has not yet responded, they remain in the pipeline and we simply proceed to allocate the subsequent patient based on whatever information is currently available.

Recall that a geometric distribution models the number of independent and identically distributed Bernoulli trials before the first success. Therefore, using a Bernoulli random variable at each stage to determine whether there has been a patient response is equivalent to assuming a geometric response time. If Y_i denotes the response time, or equivalently the delay length, of patient $i = 1, \ldots, n$, then $Y_i \sim \text{Geometric}(r)$, with probability mass function given by $(1-r)^t r$ for $t = 0, 1, \ldots$ and $0 < r \leq 1$. Note that Eick (1988b) also considered a geometric response time when investigating the one-armed bandit problem with delay.

We will now vary the success probability r, i.e. the probability of a patient responding at each stage, in order to explore the impact of random delays on the (CR)DP designs. So that the results are presented similarly to those in Section 4.2.1 for the fixed delay case, we will illustrate the performance measures for different *expected* delay lengths. Since the expected value of a geometric random variable Y_i is given by $\mathbb{E}(Y_i) = (1-r)/r$, to do this, we will choose values of $r = 1/(1+\mathbb{E}(Y_i))$ such that $\mathbb{E}(Y_i) = 0, 5, 15, 25, 50, 75$ and 100 for each i. Note that we include an expected

[2]Although it is not typical in clinical trial practice to have a binary endpoint that is randomly observed, we use it purely for the purpose of illustrating the effect of random delays on (CR)DP since it is more intuitive to interpret a random delay as being the random time from allocation to response. This set-up is also used in Hardwick *et al.* (2006). However, the equivalent — but more realistic — reformulation in terms of random arrivals with a fixed follow-up time is considered in Chapter 5.

delay length of 100 here to demonstrate that, in the random delay case, the (CR)DP gives rise to different performance measures for expected delays greater than the trial size, i.e. 75. This is in contrast to the fixed delay case in which, for all delays ≥ 75, (CR)DP mimics equal randomisation and thus behaves the same.

Figure 4.2.4 is the analogue of Figure 4.2.1 but for the random delay case. The overall trends observed for the performance measures as the expected delay lengths increase are similar to those for the fixed delay case. However, we see that there are some immediate differences as a result of the additional variability incurred by the random delay. In particular, the top right plot of Figure 4.2.4 shows that the percentage of patients allocated to the superior treatment appears to be much larger for the random delay case (explained below). The bias and MSE values are also larger when the delay is random, and there is little difference in the power as the expected delay length increases. These observations are due to a mixture of reporting averages and the fact that there is inherent variability in the results that goes beyond that of simulation error, owing to the underlying random nature of the delay.

The corresponding plot illustrating the effect of a random delay on the performance of the DP design is shown in Figure 4.5.2 of the Appendix 4.5.

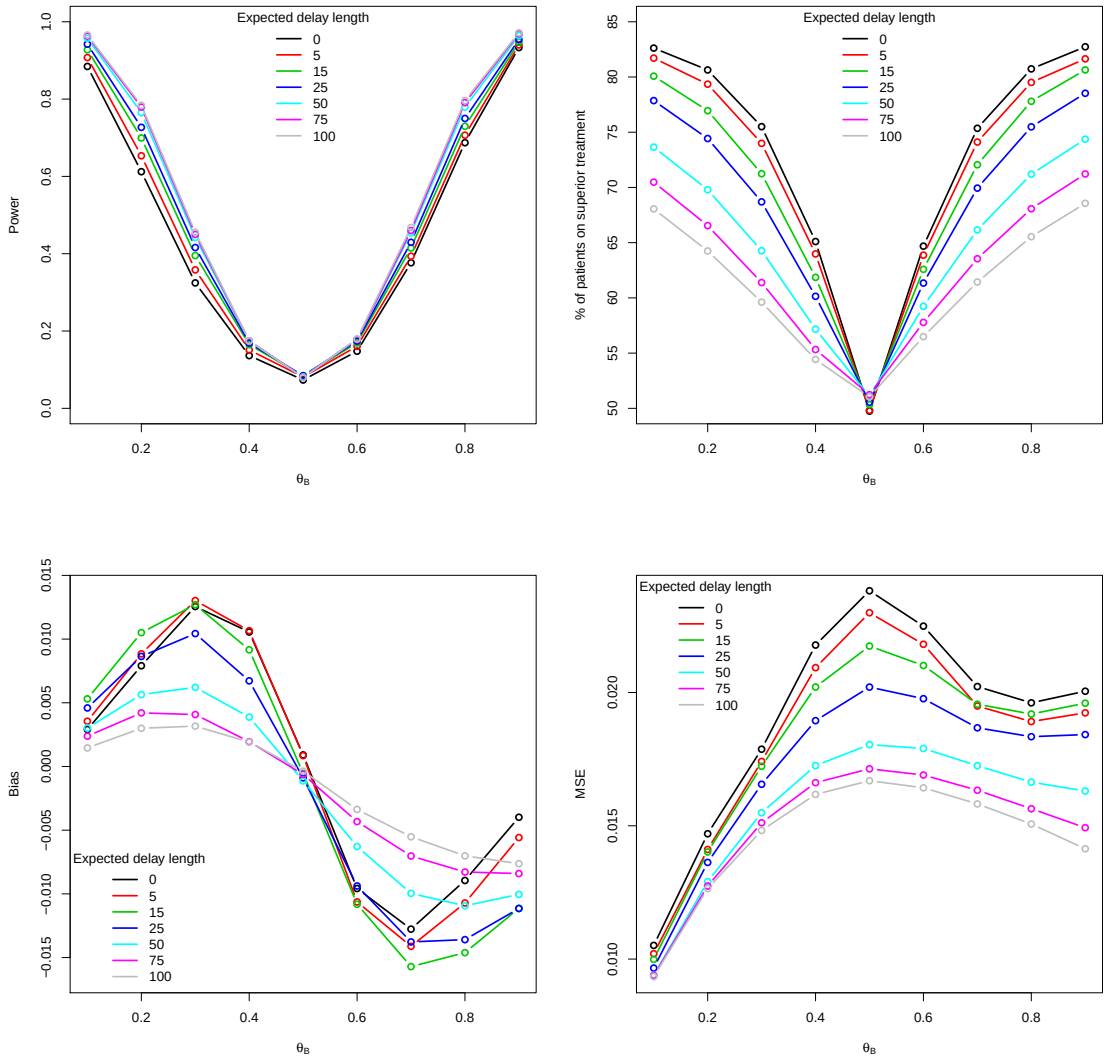


Figure 4.2.4: The changes in power (and type I error), % of patients on the superior treatment, the average bias and MSE of the treatment effect estimator for the CRDP design when $n = 75$, $\theta_A = 0.5$ and $\theta_B \in (0.1, 0.9)$ for different expected random delay lengths (estimated over $100,000$ simulations).

Comparison to Fixed Delay

We now compare the performance measures of the (CR)DP with a fixed delay versus (CR)DP with a random delay for a specific scenario in which $\theta_A = 0.5$ and $\theta_B = 0.1$ (see Figure 4.2.5). Note that this comparison is fair in the sense that we have cali-

brated the random delays so that we expect them to be the same length, on average, as the fixed delays. However, it is not fair in the fact that one has variability whilst the other does not. Thus, we use this comparison purely for illustrative purposes to highlight the differences that can occur as a result of the delay being random rather than fixed. Figure 4.2.5 shows that there is a smaller power, more patients on the superior treatment and a larger bias observed due to the additional uncertainty in the random delay. Furthermore, it is interesting to note that although we expect the percentage of patients on the superior treatment (% *on sup*) to be 50% when the expected delay length is 75, as it is for the fixed delay of 75, it is actually closer to 70% for the CRDP and 79% for the DP (see the middle plot in Figure 4.2.5).

The reason for this is that we will have some simulation runs where there is no delay, by random chance, in which case we are back to the standard (CR)DP design and will have high values for % *on sup*, and other runs where there is complete delay, in which case we are at the other extreme of pure randomisation and will consequently have values close to 50% *on sup* (and everything in between as well). Therefore, the estimates for the % *on sup* will lie somewhere in between. Similarly for the bias, which we expect to be 0 when the expected delay length is 75, but it is actually higher. Moreover, recall that the variance of a geometric random variable is given by $(1-r)/r^2$. This means that the variability increases as r decreases or, equivalently, as the expected delay length increases. Thus, the large amount of variability observed, particularly for longer expected delay lengths, is what we would expect from a geometric random variable.

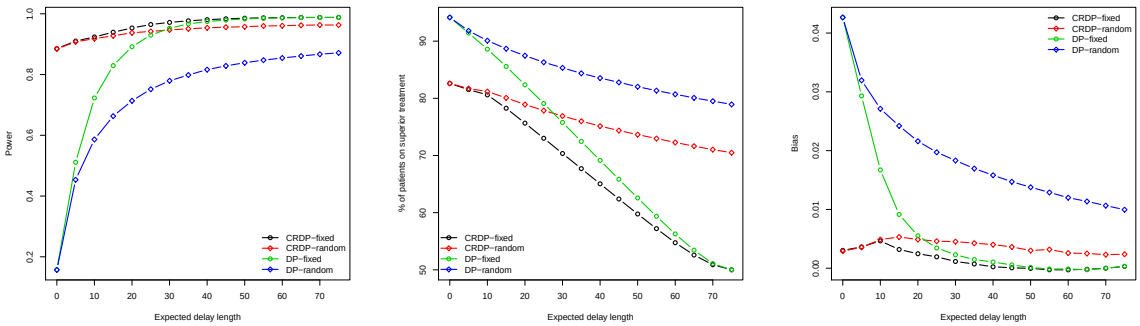
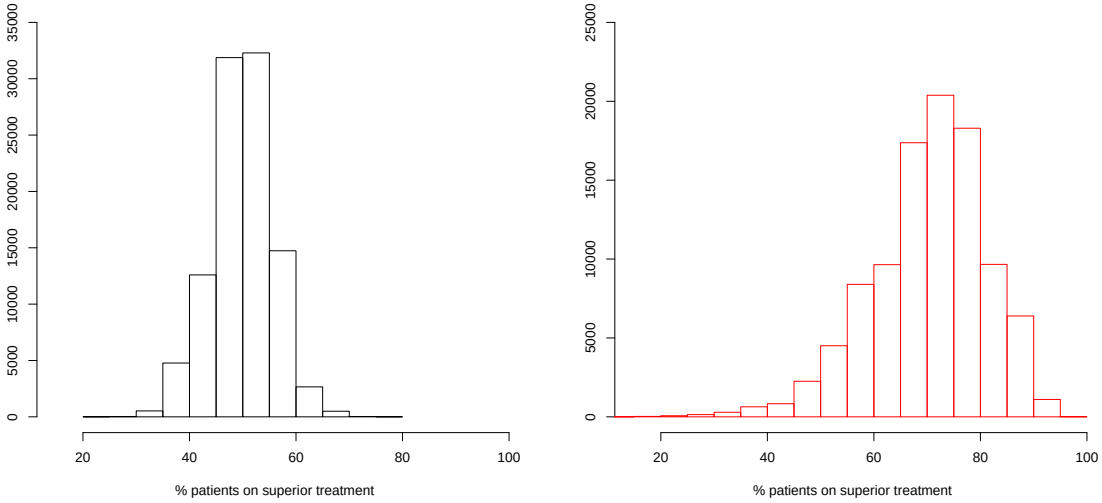


Figure 4.2.5: The changes in power, % of patients on the superior treatment and the average bias of the treatment effect estimator for the CRDP design as the fixed/expected delay length increases, when $n = 75$, $\theta_A = 0.5$ and $\theta_B = 0.1$ (estimated over $100,000$ simulations).

To confirm our justification that these differences are indeed due to the increased variability prevalent in the random delay, we consider the actual distributions of the simulations (in the form of histograms) rather than simply summarising the results as means which we have been doing thus far. Since there is a very large difference between the % *on sup* obtained for the fixed and random delays when the delay is 75, we will illustrate the corresponding histograms for this case in Figure 4.2.6.



(a) CRDP with fixed delay (b) CRDP with random delay

Figure 4.2.6: Histograms showing the distribution of the $100,000$ simulations for the % of patients on the superior treatment when the fixed/expected delay length is 75, $n = 75$, $\theta_A = 0.5$ and $\theta_B = 0.1$.

From Figure 4.2.6a, which corresponds to a fixed delay, we see that the distribution looks approximately normal with most of the 100,000 simulations centred around 50%, as expected. However, from Figure 4.2.6b, we see that the distribution covers a much wider range of values (with some of the simulations producing values below 20% and above 90%). Further, it is skewed to the right with most of the simulations concentrated around 70%. This clearly illustrates the increased variability in the results when there is a random delay and thus justifies the differences observed in the performance measures.

Comparison to DRPWR

We now compare the performance of the (CR)DP in trials with a random delay to the DRPWR, as we did in Section 4.2.1 for trials with fixed delays. Similarly, we consider how the performance measures vary with the *expected* delay length for (i) a treatment difference and (ii) no treatment difference. For an alternative illustration of how the DRPWR (with random delay) behaves for a wider range of scenarios under different expected delay lengths, see Figure 4.5.4 in Appendix 4.5.

(i) **Scenario 1:** $\theta_A = 0.5$ and $\theta_B = 0.1$

Power. The first plot in Figure 4.2.7 shows the changes in power for the (CR)DP and DRPWR as the expected delay length increases. As in the fixed delay case, the greatest changes in power for the (CR)DP designs occur for shorter expected delay lengths, although now at a slower rate. For CRDP, the power remains constant for delays expected to be greater than 65, but for DP it continues increasing for all of the expected delay lengths plotted. The power of the DRPWR, on the other hand, remains relatively stable for all expected delay lengths and attains values very close to those obtained when there is a fixed delay.

Relative to the DRPWR, the (CR)DP designs have smaller power for all expected delay lengths. Again, this difference is much more prominent for DP. For example,

when the delay length is expected to be 5, the power is 8% smaller for CRDP and 53% smaller for DP compared to the DRPWR. For expected delays over 40, the difference in power between the DRPWR and CRDP is expected to be at most 3%.

Patient benefit. The second plot in Figure 4.2.7 compares how the percentage of patients allocated to the superior treatment varies as the expected delay length increases for the (CR)DP and DRPWR. Again, we see that the % *on sup* for the (CR)DP decreases at a slower rate than when the delay is fixed, but at a faster rate than the DRPWR which only decreases by a small amount (2.7%) as the expected delay increases from 0 to 100. Moreover, the rate of change for these designs remains relatively constant. Compared to the DRPWR, the (CR)DP allocates significantly more patients to the superior treatment for all expected delay lengths considered. In particular, for an expected delay length of 5, DP and CRDP allocate approximately 30% and 20% more patients, respectively, to the superior arm than the DRPWR, which is a huge improvement and is the same as what we observed in the fixed delay case.

Bias. The third plot in Figure 4.2.7 illustrates the changes in the average bias of the treatment effect estimator as the expected delay length varies. Overall, for the CRDP design, the trend in bias appears to be decreasing, which is much more apparent for the DP. The bias values corresponding to the DRPWR do not change much with the expected delay and lie slightly closer to 0 than the CRDP for all expected delay lengths. However, it must be remembered that the scale of this plot is very small so the differences in the bias between the DRPWR and CRDP are trivial. Both the DRPWR and CRDP consistently outperform the DP, but the differences are considerably greater for shorter expected delays. For example, when the expected delay length is 5, the bias of the DP is ten times larger than that of the CRDP.

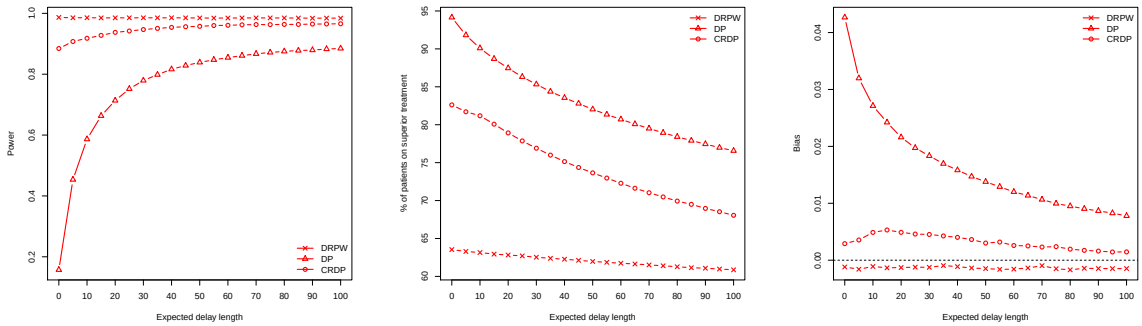


Figure 4.2.7: The changes in power, % of patients on the superior treatment and the average bias of the treatment effect estimator for the (CR)DP and DRPWR as the expected delay length increases, when $n = 75$, $\theta_A = 0.5$ and $\theta_B = 0.1$ (estimated over $100,000$ simulations).

(ii) **Scenario 2:** $\theta_A = \theta_B = 0.5$

The corresponding plots for no treatment difference are shown in Figure 4.2.8.

The first plot illustrates the changes in type I error rates for the (CR)DP and DRPWR as the expected delay increases. After an initial increase for (CR)DP, the type I error rate then seems to decrease slightly (at a slower rate than it did in the fixed delay case). The type I error for the DRPWR seems to remain relatively constant around 0.077.

The % *on sup* and bias values, illustrated in the second and third plots of Figure 4.2.8, behave as one would expect, that is, randomly jumping near 50% and 0, respectively.

4.2.3 Discussion

In the first part of this chapter, we have evaluated how the (CR)DP design performs in two-armed trials with both fixed and random delays. This is an important question in practice which has been raised several times whenever presenting the CRDP design. To summarise, we have found that we gain slightly in terms of power and bias through the delay, so in that sense delay could be viewed as a positive attribute (which

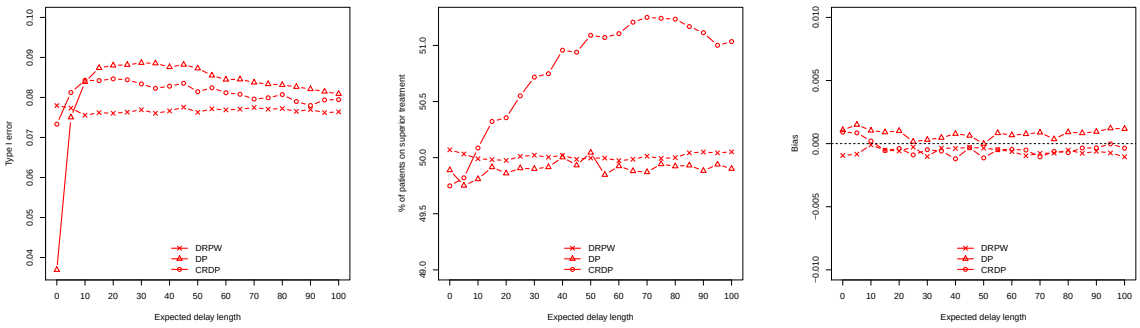


Figure 4.2.8: The changes in type I error, % of patients on the superior treatment and the average bias of the treatment effect estimator for the (CR)DP and DRPWR as the expected delay length increases, when $n = 75$ and $\theta_A = \theta_B = 0.5$ (estimated over $100,000$ simulations).

seems somewhat counter-intuitive), but we lose in terms of patient benefit which is the main advantage of using such response-adaptive designs over alternatives. However, this loss is not overly concerning and for a relatively large delay length of 25, for example, which is one third of the sample size, the percentage of patients on the superior treatment when $\theta_A = 0.5$ and $\theta_B = 0.1$ is still approximately 23% higher for CRDP than the traditional approach of fixed randomisation. Further, when compared to the performance of the most commonly proposed rule for delayed response scenarios (Hardwick *et al.*, 2006), namely the DRPWR, there are still considerable improvements with respect to the patient benefit for (CR)DP.

Therefore, this evaluation has shown that the (CR)DP designs already perform well in trials with delayed responses since they continue to maintain their patient benefit advantages over other designs for a range of (expected) delay lengths. Simulation studies for other RAR designs in the literature, mostly urn models, have similarly shown that they still reduce the expected number of failures and allocate more patients to the better treatment(s) when responses are delayed (Rosenberger and Lachin, 2016, Chapter 12). More specifically, for short to moderate delays, (CR)DP incurs only a slight loss in patient benefit (relative to the no delay case) which again reflects what has been found in the literature for other response-adaptive designs (see the pa-

pers cited in Section 4.1 for examples). Thus, the main message to convey from this simulation study is that (CR)DP is fairly robust to delays, whether fixed or random, with only a small loss in patient benefit for moderate delay lengths.

4.3 Adjusting (CR)DP for Fixed Delays

As we have seen above, (CR)DP already performs quite well in the presence of delayed responses with slight gains in power and a loss in patient benefit as the delay length increases. As such, there is not much room for improvement. However, since the patient benefit is the primary motivation behind using such bandit-based designs (see e.g. Rosenberger and Lachin, 1993; Hardwick, 1995), ideally we want to retain this feature as much as possible. Therefore, the second part of this chapter investigates whether we can minimise this loss by utilising the pipeline information in the adaptations rather than simply ignoring it as we were doing above and as most adaptive designs in the literature do. For the remainder of this chapter, we focus on the setting of a constant arrival rate and a fixed response/follow-up time.

4.3.1 Modifying the Time Horizon of (CR)DP

In Section 4.2, the time horizon used in the MDP formulation of the (CR)DP design was of size $T = n$, i.e. equivalent to the number of patients in the trial. However, when we implement this design with a fixed delay of length d, the state representing the number of unobserved patients remaining in the trial will stay the same for the first d patients because no observations accrue during this stage. Therefore, these patients are simply randomised (with equal probability) between the treatments, giving rise to an initial equal randomisation phase. It is only once we begin to receive observations, i.e. from time $d+1$ onwards, that (CR)DP allocates patients *adaptively*. This suggests that for a trial of size n, it may only be worthwhile to use the (CR)DP algorithm to

allocate patients $d+1$ to n, that is, for $n-d$ of the allocation decisions.

Therefore, we first investigate how the (CR)DP algorithm performs when it is implemented with the adjusted time horizon of $T = n - d$. Not only does this mean that we generate a smaller array of optimal actions, which is computationally quicker and requires less memory, but this will allow us to understand whether there are any non-negligible gains when optimising over the smallest possible time horizon instead.

Figure 4.3.1 illustrates the performance measures of CRDP across all scenarios when using a time horizon (TH) of $n - d$ for a range of delay lengths, which we refer to as the CRDP-TH design (represented by the dashed lines). For comparative purposes, the CRDP when using a time horizon of n is also superimposed onto these plots. In terms of the power (top left plot in Figure 4.3.1), we see that there is very little difference between the two designs, with CRDP-TH lying slightly above CRDP for shorter delay lengths. For the percentage of patients on the superior arm (top right plot in Figure 4.3.1), the differences are more pronounced and, interestingly, CRDP is found to outperform CRDP-TH for all delay lengths (excluding 0 and 75 where both designs are equivalent). A possible reason for this is discussed below. Further, since CRDP-TH results in less imbalance between the two treatment groups than CRDP, the corresponding bias and MSE values are also slightly smaller for CRDP-TH, as illustrated in the bottom two plots of Figure 4.3.1.

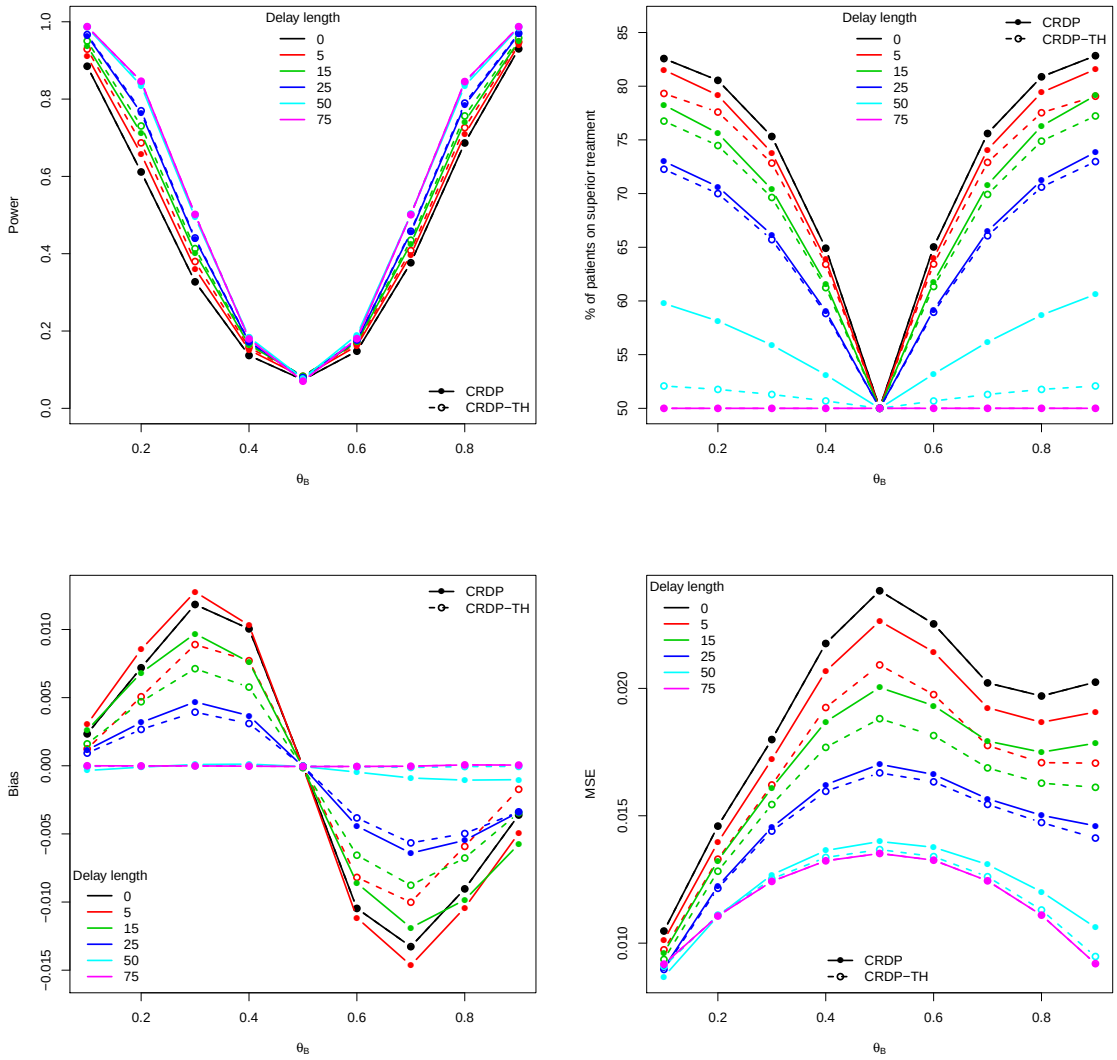


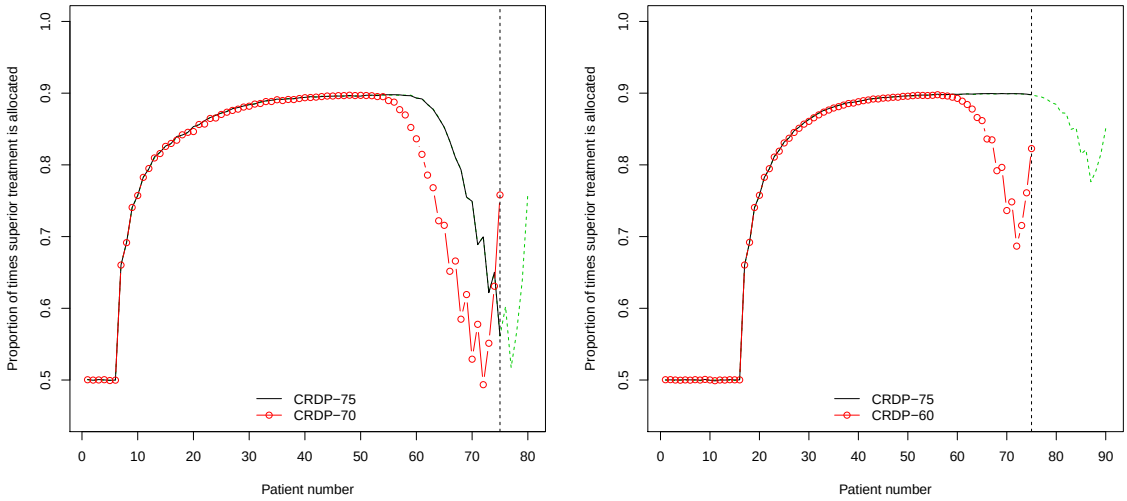
Figure 4.3.1: The changes in power (and type I error), % of patients on the superior treatment, the average bias and MSE of the treatment effect estimator for the CRDP and CRDP-TH designs when $n = 75$, $\theta_A = 0.5$ and $\theta_B \in (0.1, 0.9)$ for different delay lengths (estimated over $1,000,000$ simulations).

We now discuss why CRDP is shown to attain a larger percentage of patients on the superior arm compared to CRDP-TH with the aid of allocation plots in Figures 4.3.2 and 4.3.4. Recall that in Figure 3.4.6 of Chapter 3, we saw that, for the no delay case, the average allocation probability to the superior treatment oscillates markedly for the final 15 patients in order to satisfy the constraint, and thus the CRDP makes

an important number of allocations to the inferior arm towards the end of the trial. However, when the CRDP time horizon T is equal to the trial size n and there is a delay of length d, the final d decisions are redundant. Thus, this final exploration phase, which is illustrated by the dashed green lines in Figures 4.3.2a and 4.3.2b, is ignored. Nevertheless, CRDP will continue to allocate the required number of patients, as specified by the constraint, to the inferior arm. In fact, on average, it will "over-satisfy" the constraint because the number of allocations made to the inferior arm during the initial equal randomisation stage (as a result of the delay) will, on average, exceed those that are no longer being made at the end. This is evident from Figures 4.3.2a and 4.3.2b where it is clear that the proportion of times the superior (inferior) treatment is allocated during the "redundant" phase in green is substantially greater (smaller) than that during the equal randomisation phase.

In contrast, by using the smallest possible time horizon of $n - d$ instead, there will be even more allocations, on average, to the inferior arm because the exploration phase towards the end of the trial is still incorporated (as in the no delay case) (see the red lines in Figures 4.3.2a and 4.3.2b). Hence, we see a smaller percentage of patients on the superior treatment, and thus higher power, for CRDP-TH compared to CRDP with the longer time horizon of 75.

Note that the patient allocation plots in Figure 4.3.2 also illustrate the effect of changing the delay length d on the average allocation probabilities when using CRDP and CRDP-TH. For example, the black line in Figure 4.3.2a shows the average allocation probability to the superior treatment under the CRDP design with time horizon equal to the trial size $T = 75$, a fixed delay of $d = 5$ and a degree of constraining equal to 15% of the total sample size (approximately 12 patients on each arm). We see that near the end of the trial, by around patient number 60, the proportion of times the superior treatment is allocated decreases in order to satisfy the constraint. However, when the delay length is increased to $d = 15$, Figure 4.3.2b shows that there is no



(a) $d = 5$ (b) $d = 15$

Figure 4.3.2: Probability of allocating a patient to the superior treatment when $\theta_A = 0.5$ and $\theta_B = 0.9$ in a trial of size $n = 75$ (estimated over $1,000,000$ simulations). The black and red lines correspond to the CRDP design with time horizons $T = n$ and $T = n - d$, respectively. The dashed green lines illustrate what the remaining d allocations would look like if the CRDP was continued.

longer this decrease near the end of the trial because, in this case, it is likely that the minimum requirement on each arm will have already been fulfilled (owing to the longer delay length, and consequently the longer initial equal randomisation phase). The plots for CRDP-TH (in red) similarly show that as the delay length increases, the need to allocate as many patients to the inferior treatment at the end of the trial is reduced.

Since we expect CRDP-TH to make a greater number of allocations to the inferior arm than CRDP, it is not clear whether the observed differences in Figure 4.3.1 are due to the change in time horizon, or the fact that CRDP-TH is effectively satisfying a stricter constraint. Therefore, to isolate the impact of the time horizon alone on the performance of the design, we remove the constraint and randomisation from the design, and revert back to the original DP design. The corresponding performance measures illustrated in Figure 4.3.3 and allocation plots in Figure 4.3.4 show that DP

and DP-TH behave the same, and thus there are no non-negligible gains to be made from modifying the time horizon.

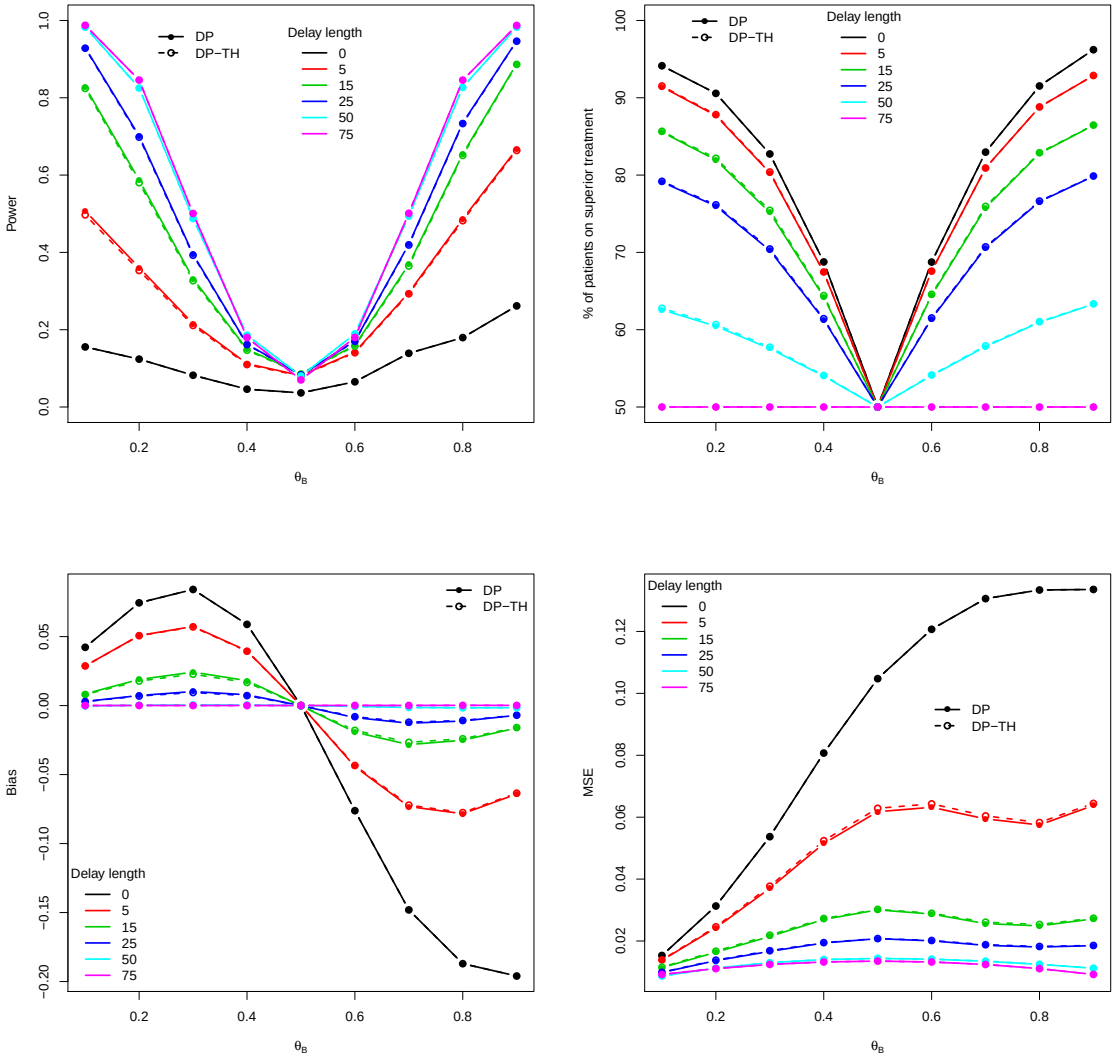
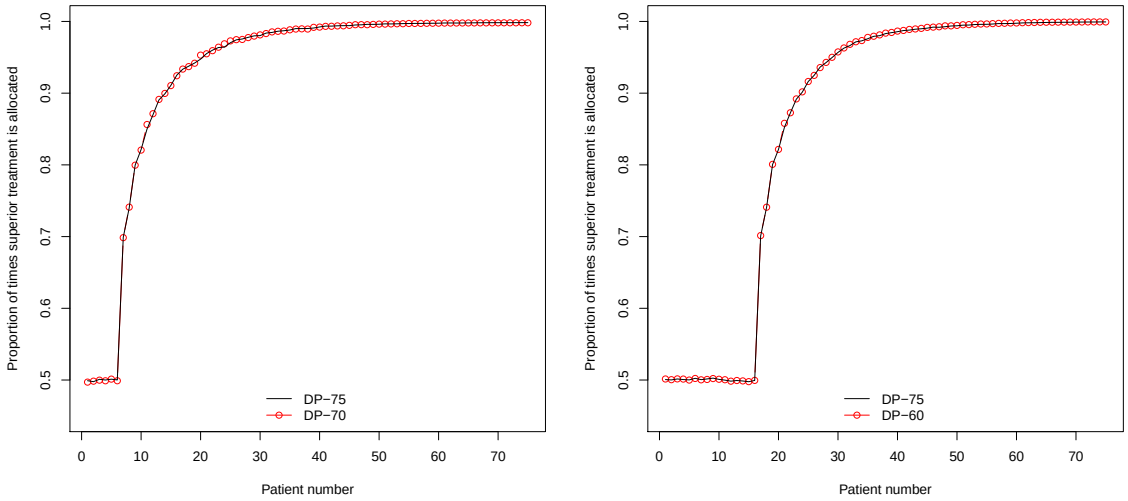


Figure 4.3.3: The changes in power (and type I error), % of patients on the superior treatment, the average bias and MSE of the treatment effect estimator for the DP and DP-TH designs when $n = 75$, $\theta_A = 0.5$ and $\theta_B \in (0.1, 0.9)$ for different delay lengths (estimated over $1,000,000$ simulations).

Next, we propose a more sophisticated way of accounting for fixed delays by incorporating data on the pipeline patients into the MDP model associated with the (CR)DP design.



(a) $d = 5$ (b) $d = 15$

Figure 4.3.4: Probability of allocating a patient to the superior treatment when $\theta_A = 0.5$ and $\theta_B = 0.9$ in a trial of size $n = 75$ (estimated over $1,000,000$ simulations). The black and red lines correspond to the DP design with time horizons $T = n$ and $T = n - d$, respectively.

4.3.2 Incorporating the Pipeline Information into (CR)DP

Model Formulation

Following Chapter 3, we consider a clinical trial where the patients arrive sequentially, one-by-one, and are allocated to either treatment A or B. We model the patient responses as independent Bernoulli random variables giving rise to binary outcomes, either a success or failure. In contrast with the model presented in Chapter 3, however, patient responses are now observed only after a *fixed delay* of length $d \in \mathbb{Z}_{\geq 0}$ (where $d = 0$ recovers the immediate response case). As before, we assign non-informative uniform prior distributions to θ_A and θ_B, the unknown success probabilities of treatments A and B respectively. Since this is a conjugate prior for the Bernoulli likelihood, the resulting posterior distribution follows a Beta distribution with parameters summarising the initial prior information plus the observed information to date. Note that for simplicity of exposition, we specify the model for the

two-arm case, yet the principles used can easily be generalised to multi-arm trials.

Similar to the approach taken in Chick *et al.* (2017), it is useful to think about this problem as being composed of the following three stages. It may also be helpful to refer to the diagram in Figure 5.2.1 of Chapter 5 for the general case.

Stage 1: allocations. This corresponds to the initial equal randomisation stage of the trial in which the first d patients are randomly allocated, with equal probability, to either treatment A or B. No responses are observed during this stage due to the delay of length d, and so these patients enter the pipeline to form, what Eick (1988b) referred to as, an *information bank*.

Stage 2: allocations and observations. During this stage, (i) patients continue to be randomised to a treatment arm and added to the pipeline, and (ii) responses from the pipeline patients are observed and used to update the states. The pipeline remains of fixed length d throughout this stage.

Stage 3: observations. This comprises the end of the trial after all n patients have been allocated. However, updating continues to take place as the remaining pipeline responses are observed.

As in Chapter 3, we formulate the problem as an MDP defined in discrete time in which each time period is indexed by $t \in \{0, 1, \ldots, n\}$, representing both time and the number of patients that have been treated (since at time t, we have treated t patients). Recall that a time period refers to the time between two allocation decisions, which will be of a fixed length throughout the trial since we are assuming that the recruitment rate is constant. The elements of the MDP corresponding to the fixed delay version of the CRDP model, which we will henceforth refer to as FCRDP for convenience, are now defined.

The *state space*, $\boldsymbol{z}_t$, which summarises all of the information available at time t when the current patient is about to be allocated (Eick, 1988b), now includes an additional parameter, $u_{A,t}$, representing the number of pipeline patients on treatment

A, that is, those patients that have been allocated to treatment A but have not yet responded. This additional parameter increases the dimension, and therefore the complexity, of the problem. Although we do not need to explicitly include another parameter in the state space for the number of pipeline patients on treatment B, since it is derived from information we already know ($u_{B,t} = d - u_{A,t}$), we include it here for completeness.

The remaining states, as before, include the number of patients in the trial remaining to be treated, $\widetilde{n} = n - t$, and the number of successes and failures observed on each treatment to date (plus the prior information), denoted by $\widetilde{s}_{A,t},\ \widetilde{f}_{A,t},\ \widetilde{s}_{B,t},\ \widetilde{f}_{B,t}$. Note that we can exclude one of $\widetilde{s}_{A,t},\ \widetilde{f}_{A,t},\ \widetilde{s}_{B,t}$ or $\widetilde{f}_{B,t}$ from the state space because $s_{A,t} + f_{A,t} + s_{B,t} + f_{B,t} + d = t$ but, again, we include it here for clarity of exposition.

Thus, the vector of states can be summarised as

$$\boldsymbol{z}_t = (u_{A,t},\ u_{B,t},\ \widetilde{s}_{A,t},\ \widetilde{f}_{A,t},\ \widetilde{s}_{B,t},\ \widetilde{f}_{B,t},\ \widetilde{n}).$$

At each $t \in \{0, 1, \ldots, T\}$, where $T = n$ is the time horizon (equivalent to the total number of patients within the trial), an action a_t is chosen from the *randomised set of actions*, $\mathcal{A} = \{1, 2\}$, such that $a_t = 1$ denotes allocating the next patient (patient $t + 1$) to treatment A with probability p and treatment B with probability $1 - p$, and $a_t = 2$ denotes allocating patient $t + 1$ to treatment B with probability p and treatment A with probability $1 - p$ (where $0.5 \leq p \leq 1$ for a two-armed trial).

The non-zero *transition probabilities*, $\mathbb{P}(\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_t, a_t)$, representing the evolution of the states from time t to $t + 1$ under action a_t, for each of the different stages of the problem are as follows:

Stage 1.

(i) When $a_t = 1$:

$$\boldsymbol{z}_{t+1} = \begin{cases} (u_{A,t}+1,\ u_{B,t},\ \widetilde{s}_{A,t},\ \widetilde{f}_{A,t},\ \widetilde{s}_{B,t},\ \widetilde{f}_{B,t},\ \widetilde{n}-1) & \text{w.p.}\quad p, \\ (u_{A,t},\ u_{B,t}+1,\ \widetilde{s}_{A,t},\ \widetilde{f}_{A,t},\ \widetilde{s}_{B,t},\ \widetilde{f}_{B,t},\ \widetilde{n}-1) & \text{w.p.}\quad 1-p. \end{cases}$$

(ii) When $a_t = 2$:

$$\boldsymbol{z}_{t+1} = \begin{cases} (u_{A,t}+1,\ u_{B,t},\ \widetilde{s}_{A,t},\ \widetilde{f}_{A,t},\ \widetilde{s}_{B,t},\ \widetilde{f}_{B,t},\ \widetilde{n}-1) & \text{w.p.}\quad 1-p, \\ (u_{A,t},\ u_{B,t}+1,\ \widetilde{s}_{A,t},\ \widetilde{f}_{A,t},\ \widetilde{s}_{B,t},\ \widetilde{f}_{B,t},\ \widetilde{n}-1) & \text{w.p.}\quad p. \end{cases}$$

Stage 2.

(i) When $a_t = 1$:

$$\boldsymbol{z}_{t+1} = \begin{cases} (u_{A,t},\ u_{B,t},\ \widetilde{s}_{A,t}+1,\ \widetilde{f}_{A,t},\ \widetilde{s}_{B,t},\ \widetilde{f}_{B,t},\ \widetilde{n}-1) & \text{w.p. } p\cdot\frac{\widetilde{s}_{A,t}}{\widetilde{s}_{A,t}+\widetilde{f}_{A,t}}\cdot\frac{u_{A,t}}{d}, \\ (u_{A,t},\ u_{B,t},\ \widetilde{s}_{A,t},\ \widetilde{f}_{A,t}+1,\ \widetilde{s}_{B,t},\ \widetilde{f}_{B,t},\ \widetilde{n}-1) & \text{w.p. } p\cdot\frac{\widetilde{f}_{A,t}}{\widetilde{s}_{A,t}+\widetilde{f}_{A,t}}\cdot\frac{u_{A,t}}{d}, \\ (u_{A,t}+1,\ u_{B,t}-1,\ \widetilde{s}_{A,t},\ \widetilde{f}_{A,t},\ \widetilde{s}_{B,t}+1,\ \widetilde{f}_{B,t},\ \widetilde{n}-1) & \text{w.p. } p\cdot\frac{\widetilde{s}_{B,t}}{\widetilde{s}_{B,t}+\widetilde{f}_{B,t}}\cdot\frac{u_{B,t}}{d}, \\ (u_{A,t}+1,\ u_{B,t}-1,\ \widetilde{s}_{A,t},\ \widetilde{f}_{A,t},\ \widetilde{s}_{B,t},\ \widetilde{f}_{B,t}+1,\ \widetilde{n}-1) & \text{w.p. } p\cdot\frac{\widetilde{f}_{B,t}}{\widetilde{s}_{B,t}+\widetilde{f}_{B,t}}\cdot\frac{u_{B,t}}{d}, \\ (u_{A,t}-1,\ u_{B,t}+1,\ \widetilde{s}_{A,t}+1,\ \widetilde{f}_{A,t},\ \widetilde{s}_{B,t},\ \widetilde{f}_{B,t},\ \widetilde{n}-1) & \text{w.p. } (1-p)\cdot\frac{\widetilde{s}_{A,t}}{\widetilde{s}_{A,t}+\widetilde{f}_{A,t}}\cdot\frac{u_{A,t}}{d}, \\ (u_{A,t}-1,\ u_{B,t}+1,\ \widetilde{s}_{A,t},\ \widetilde{f}_{A,t}+1,\ \widetilde{s}_{B,t},\ \widetilde{f}_{B,t},\ \widetilde{n}-1) & \text{w.p. } (1-p)\cdot\frac{\widetilde{f}_{A,t}}{\widetilde{s}_{A,t}+\widetilde{f}_{A,t}}\cdot\frac{u_{A,t}}{d}, \\ (u_{A,t},\ u_{B,t},\ \widetilde{s}_{A,t},\ \widetilde{f}_{A,t},\ \widetilde{s}_{B,t}+1,\ \widetilde{f}_{B,t},\ \widetilde{n}-1) & \text{w.p. } (1-p)\cdot\frac{\widetilde{s}_{B,t}}{\widetilde{s}_{B,t}+\widetilde{f}_{B,t}}\cdot\frac{u_{B,t}}{d}, \\ (u_{A,t},\ u_{B,t},\ \widetilde{s}_{A,t},\ \widetilde{f}_{A,t},\ \widetilde{s}_{B,t},\ \widetilde{f}_{B,t}+1,\ \widetilde{n}-1) & \text{w.p. } (1-p)\cdot\frac{\widetilde{f}_{B,t}}{\widetilde{s}_{B,t}+\widetilde{f}_{B,t}}\cdot\frac{u_{B,t}}{d}. \end{cases}$$

Note that $u_{A,t}$ remains the same in the first two transitions because the patient being allocated and the patient responding are both on treatment A, and similarly for treatment B in the latter two transitions.

(ii) When $a_t = 2$:

As above but with the probabilities p and $1-p$ exchanged.

Finally, the expected one-period (or immediate) *reward* after transitioning from state $\boldsymbol{z}_t$ to $\boldsymbol{z}_{t+1}$ under action a_t is given by $\mathcal{R}^{a_t}(\boldsymbol{z}_t)$. Here, the one-period reward corresponds to the first-in-pipeline patient who was allocated d time-steps ago (at time $t-d$), not the patient that was previously allocated (at time t), as in the CRDP model described in Chapter 3. Note that the reward depends on the objective function of interest, which in this case is to maximise the expected total number of patient successes in the trial, whereby we obtain a reward of 1 for a success and 0 for a failure. Since we do not know which treatment the first-in-pipeline patient received, we instead assume that they received treatment A with probability $\frac{u_{A,t}}{d}$ and treatment B with probability $\frac{u_{B,t}}{d}$. To keep track of *which* treatment each pipeline patient received, rather than just the number of pipeline patients on each treatment, a parameter would need to be introduced into the state space for every pipeline patient. Consequently, the problem would quickly become computationally infeasible as d increased.

Stage 1. For $t \in \{0, \ldots, d\}$, no reward accrues since there are no patient responses observed during this time.

Stage 2. For $t \in \{d+1, \ldots, n\}$,

(i) When $a_t = 1$:

$$\begin{aligned}\mathcal{R}^1(\boldsymbol{z}_t) = p \cdot \left\{ \frac{\widetilde{s}_{A,t}}{\widetilde{s}_{A,t} + \widetilde{f}_{A,t}} \cdot \frac{u_{A,t}}{d} + \frac{\widetilde{s}_{B,t}}{\widetilde{s}_{B,t} + \widetilde{f}_{B,t}} \cdot \frac{u_{B,t}}{d} \right\} + \\ (1-p) \cdot \left\{ \frac{\widetilde{s}_{A,t}}{\widetilde{s}_{A,t} + \widetilde{f}_{A,t}} \cdot \frac{u_{A,t}}{d} + \frac{\widetilde{s}_{B,t}}{\widetilde{s}_{B,t} + \widetilde{f}_{B,t}} \cdot \frac{u_{B,t}}{d} \right\} \\ = \frac{\widetilde{s}_{A,t}}{\widetilde{s}_{A,t} + \widetilde{f}_{A,t}} \cdot \frac{u_{A,t}}{d} + \frac{\widetilde{s}_{B,t}}{\widetilde{s}_{B,t} + \widetilde{f}_{B,t}} \cdot \frac{u_{B,t}}{d}.\end{aligned}$$

(ii) When $a_t = 2$:

$$\mathcal{R}^2(\boldsymbol{z}_t) = (1-p)\cdot\left\{\frac{\widetilde{s}_{A,t}}{\widetilde{s}_{A,t}+\widetilde{f}_{A,t}}\cdot\frac{u_{A,t}}{d}+\frac{\widetilde{s}_{B,t}}{\widetilde{s}_{B,t}+\widetilde{f}_{B,t}}\cdot\frac{u_{B,t}}{d}\right\}+$$
$$p\cdot\left\{\frac{\widetilde{s}_{A,t}}{\widetilde{s}_{A,t}+\widetilde{f}_{A,t}}\cdot\frac{u_{A,t}}{d}+\frac{\widetilde{s}_{B,t}}{\widetilde{s}_{B,t}+\widetilde{f}_{B,t}}\cdot\frac{u_{B,t}}{d}\right\}$$
$$=\frac{\widetilde{s}_{A,t}}{\widetilde{s}_{A,t}+\widetilde{f}_{A,t}}\cdot\frac{u_{A,t}}{d}+\frac{\widetilde{s}_{B,t}}{\widetilde{s}_{B,t}+\widetilde{f}_{B,t}}\cdot\frac{u_{B,t}}{d}.$$

The rewards are equivalent in this case because the response from the first-in-pipeline patient will be the same regardless of which action is taken for the subsequent patient.

Stage 3. When $\widetilde{n} = 0$,

$$\mathcal{R}(\boldsymbol{z}_t) = u_{A,t}\cdot\frac{\widetilde{s}_{A,t}}{\widetilde{s}_{A,t}+\widetilde{f}_{A,t}}+u_{B,t}\cdot\frac{\widetilde{s}_{B,t}}{\widetilde{s}_{B,t}+\widetilde{f}_{B,t}}+\epsilon,$$

where

$$\epsilon=\begin{cases}-n, & \text{if } u_{A,t}+s_{A,t}+f_{A,t}<\ell \text{ or } u_{B,t}+s_{B,t}+f_{B,t}<\ell,\\ 0, & \text{otherwise,}\end{cases}$$

which serves as a penalty to avoid choosing states that give rise to fewer than the desired number of allocations on each arm, ℓ, and $u_{A,t}\cdot\frac{\widetilde{s}_{A,t}}{\widetilde{s}_{A,t}+\widetilde{f}_{A,t}}+u_{B,t}\cdot\frac{\widetilde{s}_{B,t}}{\widetilde{s}_{B,t}+\widetilde{f}_{B,t}}$ is the expected predicted number of successes for the pipeline patients.

The Bellman equation, defined in (2.2.2) of Chapter 2, immediately follows as the expected one-period rewards, defined above, plus the expected (undiscounted) future rewards. As in Chapter 3, the objective is to maximise the expected total reward over the entire time horizon. This is solved exactly using backward induction (Sections 2.2.3 and 3.6.1) to obtain the optimal treatment allocation policy of the FCRDP design. The performance of this design is evaluated in the following section.

Simulation Results

We implement the proposed FCRDP design, along with the analogous version for the DP (which we refer to as the FDP), via simulation. For consistency, we illustrate the performance of F(CR)DP in a two-armed trial with 75 patients and a fixed delay of length d for the same scenarios that have been considered throughout this chapter. Results for other sample sizes and delay lengths showed similar patterns and the conclusions do not change.

First, we focus on how, if at all, the performance measures of FCRDP, indicated by the dashed lines in Figure 4.3.5, vary relative to CRDP, represented by the solid lines, for a selection of fixed delay lengths. Analogous results for DP versus FDP are presented in Figure 4.5.5. The top left plot in Figure 4.3.5 shows that the power of FCRDP is slightly smaller than that of CRDP, particularly for smaller delay lengths (see delay lengths of 5 and 10, for example). This is more obvious for the FDP design; see Figure 4.5.5. An alternative representation is provided in Figure 4.4.1 where the differences in power between FDP (blue line) and DP (green line) are greater than the corresponding differences between FCRDP (red line) and CRDP (black line).

The *% on sup*, displayed in the top right plot of Figure 4.3.5, is larger for FCRDP than CRDP for all delay lengths > 0, excluding 75 where both designs are equivalent (see also Figure 4.4.1). For example, when $\theta_B = 0.1$ and $d = 25$ (see blue line), an additional 1% of patients, on average, will receive the superior treatment. The larger *% on sup* for the proposed design is also reflected in the corresponding results for FDP (see Figures 4.4.1 and 4.5.5). Such gains are extremely desirable in practice (Rosenberger and Hu, 2004), particularly for trials involving life-threatening diseases.

The changes in the bias and MSE values are illustrated in the bottom two plots of Figure 4.3.5. The bias values following the FCRDP design appear to be slightly deflated relative to those for CRDP, at least for the smaller delay lengths, whereas the MSE values are slightly inflated. Since the scale of this plot is extremely small,

these observed differences are negligible.

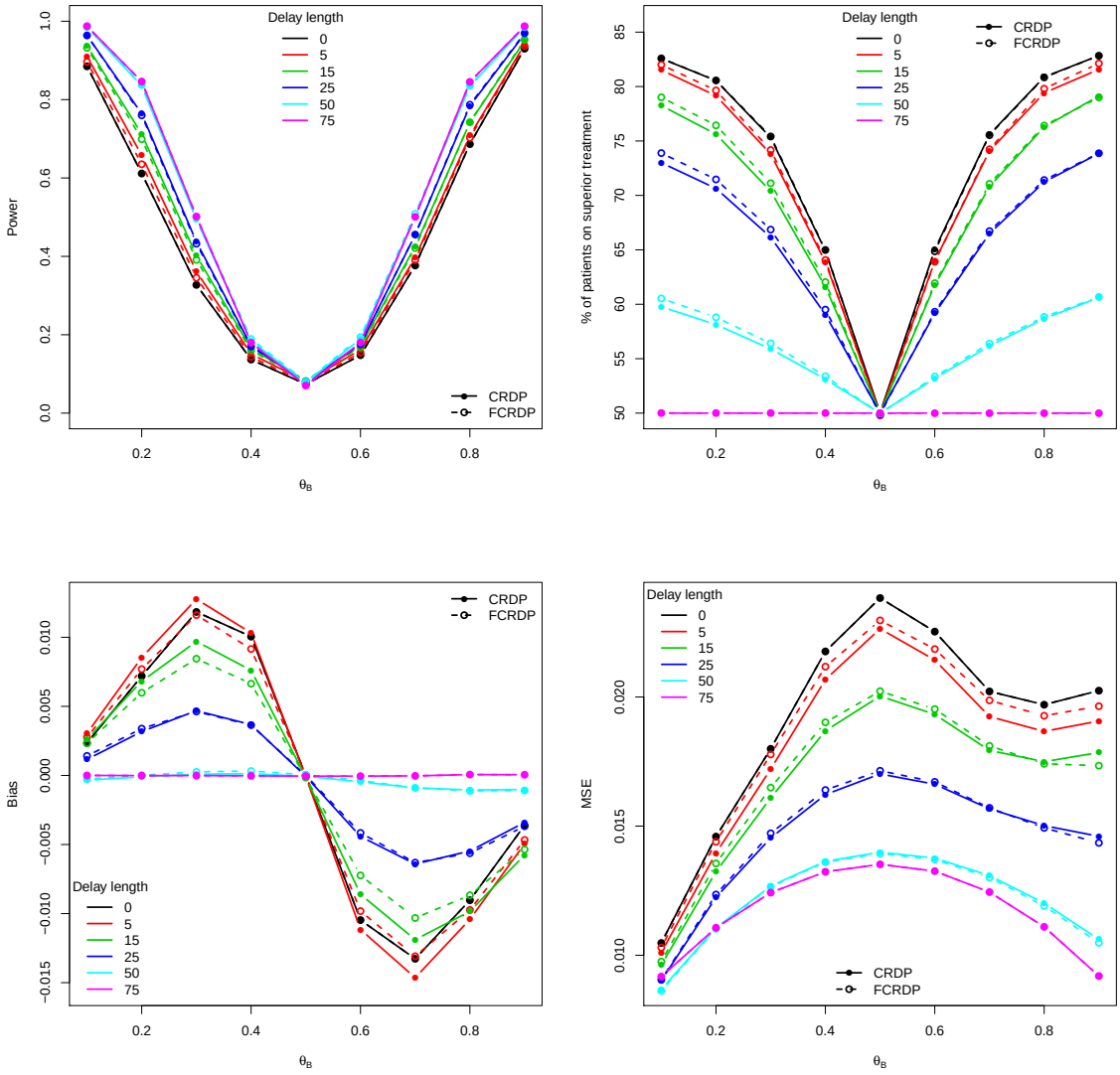


Figure 4.3.5: The changes in power (and type I error), % of patients on the superior treatment, the average bias and MSE of the treatment effect estimator for the CRDP and FCRDP designs when $n = 75$, $\theta_A = 0.5$ and $\theta_B \in (0.1, 0.9)$ for different *fixed* delay lengths (estimated over $1,000,000$ simulations).

Effect of Random Delays on F(CR)DP

Next, in the same way as we did for (CR)DP in Section 4.2.2, we investigate how the F(CR)DP design performs when the delays in response are instead random. The

results are presented in Figure 4.3.6 (see dashed lines), alongside those for CRDP (solid lines), and illustrated for the case of geometric response times with expected delay lengths as indicated.

We observe that, relative to CRDP, FCRDP continues to consistently improve the percentage of patients allocated to the superior arm, even when the delay is random. For example, in a trial where $n = 75$, $\theta_B = 0.1$ and the delay is expected to be of length 25 (see blue line), an additional 2% of patients, on average, will receive the superior treatment when implementing FCRDP over CRDP. Note that these differences tend to increase with the expected delay length. Similarly to when the delays were fixed, the corresponding changes in power, bias and MSE are minimal.

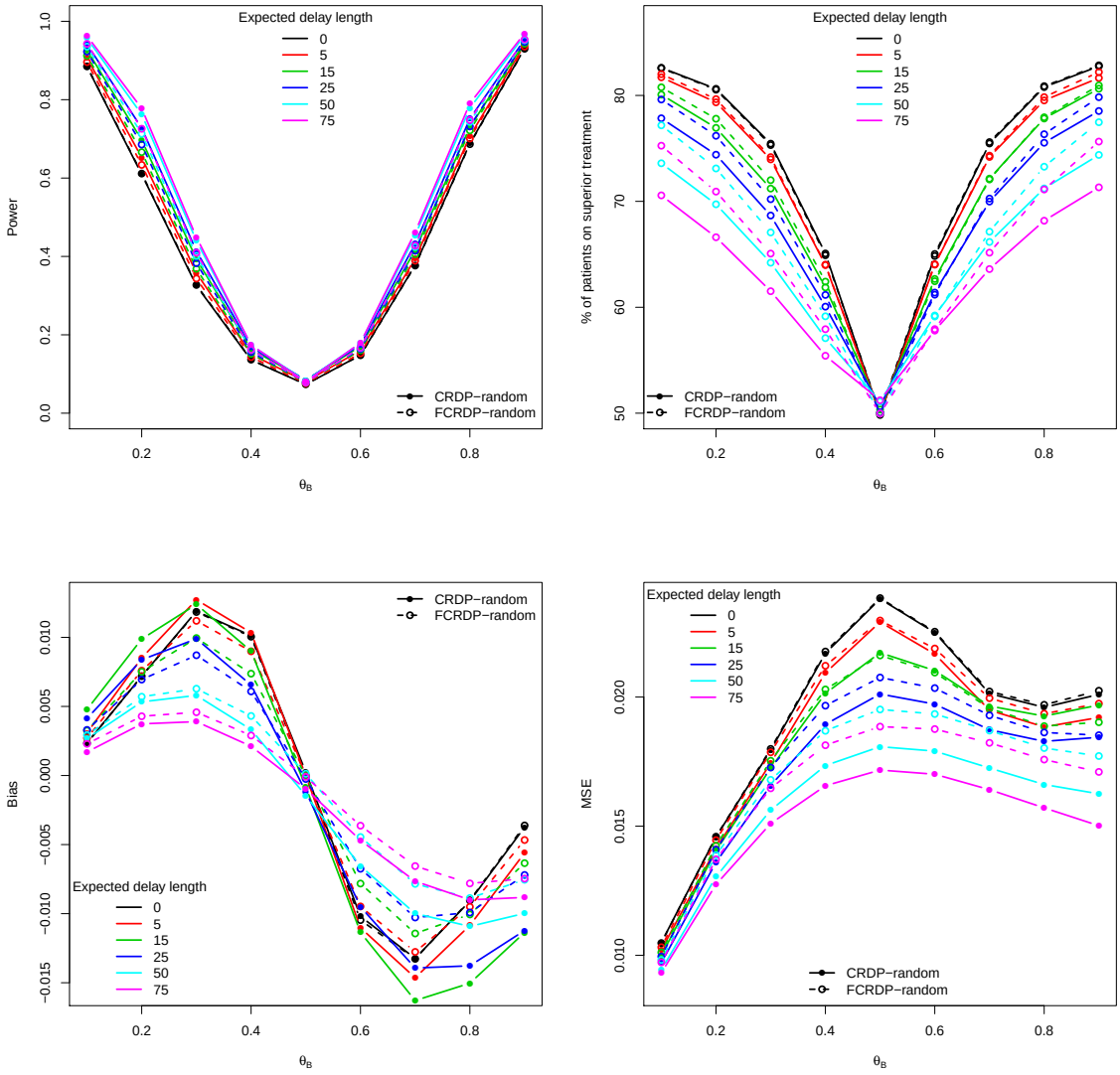


Figure 4.3.6: The changes in power (and type I error), % of patients on the superior treatment, the average bias and MSE of the treatment effect estimator for the CRDP and FCRDP designs when $n = 75$, $\theta_A = 0.5$ and $\theta_B \in (0.1, 0.9)$ for different *expected* delay lengths (estimated over $1,000,000$ simulations).

4.4 Summary

The purpose of this chapter was to gain insight into how the CRDP design, proposed in Chapter 3, behaves when responses are observed after a delay. In Section 4.2, we

demonstrated that the benefits of the CRDP, namely, a large number of patients on the superior treatment, are slightly reduced when there is a delay in observing the response. However, overall, CRDP was shown to be fairly robust to delays.

Section 4.3 provided two suggestions of how to account for a fixed delay with a view to ameliorate the associated loss in patient benefit. The first, in Section 4.3.1, was a naïve approach (referred to as CRDP-TH) which involved altering the time horizon of the corresponding MDP. Although this was actually shown to reduce the patient benefit further for CRDP, some important issues, such as the underlying interaction between the delay and constraint, were raised. Furthermore, when removing the constraint and randomisation, modifying the time horizon had practically no effect on the performance measures. An interesting topic for further research, which will help make the results more interpretable, is how to appropriately adjust the degree of constraining within the CRDP formulation to ensure that it remains the same for each design. One way to achieve this for the CRDP-TH design is to subtract the expected number of patients that are on the inferior arm during the final d allocations from the current degree of constraining. These translate to responses which the current design is "blind" to because they only become available after all allocations have been made, hence why the constraint ends up being stricter than desired.

The second approach, referred to as FCRDP and described in Section 4.3.2, involved formally extending the associated MDP to incorporate information on the pipeline patients. Therefore, rather than only using the responses once they become available, as in the previous designs, patients are still able to contribute valuable information even whilst in the pipeline. F(CR)DP was shown to consistently outperform (CR)DP, in terms of patient benefit, for all delay lengths with minimal impact on the associated power, bias and MSE.

Although the increase in patient benefit was found to be small, at least for the simulation scenarios considered here, this may be critically important in trials

where treatment failures are particularly undesirable, or even fatal (as in some life-threatening diseases), (Hu and Rosenberger, 2006, Chapter 8). Therefore, the value of such results should not be underestimated, especially since we are obtaining this worthwhile improvement at no extra cost simply by implementing a different algorithm to allocate the patients.

The key information gleaned from this chapter is summarised in Figure 4.4.1, which presents the results of (CR)DP and F(CR)DP, alongside those of standard fixed randomisation and DRPWR, all on the same plot for a specific scenario over the entire range of delay lengths.

The F(CR)DP model is formulated assuming that patients arrive sequentially and have a fixed response time, thus giving rise to a fixed number of patients in the pipeline. This is somewhat restrictive in a clinical trial setting where responses of different patients often arrive randomly. Consequently, we then evaluated how the F(CR)DP design behaves when the delay in response is instead random, and compared it to the performance of (CR)DP with random delay. The results showed that the improvements in patient benefit, achieved by F(CR)DP, persist even when implemented in a random delay setting. When implementing the random delay setting, we assumed that the delay in response is independent of the treatment. However, in practice, different treatments are likely to give rise to different delay lengths. Therefore, further investigation is required to consider the performance of (CR)DP and F(CR)DP when the response distributions vary for each treatment.

In the following chapter, we explore whether generalising the F(CR)DP model to allow for a random number of patients in the pipeline may enhance the performance of the design for the random delay setting even further.

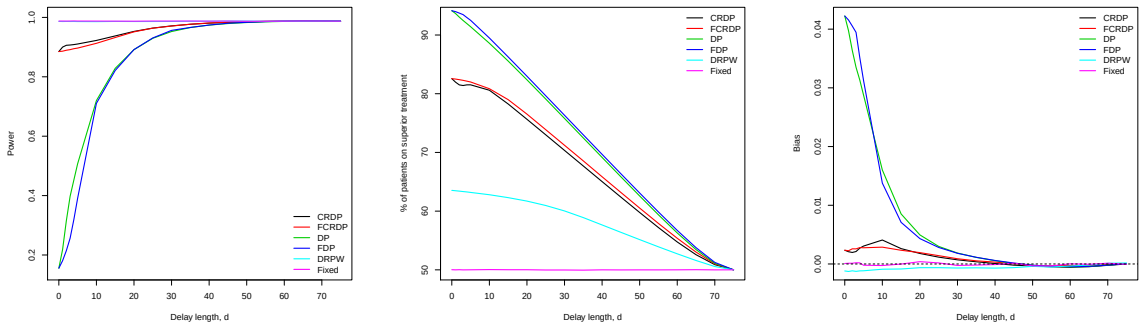


Figure 4.4.1: The changes in power, % of patients on the superior treatment and the average bias of the treatment effect estimator for (CR)DP, F(CR)DP, DRPWR and fixed randomisation as the fixed delay length increases, when $n = 75, \theta_A = 0.1$ and $\theta_B = 0.5$ (estimated over $1,000,000$ simulations).

4.5 Appendix

4.5.1 Performance Measures for DP with Fixed Delay

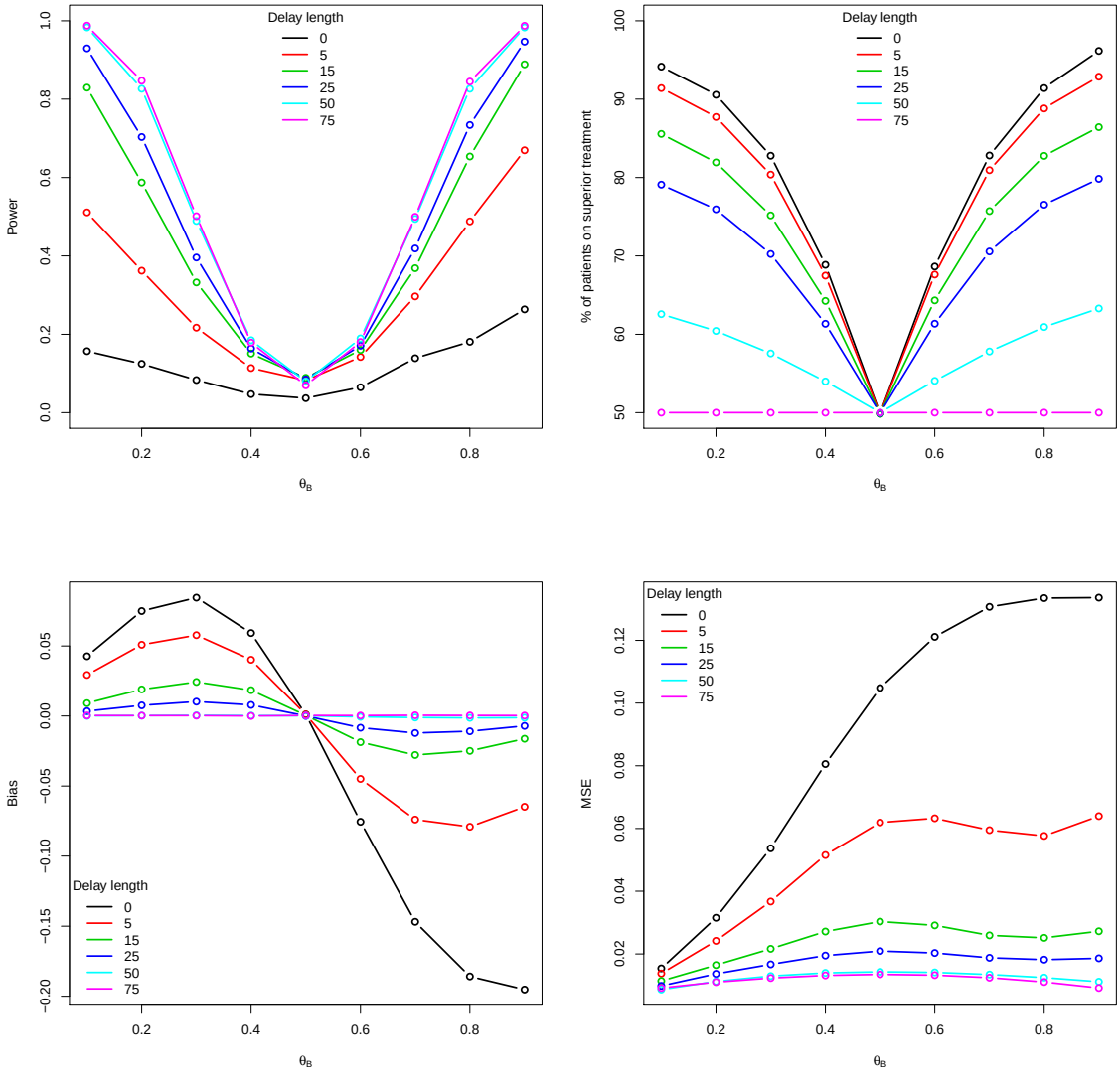


Figure 4.5.1: The changes in power (and type I error), % of patients on the superior treatment, the average bias and MSE of the treatment effect estimator for the DP design when $n = 75$, $\theta_A = 0.5$ and $\theta_B \in (0.1, 0.9)$ for different fixed delay lengths (estimated over $100,000$ simulations).

4.5.2 Performance Measures for DP with Random Delay

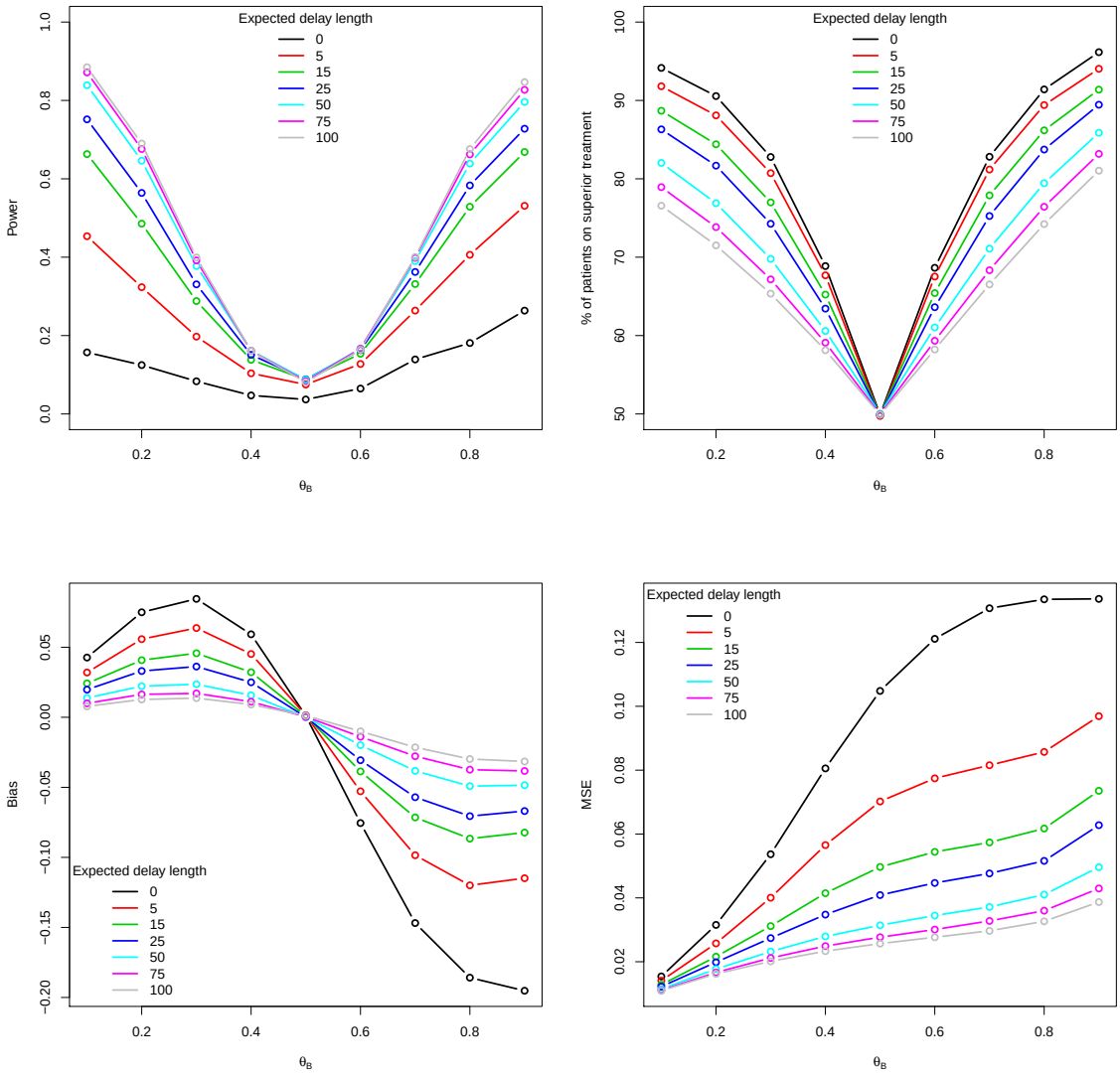


Figure 4.5.2: The changes in power (and type I error), % of patients on the superior treatment, the average bias and MSE of the treatment effect estimator for the DP design when $n = 75$, $\theta_A = 0.5$ and $\theta_B \in (0.1, 0.9)$ for different expected delay lengths (estimated over 100,000 simulations).

4.5.3 Performance Measures for DRPWR with Fixed Delay

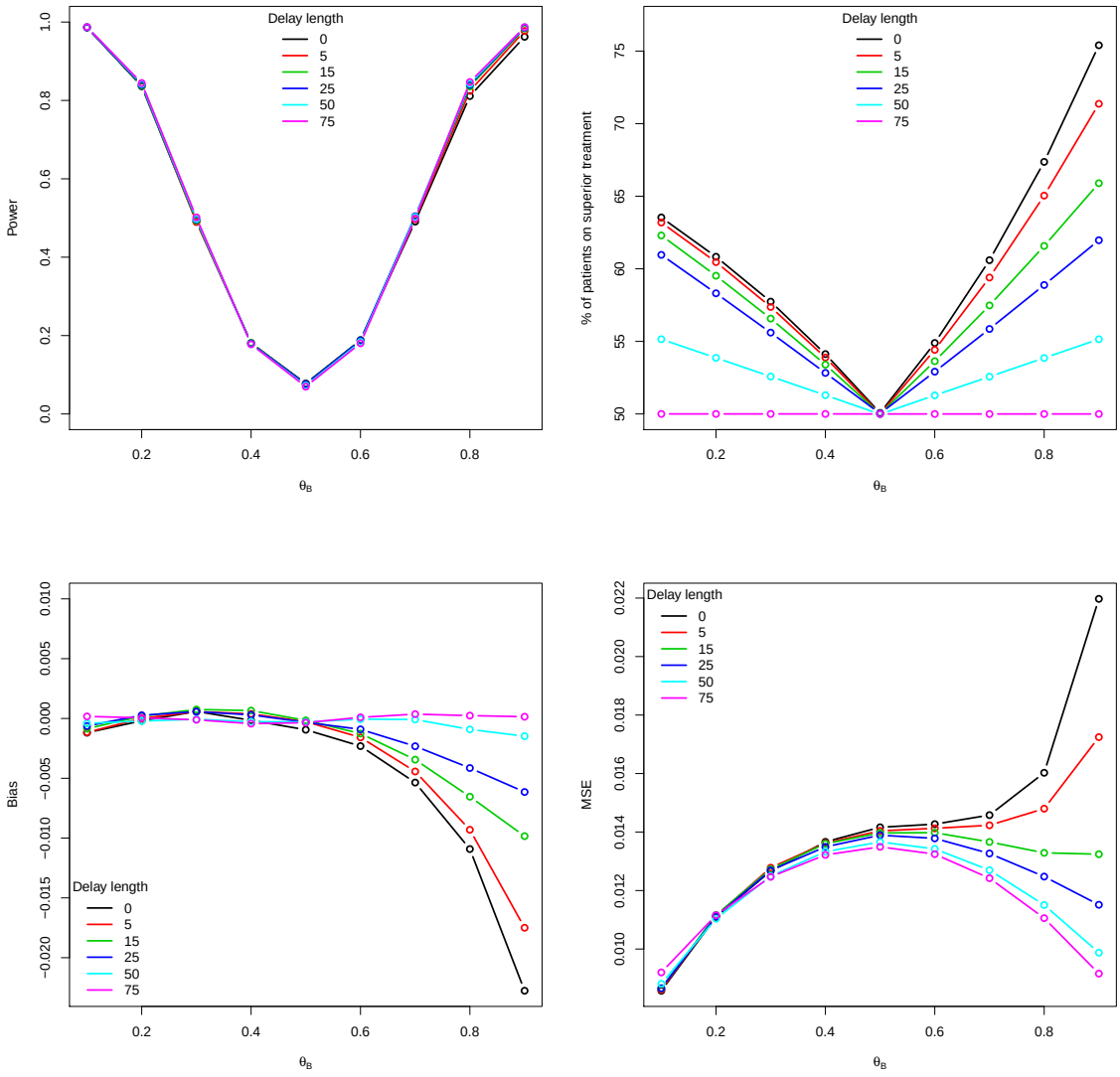


Figure 4.5.3: The changes in power (and type I error), % of patients on the superior treatment, the average bias and MSE of the treatment effect estimator for the DRPWR when $n = 75$, $\theta_A = 0.5$ and $\theta_B \in (0.1, 0.9)$ for different fixed delay lengths (estimated over $100,000$ simulations).

4.5.4 Performance Measures for DRPWR with Random Delay

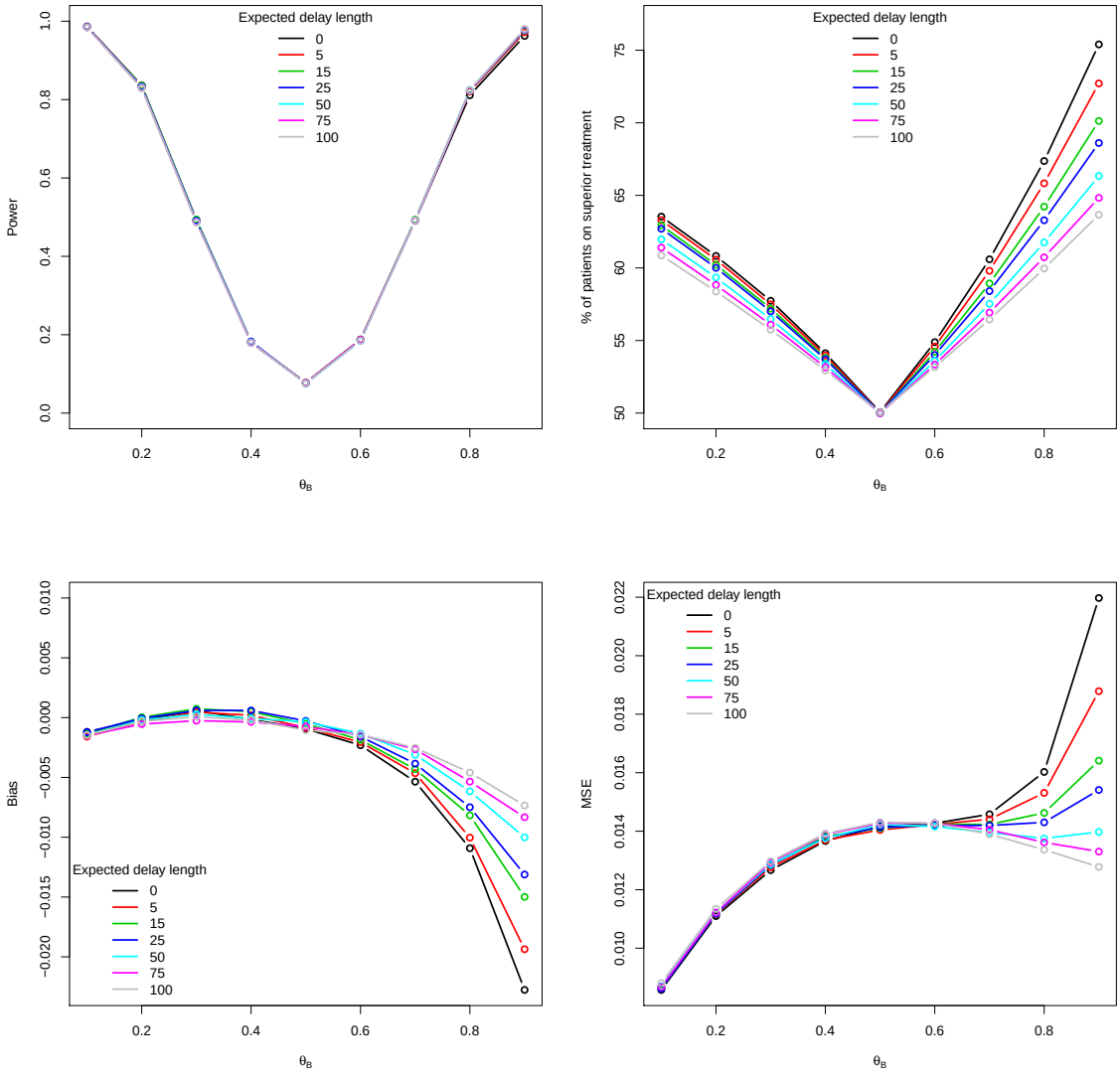


Figure 4.5.4: The changes in power (and type I error), % of patients on the superior treatment, the average bias and MSE of the treatment effect estimator for the DR-PWR when $n = 75$, $\theta_A = 0.5$ and $\theta_B \in (0.1, 0.9)$ for different expected delay lengths (estimated over $100,000$ simulations).

4.5.5 Performance Measures for DP vs. FDP

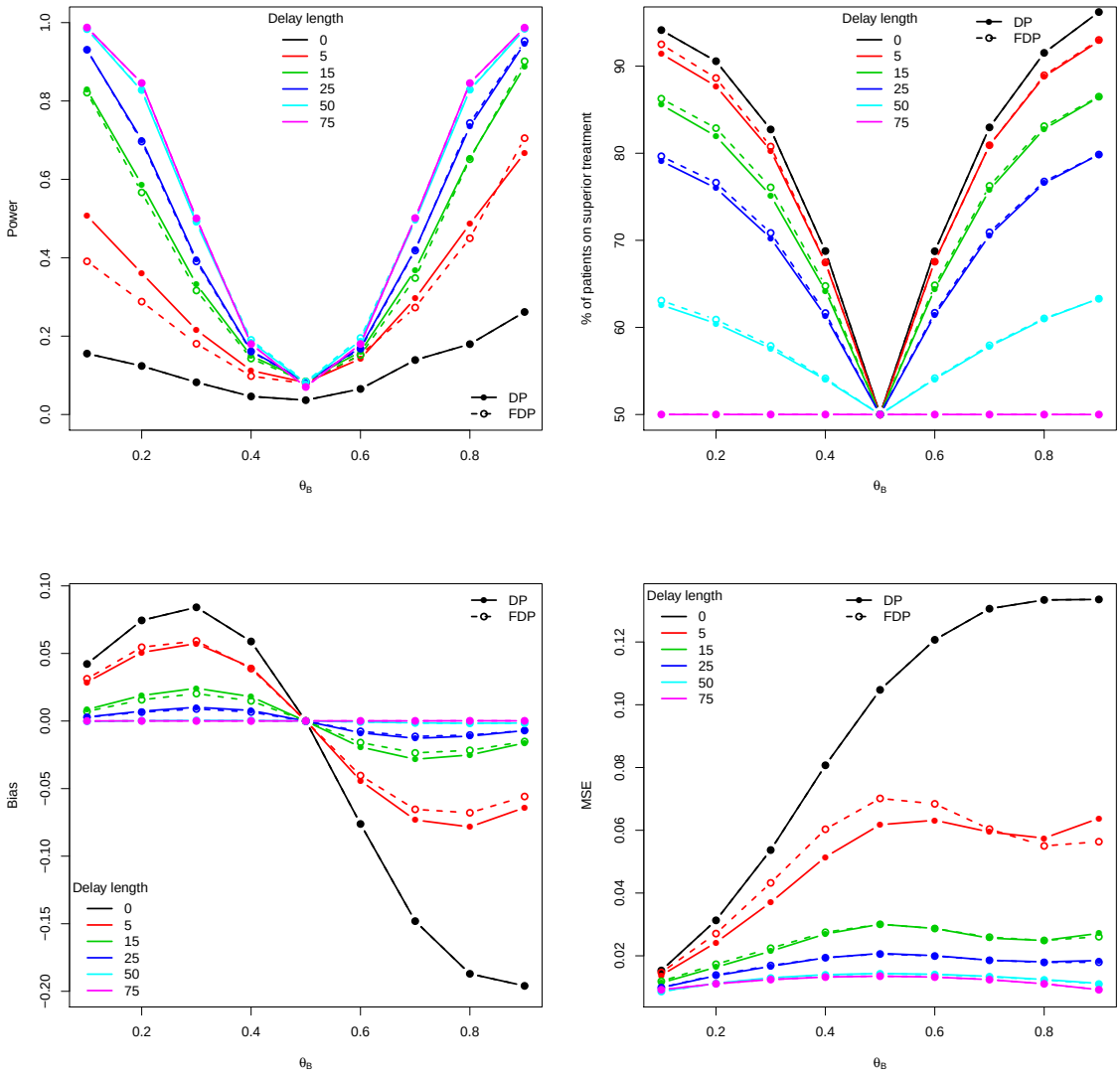


Figure 4.5.5: The changes in power (and type I error), % of patients on the superior treatment, the average bias and MSE of the treatment effect estimator for the DP and FDP designs when $n = 75$, $\theta_A = 0.5$ and $\theta_B \in (0.1, 0.9)$ for different *fixed* delay lengths (estimated over $1,000,000$ simulations).

Chapter 5

Extension to Random Arrivals

5.1 Introduction

This chapter generalises the F(CR)DP model, proposed in Chapter 4, to the more realistic setting of when there is a random, rather than fixed, number of patients in the pipeline. This encompasses a wide variety of trial contexts, including those with a: (i) constant inter-arrival time and fixed time to response, in which case the model simplifies to F(CR)DP, (ii) constant inter-arrival time and random time to response, (iii) random inter-arrival time and fixed time to response, or (iv) random inter-arrival time and random time to response.

Random arrivals of patients are most representative of clinical trial practice, especially in the rare disease setting where there is unlikely to be a constant influx of patients. Moreover, since we are concerned with a binary outcome, interest is in whether the response has been observed by a fixed follow-up time after treatment. Thus, we present and illustrate the methodology proposed in this chapter for the most pertinent trial setting of random arrivals with a fixed follow-up time. Nonetheless, it can still be easily applied to all of the aforementioned contexts. This is in contrast to many response-adaptive designs in the literature which, as Ahuja and Birge (2016)

point out as a limitation, "may not be fully applicable in all trial contexts ... e.g. when time to observation of the primary endpoint is a random variable".

Even those methods that have been developed specifically with delay in mind are not designed for all settings (i)–(iv). For example, although the (group sequential) designs proposed by Hampson and Jennison (2013) do allow for a random number of patients in the pipeline, they rely on there being a fixed delay in response and cannot accommodate stochastic delays, thus they cannot be applied to settings (ii) or (iv). It has already been mentioned in Chapter 4 that the model by Chick *et al.* (2017) is proposed under the assumption of a constant arrival rate and fixed time to response, thus only applies to setting (i). Further, although the bandit-based designs by Hardwick *et al.* (2006) are proposed for a binary trial with random arrivals and a random response time, their designs hinge on the assumption of Poisson arrivals and exponential response times. This is not the case for our design which is not limited/restricted to a specific arrival or response distribution.

5.2 Model Formulation

Recall that the MDP formulation consists of a state space, a set of actions, transition probabilities and rewards. Each of these components will now be defined for the random delay version of the CRDP model, which we will henceforth refer to as RCRDP for convenience.

Although we present this model in the two-armed setting for simplicity, it is important to note that the principles used extend to the multi-armed setting (as with the CRDP and FCRDP models defined in Chapters 3 and 4, respectively).

5.2.1 Decision Epochs and State Space

In general, decision epochs refer to the points in time at which decisions are made based upon the current state of the system (Puterman, 2014). In the current context, we define the decision epochs $t \in [0, n-1]$ to be the time when patient $t+1$ arrives and is allocated to a treatment (where arrival time and allocation time are assumed to be analogous). For example, decision epoch $t = 0$ corresponds to the arrival and allocation of the first patient which defines the start of the trial. Refer to the timeline in Figure 5.2.1 which illustrates the decision epochs as black crosses. Note that the patient allocated at each decision epoch will immediately enter the appropriate pipeline.

Since n is finite, we have a finite horizon problem in which no decisions are made after the final n^{th} patient is allocated at decision epoch $n-1$ (this is often referred to as an $n-1$ period problem). However, for completeness and the purpose of evaluating the final state of the system, we also include an "imaginary" epoch n in the model (represented by the dashed vertical line in Figure 5.2.1) at which point the observation of the last patient (and any other remaining pipeline observations) will be available.

The state space for this problem now includes two additional parameters, $u_{A,t}$ and $u_{B,t}$, which represent the number of pipeline patients on treatments A and B, respectively, just *before* patient $t+1$ is allocated at decision epoch t. Therefore, $u_{A,t}$ and $u_{B,t}$ can take values from 0 to t since it is now possible for pipeline A or pipeline B to include all of the allocated patients. Thus, the state vector, which summarises all of the information available just *before* decision epoch t, is given by

$$\boldsymbol{z}_t = (u_{A,t}, u_{B,t}, s_{A,t}, f_{A,t}, s_{B,t}, f_{B,t}, \widetilde{n}) \text{ if } 0 \leq t \leq n-1, \tag{5.2.1}$$

where $s_{A,t}, f_{A,t}, s_{B,t}, f_{B,t}$ are the total numbers of successes and failures observed on each treatment to date, and $\widetilde{n} = n - t$ is the number of patients in the trial remaining

to be treated. Note that the addition of epoch n at the end (in which no actual decision is made) leads to the definition of $\boldsymbol{z}_n = (0, 0, s_{A,n}, f_{A,n}, s_{B,n}, f_{B,n}, 0)$ (where $s_{A,n} + f_{A,n} + s_{B,n} + f_{B,n} = n$).

5.2.2 Action Set

The set of randomised actions, $\mathcal{A} = \{1, 2\}$, remains as it was for CRDP and FCRDP in Chapters 3 and 4, respectively. That is, $a_t = 1$ denotes allocating patient $t + 1$ at decision epoch t to treatment A with probability p and treatment B with probability $1 - p$, and $a_t = 2$ denotes allocating patient $t + 1$ to treatment B with probability p and treatment A with probability $1 - p$ (where $0.5 \leq p \leq 1$ for a two-armed trial).

5.2.3 State Transitions

We now need to define all the possible state transitions that can occur under action a_t during time period t, i.e. the random time between decision epochs t and $t + 1$, before we specify their corresponding probabilities, $\mathbb{P}(\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_t, a_t)$. This requires the introduction of the following notation.

Let $K_t \in \{0, \ldots, u_{A,t} + u_{B,t} + 1\}$ be the random number of responses observed (or equivalently, the number of patients leaving the pipeline) during period t (which could include the patient allocated at decision epoch t if they respond before the arrival of the next patient). We will denote the total number of patients in the pipeline just *before* decision epoch t by d_t, so that $d_t = u_{A,t} + u_{B,t}$. Out of the $K_t = k_t$ total responses observed during period t, suppose that we observe $R_t^{s_A} = r_t^{s_A}$ ($R_t^{s_B} = r_t^{s_B}$) successes and $R_t^{f_A} = r_t^{f_A}$ ($R_t^{f_B} = r_t^{f_B}$) failures on treatment A (B), where $k_{A,t} = r_t^{s_A} + r_t^{f_A}$ and $k_{B,t} = r_t^{s_B} + r_t^{f_B}$ are the total numbers of responses from arms A and B, respectively. For notational convenience, we will let $\boldsymbol{R}_t = \left(R_t^{s_A}, R_t^{f_A}, R_t^{s_B}, R_t^{f_B}\right)$ represent the random vector of responses and $\boldsymbol{r}_t = \left(r_t^{s_A}, r_t^{f_A}, r_t^{s_B}, r_t^{f_B}\right)$ be the corresponding vector of realisations.

Then, just before decision epoch $t+1$, the state vector now takes the following general form:

$$\begin{aligned}\boldsymbol{z}_{t+1} = \Big(&u_{A,t} + m_t - (r_t^{s_A} + r_t^{f_A}),\ u_{B,t} + 1 - m_t - (r_t^{s_B} + r_t^{f_B}),\\ &s_{A,t} + r_t^{s_A},\ f_{A,t} + r_t^{f_A},\ s_{B,t} + r_t^{s_B},\ f_{B,t} + r_t^{f_B},\ \widetilde{n} - 1\Big),\end{aligned} \tag{5.2.2}$$

where m_t is an indicator variable taking the value 1 if patient $t+1$ is allocated to treatment A at decision epoch t, or 0 if patient $t+1$ is allocated to treatment B at decision epoch t.

To aid with the understanding and interpretation of the model set-up, we consider the schematic in Figure 5.2.1 which clearly illustrates the ordering of events. In particular, we first update the state vector $\boldsymbol{z}_t$ and only then do we make an allocation decision (represented by the crosses). This means that if an arrival and an observation (of a different patient) happen at the same time, we first incorporate the observation and then make the allocation. We see that at epoch $t = 1$ (when we allocate patient 2): $u_{A,1} = 1$ since patient 1 was allocated to treatment A and has not yet responded so remains in pipeline A; $u_{B,1} = 0$; $k_1 = 2$ because we observe a total of two responses during period 1; $\boldsymbol{r}_1 = (1, 1, 0, 0)$ because we observe one success and one failure from A during period 1; $m_1 = 1$ because patient 2 is allocated to treatment A. At epoch t: $k_t = 3$ since we have a total of three observations during period t; $\boldsymbol{r}_t = (1, 0, 0, 2)$ because we observe one success from A and two failures from B during period t; $m_t = 0$ because patient $t+1$ is allocated to treatment B. Just before patient $t+2$ is allocated at epoch $t+1$: $u_{A,t+1} = u_{A,t} - 1$ since we observe one success from treatment A during period t, and $u_{B,t+1} = u_{B,t} + 1 - 2$ since we first add patient $t+1$ (at epoch t) to pipeline B and subsequently observe two failures from treatment B during period t. Note that it is possible for patient $t+1$ allocated at epoch t to also respond during period t, as depicted in the diagram for patients 2, $t+1$ and n.

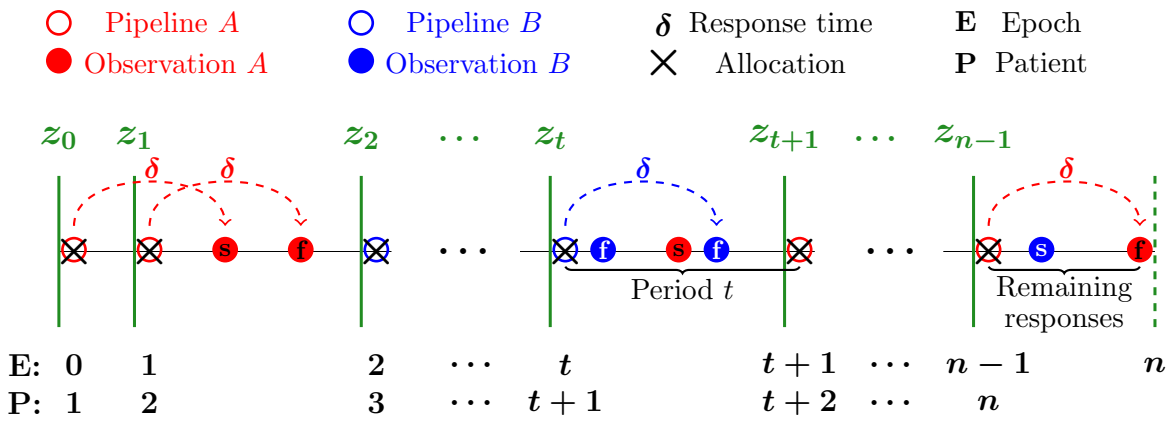


Figure 5.2.1: Schematic of model set-up showing the order in which events occur.

5.2.4 Transition Probabilities

We now turn our attention to the calculation of the transition probabilities which are conditional on the state and action at the current decision epoch t and determine the state of the system at the next decision epoch $t+1$. Under $a_t = 1$, these are given by:

$$\begin{aligned}
\mathbb{P}(\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_t, a_t = 1) &= \sum_{m=0}^{1} \mathbb{P}(\boldsymbol{R}_t = \boldsymbol{r}_t, m_t = m \mid \boldsymbol{z}_t, a_t = 1) \\
&= \sum_{m=0}^{1} \mathbb{P}(\boldsymbol{R}_t = \boldsymbol{r}_t \mid \boldsymbol{z}_t, a_t = 1, m_t = m) \cdot \mathbb{P}(m_t = m \mid \boldsymbol{z}_t, a_t = 1) \\
&= \mathbb{P}(\boldsymbol{R}_t = \boldsymbol{r}_t \mid \boldsymbol{z}_t, a_t = 1, m_t = 0) \cdot (1-p) + \\
&\qquad \mathbb{P}(\boldsymbol{R}_t = \boldsymbol{r}_t \mid \boldsymbol{z}_t, a_t = 1, m_t = 1) \cdot p,
\end{aligned}$$

where $\mathbb{P}(m_t = 0 \mid \boldsymbol{z}_t, a_t = 1) = 1 - p$ and $\mathbb{P}(m_t = 1 \mid \boldsymbol{z}_t, a_t = 1) = p$. We can simplify this further by noting that we do not need to condition on action a_t if we know the value of m_t. Thus, we have

$$\mathbb{P}(\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_t, a_t = 1) = \mathbb{P}(\boldsymbol{R}_t = \boldsymbol{r}_t \mid \boldsymbol{z}_t, m_t = 0) \cdot (1-p) + \mathbb{P}(\boldsymbol{R}_t = \boldsymbol{r}_t \mid \boldsymbol{z}_t, m_t = 1) \cdot p. \tag{5.2.3}$$

The transition probabilities under $a_t = 2$ are defined similarly to (5.2.3), but with the randomisation probabilities p and $1-p$ exchanged, that is,

$$\mathbb{P}(\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_t, a_t = 2) = \mathbb{P}(\boldsymbol{R}_t = \boldsymbol{r}_t \mid \boldsymbol{z}_t, m_t = 0) \cdot p + \mathbb{P}(\boldsymbol{R}_t = \boldsymbol{r}_t \mid \boldsymbol{z}_t, m_t = 1) \cdot (1-p). \tag{5.2.4}$$

We now show how the components of the transition probabilities in (5.2.3) and (5.2.4) can be calculated. Since $\mathbb{P}(\boldsymbol{R}_t = \boldsymbol{r}_t \mid \boldsymbol{z}_t, m_t = 0)$ and $\mathbb{P}(\boldsymbol{R}_t = \boldsymbol{r}_t \mid \boldsymbol{z}_t, m_t = 1)$ are derived in exactly the same way, we will only show how $\mathbb{P}(\boldsymbol{R}_t = \boldsymbol{r}_t \mid \boldsymbol{z}_t, m_t = 1)$ is calculated. For clarity, we will use $\cdot$ to represent $\boldsymbol{z}_t$ and $m_t = 1$ from hereafter.

By conditioning on $K_{A,t} = k_{A,t}$ and $K_{B,t} = k_{B,t}$, and applying the law of total probability, $\mathbb{P}\left(\boldsymbol{R}_t = \boldsymbol{r}_t \mid \cdot\right)$ can be expressed as follows

$$\sum_{k_{A,t}=0}^{u_{A,t}+1} \sum_{k_{B,t}=0}^{u_{B,t}} \mathbb{P}\left(\boldsymbol{R}_t = \boldsymbol{r}_t \mid K_{A,t} = k_{A,t}, K_{B,t} = k_{B,t}, \boldsymbol{z}_t\right) \cdot \mathbb{P}\left(K_{A,t} = k_{A,t}, K_{B,t} = k_{B,t} \mid \cdot\right), \tag{5.2.5}$$

where the first term in equation (5.2.5) simplifies as

$$\begin{aligned} &\mathbb{P}\left(\boldsymbol{R}_t = \boldsymbol{r}_t \mid K_{A,t} = k_{A,t}, K_{B,t} = k_{B,t}, \boldsymbol{z}_t\right) \\ &\qquad = \begin{cases} \mathbb{P}\left(R_t^{s_A} = r_t^{s_A}, R_t^{s_B} = r_t^{s_B} \mid K_{A,t} = k_{A,t}, K_{B,t} = k_{B,t}, \boldsymbol{z}_t\right), \\ \qquad\qquad \text{if } k_{A,t} = r_t^{s_A} + r_t^{f_A} \text{ and } k_{B,t} = r_t^{s_B} + r_t^{f_B}, \\ 0, \text{ otherwise.} \end{cases} \end{aligned} \tag{5.2.6}$$

and, by conditional independence, the first line in (5.2.6) becomes

$$\mathbb{P}\left(R_t^{s_A} = r_t^{s_A} \mid K_{A,t} = k_{A,t}, \boldsymbol{z}_t\right) \cdot \mathbb{P}\left(R_t^{s_B} = r_t^{s_B} \mid K_{B,t} = k_{B,t}, \boldsymbol{z}_t\right). \tag{5.2.7}$$

Since

$$R_t^{s_j} \mid K_{j,t} = k_{j,t}, \boldsymbol{z}_t \sim \text{Binomial}\left(k_{j,t}, \frac{\widetilde{s}_{j,t}}{\widetilde{s}_{j,t} + \widetilde{f}_{j,t}}\right) \text{ for } j \in \{A, B\},$$

equation (5.2.7) is simply the product of two binomial probability mass functions, as follows

$$\binom{k_{A,t}}{r_t^{s_A}} \cdot \left(\frac{\widetilde{s}_{A,t}}{\widetilde{s}_{A,t} + \widetilde{f}_{A,t}} \right)^{r_t^{s_A}} \cdot \left(\frac{\widetilde{f}_{A,t}}{\widetilde{s}_{A,t} + \widetilde{f}_{A,t}} \right)^{r_t^{f_A}} \cdot \binom{k_{B,t}}{r_t^{s_B}} \cdot \left(\frac{\widetilde{s}_{B,t}}{\widetilde{s}_{B,t} + \widetilde{f}_{B,t}} \right)^{r_t^{s_B}} \cdot \left(\frac{\widetilde{f}_{B,t}}{\widetilde{s}_{B,t} + \widetilde{f}_{B,t}} \right)^{r_t^{f_B}},$$

where, recall from Chapter 3, that $\widetilde{s}_{j,t}$ and $\widetilde{f}_{j,t}$ represent the posterior number of successes and failures, respectively, on each arm j (i.e. including the prior pseudo-observations). To calculate the second expression in (5.2.5), namely, the joint probability of $K_{A,t}$ and $K_{B,t}$ conditional on $\boldsymbol{z}_t$ and $m_t = 1$, we first note that $K_{A,t}$ and $K_{B,t}$ are not independent because they are both constrained by the total number of patients in the pipeline, d_t (thus knowledge of $k_{A,t}$ influences what values $k_{B,t}$ can take, and vice versa). Since $k_{A,t} + k_{B,t} = k_t$, it is helpful to re-express $\mathbb{P}\left(K_{A,t} = k_{A,t}, K_{B,t} = k_{B,t} \mid \cdot\right)$ as

$$\mathbb{P}\left(K_{A,t} = k_{A,t}, K_t = k_{A,t} + k_{B,t} \mid \cdot\right) = \mathbb{P}\left(K_{A,t} = k_{A,t} \mid K_t = k_{A,t} + k_{B,t}, \cdot\right) \cdot \mathbb{P}\left(K_t = k_{A,t} + k_{B,t} \mid \cdot\right). \quad (5.2.8)$$

The first term in equation (5.2.8) gives the probability of receiving $k_{A,t}$ responses from treatment A given that the total number of responses observed during period t is $k_{A,t} + k_{B,t}$, and can be calculated as follows. We will consider what happens to this probability in the cases when: (i) patient $t+1$ is observed during period t (i.e. before arrival of patient $t+2$), and (ii) patient $t+1$ is *not* observed during period t.

(i) First, if patient $t+1$ is observed, this implies that all other patients in the pipeline must have been observed[1], that is, $k_{A,t} = u_{A,t} + 1$ and $k_{B,t} = u_{B,t}$. Thus, in

[1]Patients will be observed in the same order as they are allocated owing to the fixed follow-up time.

this case

$$\mathbb{P}\left(K_{A,t} = k_{A,t} \mid K_t = k_{A,t} + k_{B,t}, \cdot\right) = \mathbb{P}\left(K_{A,t} = u_{A,t} + 1 \mid K_t = u_{A,t} + u_{B,t} + 1, \cdot\right) = 1. \tag{5.2.9}$$

Note that if $k_{B,t} \neq u_{B,t}$, the above probability will be 0.

(ii) Now, suppose that patient $t+1$ is *not* observed, that is, $k_{A,t} \leq u_{A,t}$. Then, we obtain the following probability

$$\mathbb{P}\left(K_{A,t} = k_{A,t} \mid K_t = k_{A,t} + k_{B,t}, \cdot\right) = \binom{k_{A,t} + k_{B,t}}{k_{A,t}} \cdot \frac{u_{A,t}}{u_{A,t} + u_{B,t}} \cdot \frac{u_{A,t} - 1}{u_{A,t} + u_{B,t} - 1} \cdot \ldots \cdot \frac{u_{A,t} - k_{A,t} + 1}{u_{A,t} + u_{B,t} - k_{A,t} + 1} \cdot \frac{u_{B,t}}{u_{A,t} + u_{B,t} - k_{A,t}} \cdot \ldots \cdot \frac{u_{B,t} - k_{B,t} + 1}{u_{A,t} + u_{B,t} - k_{A,t} - k_{B,t} + 1}, \tag{5.2.10}$$

which can be succinctly summarised using binomial coefficients as

$$\binom{k_{A,t} + k_{B,t}}{k_{A,t}} \cdot \binom{u_{A,t} - k_{A,t} + u_{B,t} - k_{B,t}}{u_{A,t} - k_{A,t}} \Big/ \binom{u_{A,t} + u_{B,t}}{u_{A,t}}. \tag{5.2.11}$$

This is the probability mass function of a hypergeometric distribution with parameters $u_{A,t} + u_{B,t}, k_{A,t} + k_{B,t}$ and $u_{A,t}$.

The calculation of the second term in equation (5.2.8), i.e. $\mathbb{P}\left(K_t = k_{A,t} + k_{B,t} \mid \cdot\right)$, is now discussed. First, we introduce some further notation. Suppose that we have obtained a total of $x_t = s_{A,t} + f_{A,t} + s_{B,t} + f_{B,t}$ observations just before decision epoch t. Further, let τ_i denote the random inter-arrival time of patient $i \in \{1, \ldots, n\}$ (where $\tau_1 = 0$) and δ be some constant representing the fixed follow-up time (which is the same for every patient). Thus, it follows that $\sum_{i=1}^{x_t} \tau_i$ is the arrival time of patient x_t and $\sum_{i=1}^{x_t} \tau_i + \delta$ is the corresponding observation time.

The possible events that can happen during period t (i.e. the random time between two allocations) are as follows:

(i) The first pipeline patient, i.e. patient $x_t + 1$, is observed. This happens at time

$\sum_{i=1}^{x_t+1} \tau_i + \delta$.
(ii) The second pipeline patient, i.e. patient $x_t + 2$, is observed. This happens at time $\sum_{i=1}^{x_t+2} \tau_i + \delta$.
$\vdots$

$(d_t + 1)$ The final pipeline patient, which is the patient allocated most recently at decision epoch t, i.e patient $x_t + d_t + 1$, is observed. This happens at time $\sum_{i=1}^{x_t+d_t+1} \tau_i + \delta$. The next patient to be allocated (at epoch $t+1$) is patient $x_t + d_t + 2$, which happens at time $\sum_{i=1}^{x_t+d_t+2} \tau_i$.

These allow us to derive the conditions required to observe k_t responses. In particular:

- No responses ($k_t = 0$) will be observed during period t if and only if patient $x_t + d_t + 2$ is allocated *before* the first-in-pipeline patient, patient $x_t + 1$, is observed. This occurs when

$$\sum_{i=1}^{x_t+d_t+2} \tau_i < \sum_{i=1}^{x_t+1} \tau_i + \delta, \quad \text{i.e.} \quad \sum_{i=x_t+2}^{x_t+d_t+2} \tau_i < \delta. \tag{5.2.12}$$

- $k_t = l_t$ responses will be observed during period t, where $1 \leq l_t \leq d_t$, if and only if patient $x_t + d_t + 2$ is allocated both *after* the $l_t{}^{\text{th}}$ pipeline patient, i.e. patient $x_t + l_t$, is observed and *before* patient $x_t + l_t + 1$ is observed. This occurs when

$$\begin{aligned} &\sum_{i=1}^{x_t+d_t+2} \tau_i \geq \sum_{i=1}^{x_t+l_t} \tau_i + \delta, \quad \text{i.e.} \quad \sum_{i=x_t+l_t+1}^{x_t+d_t+2} \tau_i \geq \delta \quad \text{and} \\ &\sum_{i=1}^{x_t+d_t+2} \tau_i < \sum_{i=1}^{x_t+l_t+1} \tau_i + \delta, \quad \text{i.e.} \quad \sum_{i=x_t+l_t+2}^{x_t+d_t+2} \tau_i < \delta. \end{aligned} \tag{5.2.13}$$

  Note that if patient $x_t + l_t$ does not respond, this implies that all other pipeline patients thereafter must not respond since their responses are ordered owing to

the fixed follow-up time of each patient.

- If every pipeline patient is observed during period t, we will obtain $k_t = d_t + 1$ observations. This will happen if and only if patient $x_t + d_t + 2$ is allocated *after* patient $x_t + d_t + 1$, i.e. the final pipeline patient, has been observed. This occurs when

$$\sum_{i=1}^{x_t+d_t+2} \tau_i \geq \sum_{i=1}^{x_t+d_t+1} \tau_i + \delta, \quad \text{i.e. } \tau_{x_t+d_t+2} \geq \delta. \tag{5.2.14}$$

At this point, we are now in a position where the corresponding probabilities of the events in (5.2.12), (5.2.13) and (5.2.14) can be derived for a specified inter-arrival distribution. We outline this in Example 1 below before showing that we can, in fact, condition on further information to make these probabilities closer to their true values which could be obtained if we knew the exact arrival (treatment) times. For illustrative purposes, and in keeping with related literature (e.g. Biswas and Coad, 2005; Hardwick *et al.*, 2006; Zhang and Rosenberger, 2007), we will assume from hereon that the inter-arrival times τ_i are independent and identically distributed (i.i.d.) exponential random variables with rate parameter equal to λ for all $i \in \{1, \ldots, n\}$. However, note that the same principles easily apply to other distribution types. Moreover, it is important to keep in mind that these probabilities are inherently conditioned upon the information provided by the current state vector $\boldsymbol{z}_t$, but we omit this explicit dependence from the following calculations for simplicity.

Example 1: No Further Conditioning

- Let $W = \sum_{i=x_t+2}^{x_t+d_t+2} \tau_i$. Since this is the sum of $d_t + 1$ i.i.d. exponential random variables, it follows that $W \sim \text{Gamma}(d_t + 1, \lambda)$. Thus, from (5.2.12), the probability of observing no responses during period t is given by

$$\mathbb{P}(W < \delta) = \frac{\gamma(d_t + 1, \delta\lambda)}{\Gamma(d_t + 1)}, \tag{5.2.15}$$

that is, the cumulative distribution function of W, where γ is the lower incomplete gamma function and Γ is the gamma function.

- Next, we find the probability of (5.2.13). Let $X = \tau_{x_t+l_t+1} \sim \text{Exp}(\lambda)$ and $Y = \sum_{i=x_t+l_t+2}^{x_t+d_t+2} \tau_i \sim \text{Gamma}(d_t - l_t + 1, \lambda)$. It follows that the probability of observing l_t responses during period t, where $1 \leq l_t \leq d_t$, is equivalent to

$$\mathbb{P}\left[(X + Y \geq \delta) \cap (Y < \delta)\right] = \int_0^{\delta} \int_{\delta - y}^{\infty} f_{X,Y}(x, y)\, dx\, dy = \frac{(\delta\lambda)^{d_t - l_t + 1}}{\Gamma(d_t - l_t + 2)} \cdot \exp(-\delta\lambda), \tag{5.2.16}$$

where $f_{X,Y}$ is the joint probability density function of X and Y.

- Finally, to calculate the probability of (5.2.14), that is, of observing all $d_t + 1$ responses during period t, we require $\mathbb{P}(Z \geq \delta)$, where $Z = \tau_{x_t+d_t+2} \sim \text{Exp}(\lambda)$, which is given by

$$1 - F_Z(\delta) = \exp(-\delta\lambda), \tag{5.2.17}$$

where F_Z is the cumulative distribution function of Z evaluated at δ.

We now need to check that the probabilities in (5.2.15), (5.2.16) and (5.2.17) sum to one. We have

$$\sum_{k_t=0}^{d_t+1} \mathbb{P}(K_t = k_t) = \frac{\gamma(d_t + 1, \delta\lambda)}{\Gamma(d_t + 1)} + \left\{1 + \sum_{k_t=1}^{d_t} \frac{(\delta\lambda)^{d_t - k_t + 1}}{\Gamma(d_t - k_t + 2)}\right\} \cdot \exp(-\delta\lambda). \tag{5.2.18}$$

Applying the recurrence relation $\gamma(d_t + 1, \delta\lambda) = d_t \cdot \gamma(d_t, \delta\lambda) - (\delta\lambda)^{d_t} \cdot \exp(-\delta\lambda)$ (see e.g. Jameson, 2016, p.300) iteratively, we can express $\gamma(d_t + 1, \delta\lambda)$ in terms of $\gamma(1, \delta\lambda)$ as follows

$$\gamma(d_t+1, \delta\lambda) = \Gamma(d_t+1)\cdot\gamma(1, \delta\lambda) - \Gamma(d_t+1)\cdot\left\{\sum_{k_t=1}^{d_t} \frac{(\delta\lambda)^{d_t - k_t + 1}}{\Gamma(d_t - k_t + 2)}\right\}\cdot\exp(-\delta\lambda), \tag{5.2.19}$$

where $\gamma(1, \delta\lambda) = 1 - \exp(-\delta\lambda)$. Substituting (5.2.19) into equation (5.2.18) and

simplifying gives the value 1, as required.

Further Conditioning

To make the calculations of $\mathbb{P}(K_t = k_t \mid \cdot)$ more efficient, we note that there is additional information which can be conditioned upon. In particular, recall that the current state at epoch t is given by the vector $\boldsymbol{z}_t$ (refer to the schematic in Figure 5.2.1) from which the total number of observations, x_t, and patients in the pipeline, d_t, by epoch t can be deduced. Therefore, by conditioning on $\boldsymbol{z}_t$ (which we do throughout calculation of the transition probabilities), this implies that the allocation of patient $x_t + d_t + 1$ (i.e. the patient allocated most recently at epoch t) must happen after we have observed x_t responses, but before we have observed $x_t + 1$ responses. This translates to the following pair of conditions

$$\sum_{i=1}^{x_t+d_t+1} \tau_i \geq \sum_{i=1}^{x_t} \tau_i + \delta \text{ i.e. } \sum_{i=x_t+1}^{x_t+d_t+1} \tau_i \geq \delta \text{ and}$$
$$\sum_{i=1}^{x_t+d_t+1} \tau_i < \sum_{i=1}^{x_t+1} \tau_i + \delta \text{ i.e. } \sum_{i=x_t+2}^{x_t+d_t+1} \tau_i < \delta,$$

which we will denote by the event $\mathcal{C}$. That is

$$\mathcal{C} = \left(\sum_{i=x_t+1}^{x_t+d_t+1} \tau_i \geq \delta \right) \cap \left(\sum_{i=x_t+2}^{x_t+d_t+1} \tau_i < \delta \right). \tag{5.2.21}$$

$\mathcal{C}$ represents the ideal conditioning which makes full use of the available information contained in $\boldsymbol{z}_t$ and is therefore the version that we implement[2]. This will result in a better approximation to the true transition probabilities, and hence optimal solution, which would be attained if we conditioned on the exact arrival times of the pipeline patients. However, introducing arrival times into $\boldsymbol{z}_t$ would make the problem

[2]For illustrative purposes, and to show the natural development of ideas, we present the initial event we conditioned upon, along with the derivation of the corresponding marginal probabilities, in Example 2 of Appendix 5.5.1.

intractable.

As previously mentioned, the exact evaluation of the corresponding probabilities will depend upon the distributional assumption of τ_i. Although analytical evaluation may be feasible for some distributions, not all distributions will give rise to probabilities that can be expressed in closed-form (as seen for the exponential case in Example 1 above and Example 2 of Appendix 5.5.1, which rely on the incomplete gamma function). Therefore, the probabilities of the events in (5.2.12), (5.2.13) and (5.2.14) when conditioned on $\mathcal{C}$ will be approximated using Monte Carlo simulation so that our implementation remains as general as possible.

5.2.5 Expected One-Period Rewards

To complete the formulation of the MDP, we need to define the expected one-period (or immediate) *rewards* after transitioning from state vector $\boldsymbol{z}_t$ to $\boldsymbol{z}_{t+1}$ under action a_t, which we denote by $\mathcal{R}^{a_t}(\boldsymbol{z}_t)$. This depends on the objective function of interest which, in this case, is to maximise the expected total number of patient successes in the trial[3] (formally defined in Chapter 3), whereby we obtain a reward of 1 for a success and 0 for a failure. The expected one-period reward for periods $0 \leq t \leq n-2$ under m_t is given by

$$\mathcal{R}^{m_t}(\boldsymbol{z}_t) = \sum_{r_t^{s_A}=0}^{u_{A,t}+m_t} \sum_{r_t^{f_A}=0}^{u_{A,t}+m_t-r_t^{s_A}} \sum_{r_t^{s_B}=0}^{u_{B,t}+1-m_t} \sum_{r_t^{f_B}=0}^{u_{B,t}+1-m_t-r_t^{s_B}} (r_t^{s_A} + r_t^{s_B}) \cdot \mathbb{P}(\boldsymbol{R}_t = \boldsymbol{r}_t \mid \cdot), \tag{5.2.22}$$

where we refer the reader to equation (5.2.5) for the calculation of $\mathbb{P}(\boldsymbol{R}_t = \boldsymbol{r}_t \mid \cdot)$.

After the final patient has been allocated at epoch $t = n-1$, i.e. when $\widetilde{n} = 0$, we use the information accrued during the trial to predict how many of the remaining $u_{A,n-1} + u_{B,n-1} + 1$ pipeline patients will have a success. Consequently, the expected

[3]Note that this approach applies to objective functions beyond the bandit objective of maximising reward.

one-period reward for period $n-1$ under m_t is

$$\mathcal{R}^{m_t}(\boldsymbol{z}_t) = (u_{A,t} + m_t) \cdot \frac{\widetilde{s}_{A,t}}{\widetilde{s}_{A,t} + \widetilde{f}_{A,t}} + (u_{B,t} + 1 - m_t) \cdot \frac{\widetilde{s}_{B,t}}{\widetilde{s}_{B,t} + \widetilde{f}_{B,t}} + \epsilon \tag{5.2.23}$$

for $t = n-1$, where

$$\epsilon = \begin{cases} -n, & \text{if } u_{A,t} + m_t + s_{A,t} + f_{A,t} < \ell \text{ or } u_{B,t} + 1 - m_t + s_{B,t} + f_{B,t} < \ell, \\ 0 & \text{otherwise,} \end{cases}$$

which serves as a penalty to avoid the design choosing states that give rise to fewer than the desired number of allocations on each arm, ℓ. This avoids extreme imbalance between the treatment arms (refer to Section 3.2.3).

It follows that the corresponding expected one-period rewards under action $a_t = 1$ and $a_t = 2$, respectively, for periods $0 \leq t \leq n-1$ are of the form

$$\begin{aligned} \mathcal{R}^{a_t=1}(\boldsymbol{z}_t) &= p \cdot \mathcal{R}^{m_t=1}(\boldsymbol{z}_t) + (1-p) \cdot \mathcal{R}^{m_t=0}(\boldsymbol{z}_t) \text{ and} \\ \mathcal{R}^{a_t=2}(\boldsymbol{z}_t) &= (1-p) \cdot \mathcal{R}^{m_t=1}(\boldsymbol{z}_t) + p \cdot \mathcal{R}^{m_t=0}(\boldsymbol{z}_t), \end{aligned} \tag{5.2.24}$$

with $\mathcal{R}^{m_t}$ replaced by (5.2.22) for $0 \leq t \leq n-2$ and (5.2.23) for $t = n-1$. We will take the randomisation probability p to be 0.9 in the following simulations (as in Chapter 3).

5.2.6 Obtaining the Optimal Solution

The Bellman equation, defined in (2.2.2) of Chapter 2, immediately follows as the expected one-period rewards, defined in (5.2.24), plus the expected (undiscounted) future rewards. As in Chapters 3 and 4, the corresponding optimisation problem is to maximise the expected total reward over the entire finite time horizon for a uniform prior distribution on each arm at $t = 0$, which is solved exactly using backward induction (described in Section 2.2.3 and Appendix 3.6.1). This gives rise to the

optimal treatment allocation policy which prescribes the optimal action to use in every possible combination of states for all t, along with the corresponding maximum expected total reward. For example, when the system is in state $\boldsymbol{z}_t = (u_{A,t} = 5, u_{B,t} = 8, s_{A,t} = 2, f_{A,t} = 4, s_{B,t} = 3, f_{B,t} = 3, \widetilde{n} = 35)$ just before epoch $t = 25$, the optimal action obtained from the RCRDP design is $a_{25} = 2$. For a trial with $n = 60$ patients, treatment success probabilities $(\theta_A, \theta_B) = (0.5, 0.7)$, exponential inter-arrival times with rate parameter $\lambda = 20$ and a follow-up time of $\delta = 1$, this is interpreted as follows. If there are 13 patients in the pipeline (5 on A and 8 on B), 12 patients which have currently been observed (2 successes from arm A, 4 failures from arm A, 3 successes from arm B and 3 failures from arm B) with 35 patients remaining to be treated, then the 26^{th} patient will be randomised to arm B with probability 0.9 (and arm A with probability 0.1).

5.3 Simulation Results

We implement the RCRDP design assuming that patients arrive via a Poisson process with rate λ or, equivalently, that the inter-arrival times follow i.i.d. exponential distributions, $\tau_i \sim \text{Exp}(\lambda)$ $(i = 1, \ldots, n)$, with mean $1/\lambda$. As patients arrive, they are immediately allocated to either treatment A or B based on data accrued so far. After a follow-up time of δ has elapsed, the patient's outcome (either a success or failure) is observed and used to update the states accordingly. We assume that the patient arrival rate λ and follow-up time δ is known. Note that this delay structure is purely for illustrative purposes and the RCRDP design can be applied to any appropriate arrival and/or response time distribution. As an example, results for a small trial with inter-arrival times determined by the discrete analogue of the exponential distribution instead, i.e. the geometric distribution, are illustrated in Appendix 5.5.2.

We present the simulation results for a trial with 60 patients and treatment success

probabilities as used in previous chapters, namely, $\theta_A = 0.5$ and $\theta_B \in (0.1, 0.9)$. Wason *et al.* (2019) highlight that "methodological papers often do not consider the rate of enrolment versus the length of follow-up for outcomes when quantifying the efficiency advantages of adaptive designs." However, we do consider this here by fixing the follow-up time at $\delta = 1$ and varying the arrival rates, which we take to be $\lambda = 10, 20, 30, 50$. This means that if a unit of time is interpreted as one week, say, then λ patients are expected to enter the trial per week, each of which are followed up exactly one week later. Also included are the results corresponding to values of $\lambda \to 0$ and $\lambda \to \infty$, which represent the two extreme situations of immediate response (IR) and equal fixed randomisation (FR), respectively.

For comparative purposes, we plot the analogous results for CRDP (introduced in Chapter 3) and FCRDP (introduced in Chapter 4) alongside those for RCRDP, all of which are the average of $1,000,000$ simulation runs. These are displayed in Figure 5.3.1 which is discussed below. Moreover, to check that the differences between CRDP, FCRDP and RCRDP are not attributed to any underlying interaction between the delay length, constraint and randomisation (refer to Chapter 4), we also consider the optimal version of each design which has the constraint and randomisation removed. These are referred to as DP, FDP and RDP, and the corresponding results are presented in Appendix 5.5.3 where they are shown to exhibit similar patterns to those observed for CRDP, FCRDP and RCRDP, respectively.

First note that as the arrival rate λ increases, the expected inter-arrival time decreases and, since the follow-up time remains fixed, the number of patients in the pipeline will accumulate. In other words, the "delay" length increases with λ.

The top left plot in Figure 5.3.1 shows that there is practically no difference between the power values obtained for all three versions of the design. As λ (and hence the delay length) increases, we see that the power also increases which reaffirms what was found in Chapter 4. When $\lambda = 50$ (see red line), the resulting power is

approximately the same as that achieved by FR (i.e. when $\lambda \to \infty$).

In terms of the percentage of patients allocated to the superior treatment (top right plot in Figure 5.3.1), which is indicative of patient benefit, we observe that RCRDP offers an improvement over CRDP for all values of λ (excluding the two extremes in which RCRDP reduces to FR and CRDP). For example, when $\theta_B = 0.1$ and $\lambda = 20$ (see blue line), an additional 1% of patients, on average, will receive the superior treatment. Similarly for an arrival rate which is half the trial size, $\lambda = 30$ (see green line). However, when RCRDP (see black, dot-dashed lines) is compared to FCRDP (see coloured, dashed lines), the difference between their performance is very small.

Further note that as λ increases, the patient benefit decreases (which is what we expect from Chapter 4). In particular, the best case occurs when all responses have been observed immediately so that full information is retained (see pink line), whilst the "worst" case occurs when no responses have been observed until after all patients have been allocated, i.e. FR (see grey line).

The additional gains in patient benefit achieved by RCRDP over CRDP are also clearly demonstrated in Figure 5.5.1 of Appendix 5.5.2 for the geometric case.

The bottom two plots in Figure 5.3.1 illustrate the changes in the average bias and MSE of the treatment effect estimator, which decrease as the arrival rate λ increases. The observed differences in bias and MSE between the designs are negligible (note the extremely small scale of the plot), with RCRDP and FCRDP almost identical.

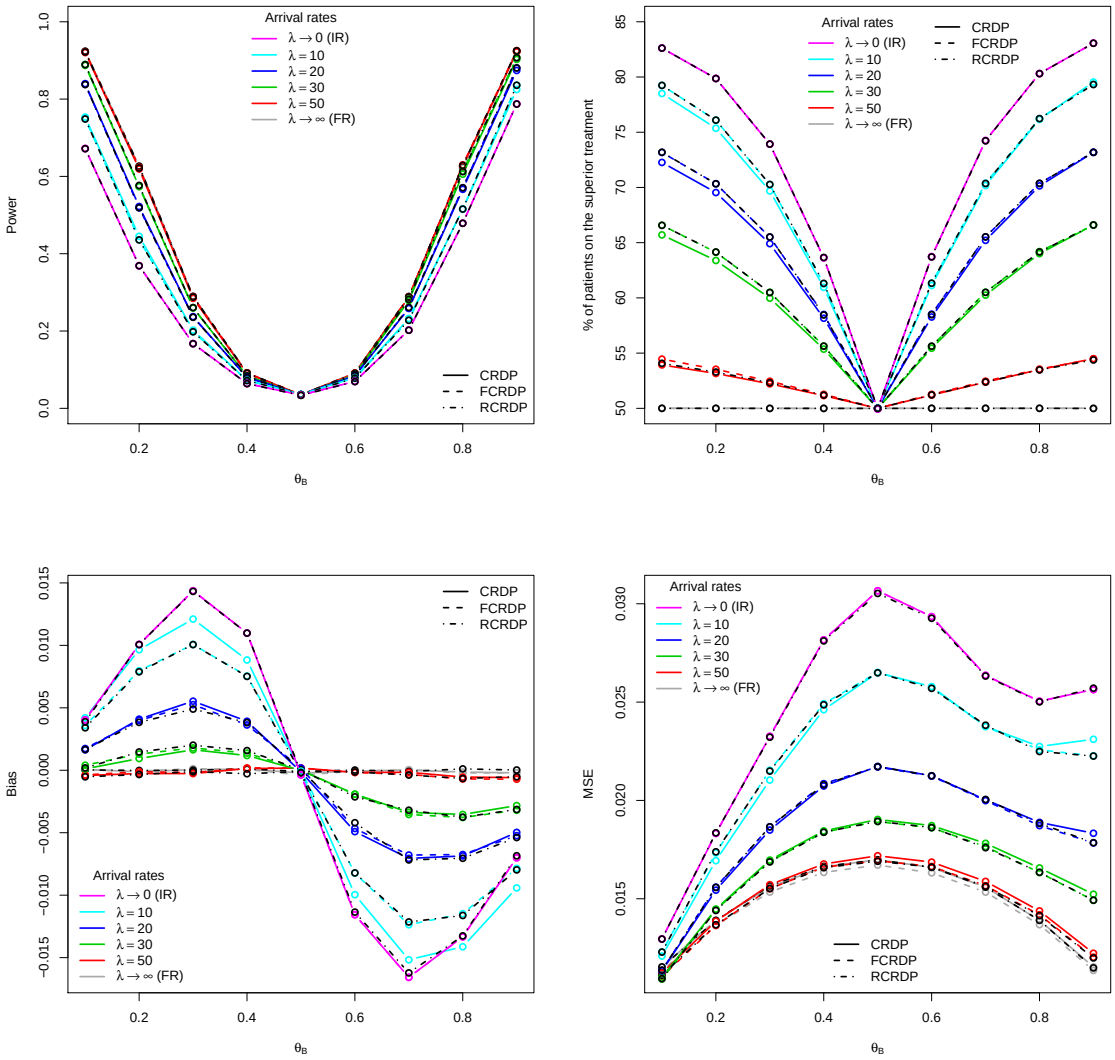


Figure 5.3.1: The changes in power (and type I error), % of patients on the superior treatment, the average bias and MSE of the treatment effect estimator for the CRDP, FCRDP and RCRDP designs when $n = 60$, $\theta_A = 0.5$, $\theta_B \in (0.1, 0.9)$, $\tau_i \sim \text{Exp}(\lambda)$ and $\delta = 1$ (estimated over $1,000,000$ simulations). IR, immediate response and FR, fixed randomisation.

5.4 Summary

In this chapter, we have built upon the ideas introduced in Chapters 3 and 4 to present a general model, R(CR)DP, for the bandit problem with delayed information using

the Bayesian MDP framework and solution by dynamic programming. Incorporating delayed information provides a powerful modelling framework which can not only be used for a greater variety of real life clinical trials but, as discussed in Caro and Yoo (2010), "can be generalised to aid decision making in many [other] application areas." Examples include, but are not limited to, online advertising (Chapelle, 2014; Vernade *et al.*, 2017), dynamic assortment (Caro and Gallien, 2007) and bandwidth allocation (Ehsan and Liu, 2004).

We illustrated the workings of RCRDP for random arrivals with fixed response times (since this is most pertinent to clinical trials with a binary endpoint), but the fundamental principles apply to the most general structure of random arrivals with random response times (as in survival trials, for example, which are beyond the scope of this thesis). In particular, the state space (and hence dimension/computational complexity of the problem), action set, state transitions and specification of the corresponding Bellman equation will remain the same. The transition probabilities, and consequently the reward function, will take a slightly different form because the order in which responses are observed will no longer be known (which will affect the derivation of equation (5.2.8)). However, these can still be computed using Monte Carlo simulation.

Moreover, although we implemented RCRDP assuming exponential inter-arrival times, the underlying DP formulation will remain the same regardless of the arrival and/or response distribution which can simply be adjusted within the Monte Carlo simulation. This is an advantage over the DP solution implemented by Hardwick *et al.* (2006, Approach II) which is more restrictive and applies specifically to the exponential delay model. In particular, Hardwick *et al.* (2006) comment that, "unfortunately, optimising and evaluating different arrival and response delay models can involve significantly different recursive equations, and the computational requirements can vary dramatically."

The first approach suggested by Hardwick *et al.* (2006) is the optimal design for the standard two-armed bandit problem with delay, based upon the DP approach. The corresponding Bellman equation is stated in Hardwick *et al.* (2006, Appendix A) and takes a similar form to ours. However, there is no explanation provided as to how the probabilities t, q_1, q_2, equivalent to our $\mathbb{P}\left(K_{A,t} = k_{A,t}, K_{B,t} = k_{B,t} \mid \cdot\right)$, $\mathbb{P}\left(R_t^{s_A} = r_t^{s_A} \mid K_{A,t} = k_{A,t}, \boldsymbol{z}_t\right)$, $\mathbb{P}\left(R_t^{s_B} = r_t^{s_B} \mid K_{B,t} = k_{B,t}, \boldsymbol{z}_t\right)$ (see equation (5.2.5)), can be calculated. Furthermore, they do not implement this solution because, at the time of publication, it was computationally infeasible. In contrast, we are able to implement our DP solution to the bandit problem with delay (using the programming language `R`) for sample sizes up to 100 on a standard laptop with 16GB of RAM. Recall that this only needs to be computed once and can then be stored for future use.

Our results showed that RCRDP consistently improved patient benefit compared to CRDP, with an inconsequential effect on the corresponding power, bias and MSE. However, similar gains were also achieved by the FCRDP design, at least for the setting considered here. Therefore, given the additional complexity and increased computational requirements associated with RCRDP, the FCRDP design is preferable[4]. It would be interesting to see if the FCRDP and RCRDP designs continue to perform similarly for other inter-arrival distributions that do not possess the memoryless property (e.g. the Weibull distribution), as well as for more general settings involving random arrivals and random response times. This forms a topic for future work.

The results also illustrated that the reformulation of the problem in terms of random arrivals and a fixed follow-up time provides the same conclusions as those obtained in Chapter 4 for the case of sequential arrivals and a random response time.

In this work, we have assumed that only information on the primary endpoint is

[4]This point has been highlighted in Wason *et al.* (2019), for example, which states that "it is also important to carefully consider reducing the complexity of an adaptive design when the efficiency gains are marginal".

available for updating the allocation probabilities. However, many clinical trials include a short-term or surrogate endpoint which is correlated with, and observed more quickly than, the primary endpoint (e.g. Tamura *et al.*, 1994). Therefore, an "alternative is to adapt on the basis of some surrogate measure" (Rosenberger and Lachin, 1993). This approach has been considered in the group sequential setting by Hampson and Jennison (2013), for example, who demonstrated that the loss of efficiency caused by a delayed response can be ameliorated by incorporating information on the short-term endpoint. This raises the question of how data on a short-term endpoint can be incorporated into the proposed designs. The inclusion of short-term endpoints into RAR methods has not received much attention in the literature (Nowacki *et al.*, 2017) and hence provides an opportune area for further research.

An alternative approach that can be implemented to dilute the effects of delayed responses is block (or cohort) RAR, in which the allocation probabilities are updated only after groups of patients respond, rather than after each individual patient responds (Rosenberger and Lachin, 1993; Karrison *et al.*, 2003; Sverdlov *et al.*, 2012; Perchet *et al.*, 2016). A block RAR design, which is also less computationally intensive than the DP-based designs discussed so far, is proposed in the following chapter.

5.5 Appendix

5.5.1 Example 2: Initial Version of Further Conditioning

As discussed in Section 5.2.4, obtaining the transition probabilities requires evaluation of $\mathbb{P}\left(K_t = k_t \mid \cdot\right)$, where $k_t = 0, \ldots, d_t + 1$, for every possible value of d_t. This can be achieved by calculating the probabilities of the events defined in (5.2.12), (5.2.13) and (5.2.14) (which was illustrated in Example 1 for exponential inter-arrival times). However, we can condition on further information which will improve the accuracy of the probability estimates. Here, we provide the initial version that was considered

before implementing the stricter conditioning defined in (5.2.21). This example also includes the exact derivation of the corresponding probabilities for exponential inter-arrival times.

Initially, we conditioned on the fact that the allocation of patient $x_t + d_t + 2$ must happen after we have observed x_t responses. That is,

$$\sum_{i=1}^{x_t+d_t+2} \tau_i \geq \sum_{i=1}^{x_t} \tau_i + \delta, \quad \text{i.e.} \quad \sum_{i=x_t+1}^{x_t+d_t+2} \tau_i \geq \delta. \tag{5.5.1}$$

- To calculate the probability that $k_t = 0$, let $W = \sum_{i=x_t+2}^{x_t+d_t+2} \tau_i$. Since this is the sum of $d_t + 1$ i.i.d. exponential random variables, it follows that $W \sim \text{Gamma}(d_t + 1, \lambda)$. Thus, from (5.2.12), the probability of no observations conditional on the event in (5.5.1) is given by

$$\mathbb{P}(W < \delta \mid W + \tau_{x_t+1} \geq \delta) = \frac{\mathbb{P}(W < \delta, W + \tau_{x_t+1} \geq \delta)}{\mathbb{P}(W + \tau_{x_t+1} \geq \delta)}. \tag{5.5.2}$$

To calculate the numerator of (5.5.2), let $V = \tau_{x_t+1} \sim \text{Exp}(\lambda)$. Thus,

$$\mathbb{P}(W < \delta, W + V \geq \delta) = \int_0^{\delta} \int_{\delta-w}^{\infty} f_{V,W}(v, w)\, dv\, dw = \frac{(\delta\lambda)^{d_t+1} \cdot \exp(-\delta\lambda)}{\Gamma(d_t + 2)}, \tag{5.5.3}$$

where $f_{V,W}$ is the joint probability density function of V and W.

The denominator of (5.5.2) is given by

$$\mathbb{P}(W + V \geq \delta) = 1 - \mathbb{P}(W + V < \delta) = 1 - \frac{\gamma(d_t + 2, \delta\lambda)}{\Gamma(d_t + 2)}, \tag{5.5.4}$$

where $V + W \sim \text{Gamma}(d_t + 2, \lambda)$ and γ is the lower incomplete gamma function.

Therefore, from (5.5.3) and (5.5.4), it follows that the **probability of obtain-**

ing no observations during period t is given by

$$\frac{(\delta\lambda)^{d_t+1} \cdot \exp(-\delta\lambda)}{\Gamma(d_t+2) - \gamma(d_t+2, \delta\lambda)}. \tag{5.5.5}$$

- Next, we use the events defined in (5.2.13) to find the probability that $k_t = l_t$, conditional on the event in (5.5.1), where $1 \leq l_t \leq d_t$. Let $X = \sum_{i=x_t+1}^{x_t+l_t} \tau_i$, $Y = \tau_{x_t+l_t+1}$ and $Z = \sum_{i=x_t+l_t+2}^{x_t+d_t+2} \tau_i$, then the required probability can be expressed as $\mathbb{P}\left\{(Z < \delta, Y + Z \geq \delta) \mid (X + Y + Z \geq \delta)\right\}$, which is equivalent to

$$\frac{\mathbb{P}(Z < \delta, Y + Z \geq \delta, X + Y + Z \geq \delta)}{\mathbb{P}(X + Y + Z \geq \delta)} = \frac{\mathbb{P}(Z < \delta, Y + Z \geq \delta)}{\mathbb{P}(X + Y + Z \geq \delta)}. \tag{5.5.6}$$

First, we calculate the numerator of (5.5.6) as follows

$$\mathbb{P}(Z < \delta, Y + Z \geq \delta) = \int_0^{\delta} \int_{\delta - z}^{\infty} f_{Y,Z}(y, z)\, dy\, dz = \frac{(\delta\lambda)^{d_t - l_t + 1} \cdot \exp(-\delta\lambda)}{\Gamma(d_t - l_t + 2)}. \tag{5.5.7}$$

The denominator, i.e. the probability of the event (5.5.1) being conditioned upon, has already been calculated in equation (5.5.4). Thus, it follows that the **probability of observing l_t responses**, where $1 \leq l_t \leq d_t$, during period t is given by

$$\frac{\Gamma(d_t+2) \cdot (\delta\lambda)^{d_t - l_t + 1} \cdot \exp(-\delta\lambda)}{\Gamma(d_t - l_t + 2) \cdot \{\Gamma(d_t+2) - \gamma(d_t+2, \delta\lambda)\}}. \tag{5.5.8}$$

- Finally, to calculate the probability that *all* pipeline patients are observed during period t, that is, we receive $d_t + 1$ observations during period t, we require the probability of the event in (5.2.14) conditional on the event in (5.5.1), which is

expressed by

$$\mathbb{P}\left(\tau_{x_t+d_t+2} \geq \delta \;\middle|\; \sum_{i=x_t+1}^{x_t+d_t+2} \tau_i \geq \delta\right) = \frac{\mathbb{P}\left(\tau_{x_t+d_t+2} \geq \delta, \sum_{i=x_t+1}^{x_t+d_t+2} \tau_i \geq \delta\right)}{\mathbb{P}\left(\sum_{i=x_t+1}^{x_t+d_t+2} \tau_i \geq \delta\right)} = \frac{\mathbb{P}\left(\tau_{x_t+d_t+2} \geq \delta\right)}{\mathbb{P}\left(\sum_{i=x_t+1}^{x_t+d_t+2} \tau_i \geq \delta\right)},$$

where $\mathbb{P}(\tau_{x_t+d_t+2} \geq \delta) = 1 - \mathbb{P}(\tau_{x_t+d_t+2} < \delta) = \exp(-\delta\lambda)$ and the denominator is as calculated previously in (5.5.4). Therefore, the **probability of obtaining $d_t + 1$ observations** during period t is given by

$$\frac{\Gamma(d_t+2) \cdot \exp(-\delta\lambda)}{\Gamma(d_t+2) - \gamma(d_t+2, \delta\lambda)}. \tag{5.5.9}$$

The probabilities in (5.5.5), (5.5.8) and (5.5.9) sum to one, as required.

5.5.2 Results for CRDP and RCRDP with Geometric Inter-Arrival Times

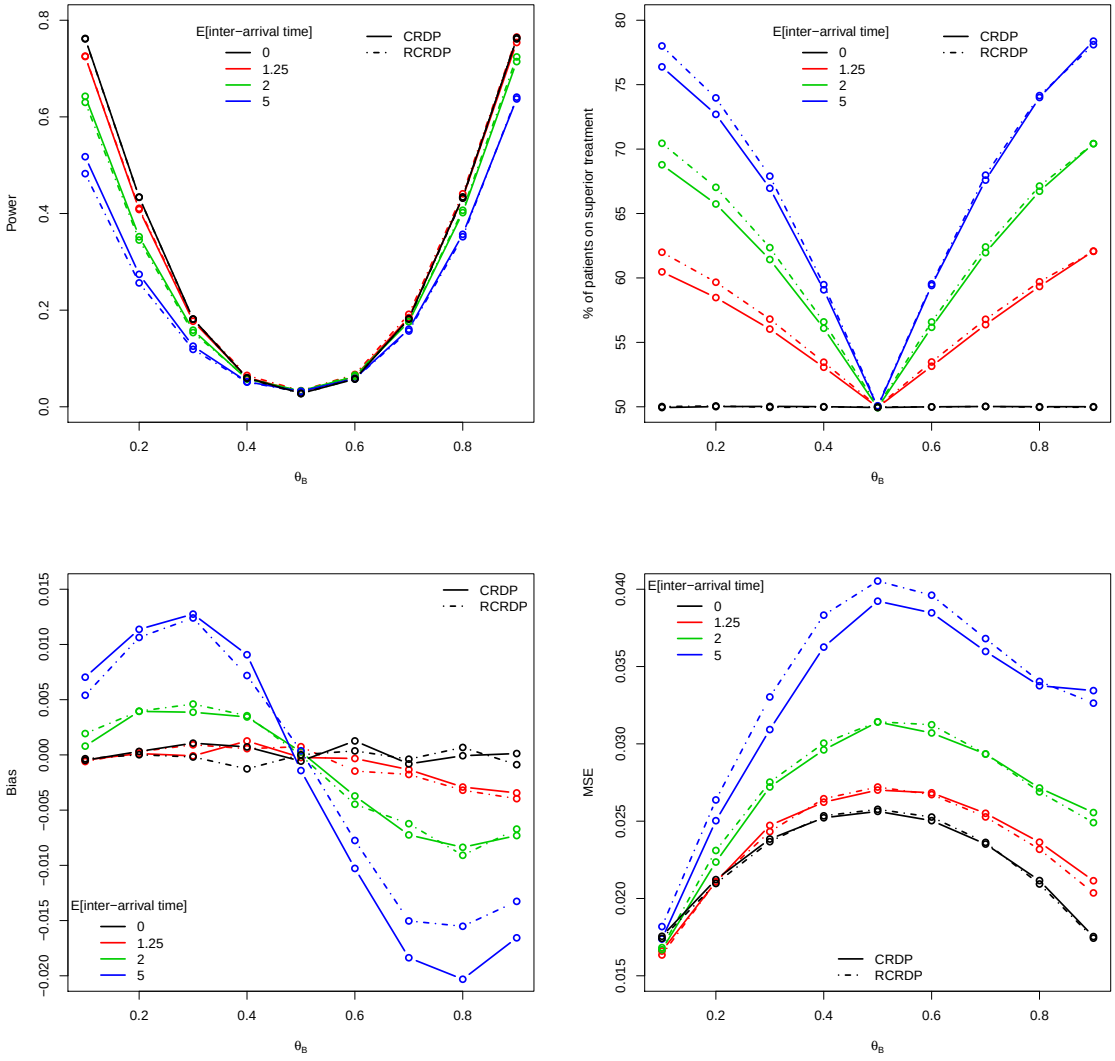


Figure 5.5.1: The changes in power (and type I error), % of patients on the superior treatment, bias and MSE for RCRDP (dot-dashed line) and CRDP (solid line) with geometric inter-arrival times when $n = 40$, $\delta = 30$, $\theta_A = 0.5$ and $\theta_B \in (0.1, 0.9)$ (estimated over $100,000$ simulations).

5.5.3 Results for the DP Variants with Exponential Inter-Arrival Times

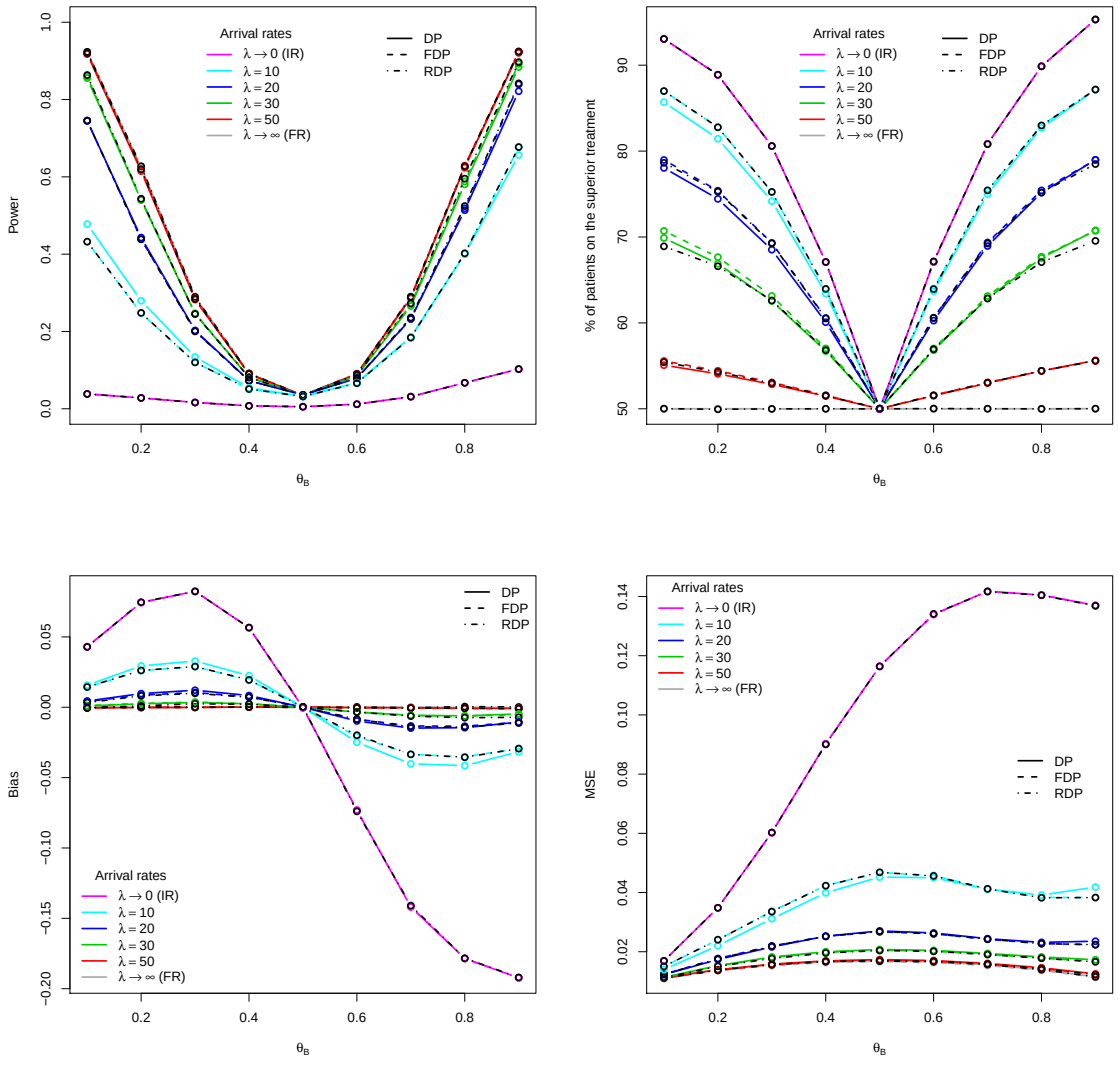


Figure 5.5.2: The changes in power (and type I error), % of patients on the superior treatment, the average bias and MSE of the treatment effect estimator for the DP, FDP and RDP designs when $n = 60$, $\theta_A = 0.5$, $\theta_B \in (0.1, 0.9)$, $\tau_i \sim \text{Exp}(\lambda)$ and $\delta = 1$ (estimated over $100,000$ simulations). IR, immediate response and FR, fixed randomisation.

Chapter 6

A Response-Adaptive Randomisation Procedure for Multi-Armed Clinical Trials with Normally Distributed Outcomes

Whereas the previous two chapters have focused on the dynamic programming approach to the bandit problem for binary endpoints within a two-arm clinical trial, this chapter considers the alternative Gittins index approach to the bandit problem for normally distributed endpoints within a multi-armed trial. Furthermore, we now consider the use of block randomisation rather than sequential randomisation.

6.1 Introduction

Response-adaptive randomisation (RAR) has been widely developed ever since the idea was first suggested by Thompson (1933) (Hu and Rosenberger, 2006). The usual motivation behind RAR is to achieve a patient benefit objective, e.g. to reduce exposure to inferior treatments by skewing the allocation towards superior treatments

based on observed responses. Incorporating such an objective into a trial design is particularly important when the disease under study is rare — in which case a substantial proportion of patients in the population will be included in the trial — and when an inferior treatment could result in a fatal outcome.

Despite the vast array of RAR procedures proposed in the literature, most of them: (a) assume *binary* responses, (b) are defined for trials with only *two treatments*, and (c) are *myopic*. However, many clinical trials have *continuous* primary outcomes and include more than two (i.e. multiple) arms. Wason and Trippa (2014) report that 39% of all multi-arm clinical trials published in four major medical journals during 2012 had normally distributed primary outcomes. Although most RAR procedures for binary responses are not easily extended to the continuous case, particularly those based on urn models (Atkinson and Biswas, 2014), several RAR procedures for continuous outcomes have been proposed (e.g. Zhu and Hu, 2009); a review of these can be found in Atkinson and Biswas (2014, Chapter 4), and Biswas and Bhattacharya (2016). Moreover, a "shortage of RAR methodology to handle cases with multiple treatments" (Zhang *et al.*, 2011) persists, despite the fact that RAR has the greatest potential for efficiency and patient benefit gains in multi-armed trials (Berry, 2011), which considerably limits its use in practice.

Furthermore, almost all procedures in the RAR literature (for binary or continuous outcomes) use only *past* observations (allocations and responses) to influence the decision for the next patient, without considering the number of patients remaining to be treated (inside or outside the trial) or the information they could provide. Such *myopic* strategies are not optimal in general (Berry and Fristedt, 1985). An optimal approach, in terms of patient benefit, is based on the multi-armed bandit problem (MABP) which considers *all* possible sequences of trial observations, and the sequence that maximises patient response is selected. As a result, the traditional dynamic programming approach used to solve the MABP is much more computation-

ally intensive than myopic procedures, which is the predominant reason why the latter have been favoured in the literature. Recent work proposing *non*-myopic bandit-based RAR procedures for binary responses includes Villar *et al.* (2015b), Williamson *et al.* (2017), and Villar and Rosenberger (2018). We will refer to *non*-myopic procedures as *forward-looking* hereafter to be consistent with the terminology used in previous papers.

Examples of forward-looking adaptive allocation rules for continuous endpoints relevant to this chapter are Coad (1991b), Wang (1991a), Coad (1995) and Smith and Villar (2018), all of which use the Gittins index for normally distributed outcomes. However, the main limitation of these designs from a clinical trials perspective is their deterministic nature. Randomisation is essential in order to remove various sources of bias and it additionally provides a basis for inference (Rosenberger and Lachin, 2016).

Motivated by the above considerations, we propose a novel bandit-based allocation rule that (a) applies to continuous outcomes, assumed to be normally distributed; (b) applies when the outcome variance is assumed unknown; (c) is defined for multi-armed trials; (d) is forward-looking and thus is orientated towards a patient benefit objective; (e) is computationally feasible, and (f) is randomised. Additionally, we investigate the impact on patient benefit of dichotomising a continuous endpoint, which is a widely adopted approach in clinical research that has received considerable attention in the literature (Royston *et al.*, 2006). A common reason for this practice is to deal with complete responses and missing data (due to death or dropout, for example) since these naturally fall into success and failure categories, respectively. However, dichotomisation comes at an efficiency cost (either a reduced power or larger sample size) (Lavin, 1981; Wason *et al.*, 2011).

Dealing with complete responses and missing data poses an extra challenge that is exclusive to the implementation of RAR in a trial. The imputation method suggested in Karrison *et al.* (2007), which is the only one that has shown moderate uptake in

practice (Wason and Jaki, 2016), imputes unobserved responses using the distribution of data collected at the *end* of the trial and therefore, the imputed data cannot be used to perform any adaptations. In this chapter, we suggest a simple modification of the procedure by Karrison *et al.* (2007) which permits the use of RAR to allocate patients dynamically during the trial.

In Section 6.2, we present our forward-looking rule for continuous endpoints with unknown variance using a simple example to illustrate its implementation. In Section 6.3, we report extensive comparative simulation studies in the context of a real phase II cancer trial. We discuss the costs of dichotomisation in Section 6.4, and present our method to accommodate missing data due to deaths, dropouts and complete responses in Section 6.5. We draw conclusions in Section 6.6.

6.2 The Forward-Looking Gittins Index (FLGI) Rule for Continuous Endpoints

We now define a RAR procedure for continuous endpoints, assumed to be normally distributed, which augments the Forward-Looking Gittins Index (FLGI) rule proposed in Villar *et al.* (2015b) for binary endpoints. Following the notation in that paper, we consider a clinical trial that will test the effectiveness of K experimental treatments against a control treatment on a sample of T patients, with K and T fixed. Patients are labelled by t ($t = 1, \ldots, T$) and treatments by k ($k = 0, \ldots, K$), where $k = 0$ denotes the control. The response of patient t allocated to treatment k is a random variable denoted by $Y_{k,t}$, now assumed to follow a normal distribution, $Y_{k,t} \sim N(\mu_k, \sigma_k^2)$. Without loss of generality, we also assume that a *larger* response is desired and that σ_k^2 is unknown.

In order to derive our FLGI rule, we need to obtain the Gittins index for the MABP associated with this trial design problem. A detailed explanation of the problem's

assumptions and its exact formulation appears in Appendix 6.7.1. The Gittins index for a treatment with posterior mean $\tilde{y}_{k,t}$ and posterior standard deviation $\tilde{s}_{k,t}$, after having observed $n_{k,t}$ responses from treatment k, $\mathcal{G}(\tilde{y}_{k,t}, \tilde{s}_{k,t}, n_{k,t})$, can be written as

$$\mathcal{G}(\tilde{y}_{k,t}, \tilde{s}_{k,t}, n_{k,t}) = \tilde{y}_{k,t} + \tilde{s}_{k,t}\mathcal{G}(0, 1, n_{k,t} + 2, d), \tag{6.2.1}$$

where $\mathcal{G}(0, 1, n_{k,t} + 2, d)$ denotes the Gittins index value of a standardised bandit problem with posterior mean 0, posterior standard deviation 1, $n_{k,t}$ observations, an implicit (prior) sample size of 2 (refer to Appendices 6.7.1 and 6.7.3 for details), and discount factor $0 \leq d < 1$. In this chapter, we choose d as recommended in Wang (1991b) (Appendix 6.7.2 provides further details).

Notice that in this case we have two unknown parameters, μ_k and σ_k^2, which we assume have the hierarchical conjugate priors $\mu_k \mid \sigma_k^2 \sim N\left(0, \frac{\sigma_k^2}{2}\right)$ and $\sigma_k^2 \sim IG\left(\frac{1}{2}, \frac{1}{2}\right)$, that is, the normal-inverse-gamma joint prior $(\mu_k, \sigma_k^2) \sim NIG\left(0, 2, \frac{1}{2}, \frac{1}{2}\right)$ when $n_{k,t} = 0$. The choice of prior and its effect on performance measures is explored in Appendix 6.7.3. As in Smith and Villar (2018), we implement the solution in (6.2.1) at a very low computational cost by calculating the values of $\mathcal{G}(0, 1, n_{k,t} + 2, d)$ in advance[1] and interpolating from Table 8.3 in Gittins *et al.* (2011, p.263). Details on how to compute these indices, first computed by Jones (1975), can be found in Gittins *et al.* (2011, Chapters 7 and 8).

In order to derive a response-adaptive rule that will sequentially randomise the next b patients among the $K + 1$ treatments at stage j $(j = 1, \ldots, J)$, given the data up to and including block $j - 1$, according to what the Gittins index rule would do, we assume that patients are enrolled in groups, or blocks, of size b over J stages, so that $J \times b = T$. Using (6.2.1) and the Gittins index rule, which states that it is optimal to allocate the treatment with the highest index value (breaking ties at random), we can compute the FLGI probabilities for the case of a normally distributed endpoint (with

[1] These values are provided in Table 6.7.1.

unknown variance) using equation (3) in Villar *et al.* (2015b). The main difference here is that the optimal action probabilities in equation (3) of Villar *et al.* (2015b) can no longer be matched to the probabilities of the (binary) outcome and must be computed for different ranges of the continuous outcome.

Example

To illustrate the proposed rule, we derive the FLGI probabilities for the simplest possible case of a two-arm trial testing a control treatment ($k = 0$) against an experimental treatment ($k = 1$) with a block of size two ($b = 2$).

For both k, we assume the following hierarchical (conjugate) prior structure at the start of the trial: $\mu_k \mid \sigma_k^2 \sim N\left(0, \frac{\sigma_k^2}{2}\right)$ and $\sigma_k^2 \sim IG\left(\frac{1}{2}, \frac{1}{2}\right)$, so that $(\mu_k, \sigma_k^2) \sim NIG\left(0, 2, \frac{1}{2}, \frac{1}{2}\right)$. Suppose further that both patients are randomly allocated to the control treatment in the first block of the trial, resulting in responses $y_{0,1} = 3.1$ and $y_{0,2} = -0.4$. Thus, the three relevant parameters required to obtain the corresponding Gittins index for the *control* treatment are: the posterior mean $\widetilde{y}_{0,2} = 0.675$, the posterior standard deviation $\widetilde{s}_{0,2} = 1.727$, and the number of observations $n_{0,2} = 2$ (see equation (6.7.1) in Appendix 6.7.1). For the *experimental* treatment, the relevant parameters are: $\widetilde{y}_{1,2} = 0$, $\widetilde{s}_{1,2} = 1$, and $n_{1,2} = 0$. From equation (6.2.1), setting $d = 0.995$ and using Table 6.7.1 of Appendix 6.7.1, the Gittins index for the control and experimental treatment, respectively, is $\mathcal{G}_0(0.675, 1.727, 2) = 0.675 + 1.727 \times 1.8126 = 3.805$ and $\mathcal{G}_1(0, 1, 0) = 0 + 1 \times 65.5848 = 65.585$.

Figure 6.2.1 illustrates how the FLGI probabilities for block two, given the data in block one, are computed via a probability tree. Given that the experimental treatment has the unique maximum Gittins index, the first patient of the second block is allocated to the experimental treatment with probability 1. When the second patient of the second block is to be allocated, we need to have observed the (random) outcome of the first patient in this block, denoted by $Y_{1,3}$, in order to update the

indices and determine the optimal action. The updated prior parameters for the experimental treatment, as a function of the observed information on this treatment and given the previous optimal action, are: $\widetilde{Y}_{1,3} = \frac{Y_{1,3}}{n_{1,3}+2}$, $\widetilde{S}_{1,3} = \left(\frac{1}{2} + \frac{Y_{1,3}^2}{n_{1,3}+2}\right)^{1/2}$, and $n_{1,3} = 1$. Thus, the index for the experimental treatment can be expressed as a function of the random outcome from patient three as follows: $\mathcal{G}_1(\widetilde{Y}_{1,3}, \widetilde{S}_{1,3}, n_{1,3} = 1) = \frac{Y_{1,3}}{3} + \left(\frac{1}{2} + \frac{Y_{1,3}^2}{3}\right)^{1/2} \mathcal{G}_1(0, 1, 3, 0.995)$, with $\mathcal{G}_1(0, 1, 3, 0.995) = 4.6049$.

For the control treatment, we have no new information and so its index remains unchanged at $\mathcal{G}_0(\widetilde{Y}_{0,3}, \widetilde{S}_{0,3}, n_{0,3}) = 3.805$. According to the Gittins index rule, it is optimal to allocate the control treatment to the second patient in the block if and only if $\mathcal{G}_1(\widetilde{Y}_{1,3}, \widetilde{S}_{1,3}, n_{1,3}) < \mathcal{G}_0(\widetilde{Y}_{0,3}, \widetilde{S}_{0,3}, n_{0,3})$, which happens when $-0.9508 < Y_{1,3} < 0.5862$. Since $Y_{1,3}$ is a standard normal random variable, this happens with probability 0.5503, that is, $\mathbb{P}(Y_{1,3} \leq 0.5862) - \mathbb{P}(Y_{1,3} \leq -0.9508) = 0.5503$. If $Y_{1,3} < -0.9508$ or $Y_{1,3} > 0.5862$, which happens with probability 0.4497, then $\mathcal{G}_1(\widetilde{Y}_{1,3}, \widetilde{S}_{1,3}, n_{1,3}) > \mathcal{G}_0(\widetilde{Y}_{0,3}, \widetilde{S}_{0,3}, n_{0,3})$ and the second patient in the second block is optimally allocated to the experimental treatment. Notice that if $Y_{1,3} = -0.9508$ or $Y_{1,3} = 0.5862$, the index values are equal and it is optimal to allocate any of the two treatments. In theory, this would happen with probability 0 since $Y_{k,t}$ is a continuous variable. However, in practice, if this were to happen, we would randomise with probability 0.5. Hence, the normal FLGI procedure would randomise both patients in this block to receive the experimental treatment with probability $\frac{1+(1\times 0.4497)}{2} = 0.7249$, and the control treatment with probability $\frac{0+(1\times 0.5503)}{2} = 0.2751$. Continuing this example for larger block sizes using Monte Carlo simulation, the allocation probabilities to the experimental and control arm, respectively, are $(0.6565, 0.3435)$ for $b = 3$, $(0.5151, 0.4849)$ for $b = 4$, $(0.4370, 0.5630)$ for $b = 5$, and $(0.3051, 0.6949)$ for $b = 10$.

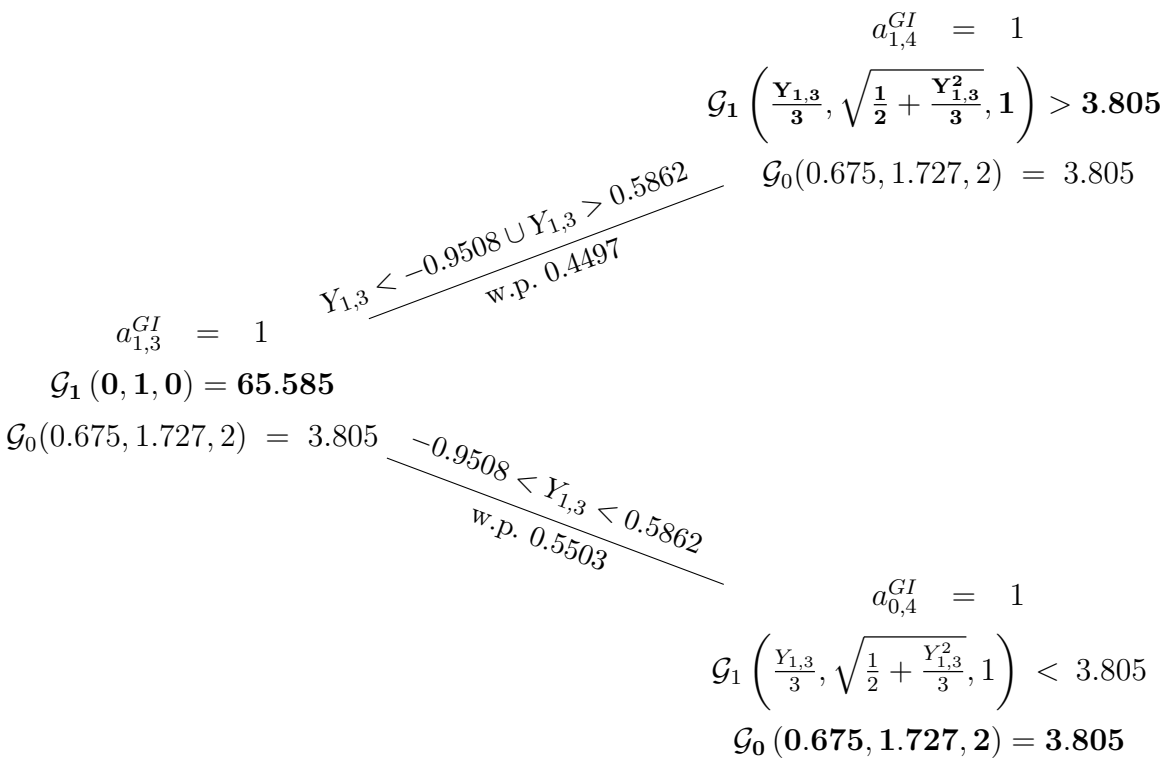


Figure 6.2.1: The FLGI rule and a probability tree of all trial histories using the Gittins index rule when $K + 1 = 2$, $b = 2$, $d = 0.995$, the outcome $Y_{k,t}$ is normally distributed with unknown mean and variance, and parameters $(\widetilde{y}_{k,2}, \widetilde{s}_{k,2}, n_{k,2})$ are given by $(0.675, 1.727, 2)$ for $k = 0$ and $(0, 1, 0)$ for $k = 1$. Bold text indicates the allocated treatment under the Gittins index rule $\{a^{GI}_{k,t}\}$. Note that the FLGI probabilities in this case are 0.7249 and 0.2751 for the experimental and control arm, respectively. (For simplicity of the illustration, we have omitted the branch corresponding to the cases $Y_{1,3} = -0.9508$ or $Y_{1,3} = 0.5862$ since, theoretically, this would happen with probability 0).

6.3 Simulation Study

6.3.1 Alternative Designs and Performance Measures

Next, we will report simulations that compare the FLGI for a normally distributed endpoint (with unknown variance) against the following existing randomisation procedures:

(1) Equal Randomisation[2] (ER), where each patient is randomly allocated to one

[2]Note that this is referred to as *fixed randomisation* in previous chapters. The two terms are taken to be synonymous throughout this thesis.

of the $K+1$ arms with equal probability, $1/(K+1)$. ER is predominant in practice (implemented, for example, by permuted-block randomisation), thus it will be used as a reference to compare all designs.

(2) Modified Zhang and Rosenberger (MZR), introduced by Zhang and Rosenberger (2006) and later modified by Biswas and Bhattacharya (2009) to allow for negative mean responses. The rule aims at minimising the total of inverse mean responses, that is, $n_{0,T}/\mu_0 + n_{1,T}/\mu_1$. This design results in the following optimal allocation proportion ρ^*:

$$\rho^* = \begin{cases} c & \text{if } \{\mu_0, \mu_1 > 0 \text{ and } \rho_c < c\} \text{ or } \left\{\mu_0, \mu_1 < 0,\ \frac{\sigma_0}{\sigma_1} > \sqrt{\frac{\mu_1}{\mu_0}}\right\} \text{ or } \{\mu_0 < 0,\ \mu_1 > 0\}, \\ \rho_c & \text{if } \{\mu_0, \mu_1 > 0,\ c \leq \rho_c \leq 1-c\}, \\ 1-c & \text{if } \{\mu_0, \mu_1 > 0,\ \rho_c > 1-c\} \text{ or } \left\{\mu_0, \mu_1 < 0,\ \frac{\sigma_0}{\sigma_1} < \sqrt{\frac{\mu_1}{\mu_0}}\right\} \text{ or } \{\mu_0 > 0,\ \mu_1 < 0\}, \end{cases}$$

where $\rho_c = \sigma_0\sqrt{\mu_0}\,/\left(\sigma_0\sqrt{\mu_0} + \sigma_1\sqrt{\mu_1}\right)$ and $c \in [0, 1/2]$. The initial parameter estimates are obtained by allocating the first n^{ER} patients using ER. After that, estimates of the unknown parameters μ_k and σ_k are sequentially updated based on the current data available.

(3) Constrained Gittins Index (GI) Rule is a procedure based on Gittins indices proposed by Wang (1991a) and further studied by Coad (1991b, 1995). However, unlike the FLGI, Constrained GI is not implemented in terms of probabilities, and hence is not randomised. This is a practical limitation and explains why Constrained GI has been neglected as a comparator within the RAR literature. The rule is defined as follows: if $n^c_{0,t} < n_{1,t}$, allocate the next patient to arm 0; if $n^c_{1,t} < n_{0,t}$, allocate the next patient to arm 1; else, allocate the next patient to the treatment with the largest Gittins index (randomising if they are equal). The parameter $c \geq 1$ is a *tuning* parameter; $c = 1$ corresponds to ER, and the Gittins index is eventually recovered as $c \to \infty$. Following Wang (1991a), we fix $c = 2$ in our simulations.

(4) Thompson Sampling (TS) randomises patients to arms based on their posterior

probability of being the "best" arm. Specifically, we consider a version of Thompson sampling suggested by Thall and Wathen (2007), where the probability of allocating treatment k to patients in block j is computed as

$$\frac{\mathbb{P}\left(\max_i \ \mu_i = \mu_k \mid \widetilde{\mathbf{x}}_{(j-1)b}\right)^c}{\sum_{k=0}^{K} \mathbb{P}\left(\max_i \ \mu_i = \mu_k \mid \widetilde{\mathbf{x}}_{(j-1)b}\right)^c},$$

where $\widetilde{\mathbf{x}}_t = (\widetilde{y}_{0,t}, \widetilde{s}_{0,t}, n_{0,t}, \ldots, \widetilde{y}_{K,t}, \widetilde{s}_{K,t}, n_{K,t})$ and $c = (j-1)b/2T$ is a *tuning* parameter that recovers ER when $c = 0$ and TS when $c = 1$.

(5) Trippa *et al.* (2012) Procedure (TP) randomises patients similarly to TS, but also protects allocation to the control arm. We have implemented TP as in Villar *et al.* (2015b)[3].

(6) Controlled FLGI (CFLGI) is a variant of the FLGI design proposed in Villar *et al.* (2015b) which, similarly to TP, protects the allocation to the control arm by ensuring that the corresponding allocation probability is always at least $1/(K+1)$.

(7) Gwise *et al.* (2011) propose a design for comparing $K+1$ arms with heteroscedasticity. After an initial ER phase, patient $t+1$ is allocated to arm k with probability

$$\frac{\hat{\sigma}^2_{k,t}/n_{k,t}}{\hat{\sigma}^2_{0,t}/n_{0,t} + \cdots + \hat{\sigma}^2_{K,t}/n_{K,t}},$$

where $\hat{\sigma}^2_{k,t}$ is the estimated sample variance of the first $n_{k,t}$ responses on arm k.

Note that MZR and Constrained GI are fully sequential and will only be implemented in the two-armed case (see Coad (1995) for the multi-arm version of Constrained GI). TP and CFLGI apply only to the multi-armed case. For all of the rules which require specification of a joint prior distribution on μ_k and σ^2_k, we take the same approach as with the FLGI. For the index-based designs, a discount factor of $d = 0.995$ is used, and the allocation probabilities defined in the FLGI designs, TS and TP are computed using their empirical estimates from 100 Monte Carlo repli-

[3]Refer to Section 2.1.3 for a description of TP.

cates. Additionally, we implement the doubly adaptive biased coin design by Hu and Zhang (2004)[4] with the target allocation proportions taken to be the corresponding FLGI probabilities for $b = T$ under the null and alternative hypotheses, $\mathcal{H}_0$ and $\mathcal{H}_1$, defined below.

To evaluate the performance of all designs, we consider patient benefit and the usual inferential measures. The former includes: (a) the expected proportion of patients in the trial allocated to the superior treatment, $\mathbb{E}(p^*)$, and (b) the percentage change in expected total outcome for rule r (ETO_r) relative to the theoretical expected total outcome for ER (ETO_{ER}), computed as $100 \times (ETO_r - ETO_{\text{ER}})/ETO_{\text{ER}}$ and denoted in the tables of results by Rel$ETO\%$. For the inferential measures, we focus on standard operating characteristics, including: power, $1 - \beta$; type I error rate, α; and bias in the maximum likelihood estimator of the treatment effect, $\mathbb{E}(\hat{\Delta} - \Delta)$, with $\Delta = \mu_k - \mu_0$ and $\hat{\Delta} = (\hat{\mu}_k - \hat{\mu}_0)$. For the multi-armed case, we report both the marginal power (i.e. power to reject $\mathcal{H}_{0,k^*}$, where k^* is the best arm) and the bias for the best experimental arm under $\mathcal{H}_1$. Note that under $\mathcal{H}_0$, we take k^* to be the control arm.

We consider the following hypotheses: $\mathcal{H}_0 : \mu_0 = \mu_k \ \forall \ k$ versus the one-sided alternatives, $\mathcal{H}_{1,k} : \mu_0 < \mu_k$ for some $k > 0$ considered the best arm. We will use the test statistic $T_k = \left(\overline{Y}_k - \overline{Y}_0\right) \Big/ \sqrt{\frac{\hat{\sigma}_k^2}{n_{k,T}} + \frac{\hat{\sigma}_0^2}{n_{0,T}}}$ for $k = 1, \ldots, K$, where $\overline{Y}_k$ and $\hat{\sigma}_k^2$ are the sample mean and sample variance, respectively, of arm k at the end of the trial. In the multi-armed case, we consider the joint distribution of $T_1, \ldots, T_K$ and use a critical value, $t_{1-\alpha}$, to achieve a family-wise type I error rate (FWER) close to the specified α, where FWER is defined as the probability of obtaining at least one false positive, or type I error, within the family of null hypotheses $\mathcal{H}_0$.

[4]Refer to (2.1.1) for details.

6.3.2 A Two-Armed Trial

To motivate this scenario, we use the example in Karrison *et al.* (2007) of a two-armed phase II cancer trial, in which the primary endpoint is the ratio of tumour size at the time of follow-up to that at baseline for patient t under treatment k, that is, the change in tumour size, denoted by $C_{k,t}$. After a log-transformation, $C_{k,t}$ is continuous and approximately normally distributed, as shown by Lavin (1981). In keeping with our assumption that a larger outcome is desirable, we add a minus sign to re-express the endpoint as a measure of tumour reduction. Under the assumption that $Y_{0,t} = -\log(C_{0,t}) \sim N(0.155, 0.64^2)$ and $Y_{1,t} = -\log(C_{1,t}) \sim N(0.529, 0.64^2)$, the total sample size required to detect this treatment difference with approximately 80% power at the $\alpha = 0.05$ significance level and assuming complete observations is $T = 72$.

Results

Table 6.3.1 displays the results from $50,000$ replications of the trial when we assume unknown variance. As expected, under $\mathcal{H}_0$ all the designs are equal in terms of patient benefit (Rel$ETO\% \approx 0$ and $\mathbb{E}(p^*) \approx 0.50$). The main difference between designs under the null is the variability of the allocations, represented by the standard deviations (s.d.) of p^*, with ER and FLGI (for $b = 1$) being the least and most variable, respectively. As the block size increases, changes in the allocation probabilities are based on more data and the FLGI becomes less variable. The index-based procedures tend to be more variable because they aim at maximising patient response. For example, the Constrained GI also has a large variability which is comparable to that of the FLGI. For the MZR design, the variability of the allocations decreases as the size of the initial ER period, n^{ER}, increases. The variability of the FLGI is also markedly reduced when implemented using Hu and Zhang (2004), labelled as FLGI-HZ in Table 6.3.1. In terms of the bias of the treatment effect estimator, all are (on

average) unbiased under $\mathcal{H}_0$. Note that we have used adjusted t-critical values to control type I error rates for all designs following the approach used in Smith and Villar (2018). The (unreported) type I error inflation incurred for the FLGI when using the usual $t_{0.95}$ critical value is approximately 11% for $b = 1$ and it decreases as the block size grows, as expected. A similar level and pattern of inflation occurs for TS.

The results under $\mathcal{H}_1$, in which we are testing for superiority of arm 1 (the experimental arm), show more contrasts amongst designs. First, we focus on the FLGI design and the effect of varying the block size on the power versus patient benefit trade-off. When $b = 1$, the FLGI design is statistically identical to the fully sequential Gittins index rule and so favours patient response. At the other extreme, when $b = T$, the FLGI design is equivalent to ER and therefore favours power. Thus, consistent with the findings for the binary case, Table 6.3.1 shows that as b increases under $\mathcal{H}_1$, the patient benefit measures (and corresponding standard deviations) decrease, whilst the power increases (at a faster rate) which illustrates the natural tension between these two conflicting goals. This relationship is depicted visually in Figure 6.5.1 for $T = 128$.

In terms of the patient benefit measures, the index-based designs (namely the FLGI and Constrained GI) perform the best out of all the designs considered. Relative to ER, for a moderate block size of $b = 9$, the FLGI allocates approximately 34% more patients to the superior treatment (equivalent to 25 patients). Moreover, the expected total tumour size reduction is just over 37% greater than that obtained when using ER. Even for a large block size of $b = 36$, the FLGI allocates approximately 21% more patients to arm 1 and achieves an expected total tumour size reduction 23% larger than ER. All other block sizes for the FLGI have a total tumour size reduction at least 30% greater than ER, on average. The Constrained GI is shown to perform similarly to the FLGI when $b = 9$. TS has a total tumour size reduction rate of at

least 20% greater than ER for small b, on average, whereas MZR falls below this for all n^{ER}.

As mentioned above, the cost of these patient benefit gains is a severe reduction in the power compared to that of ER. However, this is ameliorated as b increases or by implementing the FLGI probabilities using Hu and Zhang (2004). The ER design attains an unbiased treatment effect estimator, as expected, with the largest relative bias exhibited by the FLGI design when $b = 1$ (i.e. the GI design). This makes sense because this is the design with the biggest imbalance in favour of arm 1. As a result, $\hat{\mu}_0$ will be substantially underestimated giving rise to an overestimated $\hat{\Delta}$ (and positive bias of treatment effect). As b increases, and consequently the number of observations on arm 0 increases, the bias (and associated standard deviations) of the treatment effect estimator decreases.

These results emphasise the very important point that, in a two-armed setting, none of the designs are uniformly better than the others for *every* performance measure since each design is tailored towards a different competing objective. This makes direct comparisons between such designs infeasible and motivates our main interest in the multi-armed case.

Table 6.3.1 also shows the results attained by the FLGI rule when assuming the correct variance in both arms (see FLGI-known). As expected, FLGI-known marginally outperforms the FLGI with unknown variance in terms of patient benefit (and reduces the power) due to the additional uncertainty present in the latter. However, in practice, this is unrealistic since the true variance of the outcome is seldom known at the start of a trial. Therefore, in Tables 6.3.2 and 6.3.3, we illustrate the effect of assuming an *incorrect* variance (on one, or both, of the arms) on the performance of the FLGI relative to when assuming an unknown variance. Although misspecifying the variance does not always have a negative impact on the results, and the performance may be comparable to that when assuming an unknown variance (as in Scenarios

(i)–(iv)), it is important to be aware that it can sometimes lead to a considerable loss in patient benefit. This is evident in Scenario (vii) of Table 6.3.3, for example, where 6.4% fewer patients are allocated to the superior treatment (for $b = 1$) as a consequence of underestimating σ_1^2. As such, the robustness and flexibility attained by the FLGI with unknown variance makes this design more suited to practice.

		$\mu_0 = \mu_1 = 0.155$					$\mu_0 = 0.155,\ \mu_1 = 0.529$			
Design		$t_{1-\alpha}$	α	$\mathbb{E}(p^*)$ (s.d.)	RelETO% (s.d.)	Bias (s.d.)	$1-\beta$	$\mathbb{E}(p^*)$ (s.d.)	RelETO% (s.d.)	Bias (s.d.)
ER	$b = 1$	1.654	0.0518	0.4999 (0.06)	-0.19 (5.45)	-0.0005 (0.15)	0.7884	0.5005 (0.06)	0.20 (5.66)	-0.0013 (0.15)
FLGI-known	$b = 1$	1.991	0.0526	0.4997 (0.33)	-0.06 (5.42)	-0.0004 (0.49)	0.2290	0.8823 (0.16)	41.69 (7.01)	0.2167 (0.45)
	$b = 2$	1.969	0.0505	0.5002 (0.32)	0.34 (5.42)	-0.0003 (0.44)	0.2701	0.8777 (0.16)	41.27 (6.85)	0.1870 (0.41)
	$b = 6$	1.911	0.0481	0.4987 (0.29)	-0.03 (5.45)	0.0010 (0.34)	0.3599	0.8605 (0.14)	39.38 (6.59)	0.1147 (0.30)
	$b = 9$	1.864	0.0492	0.4983 (0.28)	-0.15 (5.43)	0.0008 (0.30)	0.4235	0.8483 (0.13)	38.14 (6.53)	0.0843 (0.26)
	$b = 18$	1.766	0.0513	0.5017 (0.24)	0.14 (5.44)	-0.0010 (0.23)	0.5653	0.8074 (0.12)	33.72 (6.30)	0.0389 (0.20)
	$b = 36$	1.682	0.0495	0.5000 (0.19)	0.26 (5.45)	0.0001 (0.17)	0.7124	0.7139 (0.09)	23.25 (5.98)	0.0087 (0.17)
FLGI	$b = 1$	2.1820	0.0525	0.5013 (0.29)	-0.15 (5.43)	0.0013 (0.29)	0.3289	0.8712 (0.12)	40.62 (6.39)	0.0955 (0.27)
	$b = 2$	2.159	0.0497	0.5016 (0.28)	0.21 (5.45)	0.0014 (0.28)	0.3432	0.8651 (0.12)	39.95 (6.41)	0.0902 (0.26)
	$b = 6$	2.118	0.0477	0.4985 (0.27)	0.18 (5.42)	-0.0008 (0.26)	0.3790	0.8521 (0.12)	38.48 (6.36)	0.0801 (0.25)
	$b = 9$	2.045	0.0514	0.5011 (0.26)	-0.01 (5.41)	0.0019 (0.25)	0.4236	0.8412 (0.12)	37.13 (6.36)	0.0698 (0.24)
	$b = 18$	1.898	0.0517	0.5008 (0.24)	-0.04 (5.42)	0.0005 (0.22)	0.5277	0.8047 (0.12)	33.30 (6.23)	0.0356 (0.20)
	$b = 36$	1.733	0.0505	0.4997 (0.18)	0.01 (5.43)	-0.0009 (0.18)	0.6973	0.7128 (0.09)	23.23 (6.00)	0.0097 (0.17)
FLGI-HZ ($\gamma = 2$)	$b = 1$	1.658	0.0509	0.5000 (0.05)	0.12 (5.43)	0.0008 (0.15)	0.6510	0.7784 (0.04)	30.46 (5.51)	0.0001 (0.18)
	$b = 2$	1.660	0.0509	0.5003 (0.05)	-0.14 (5.44)	0.0004 (0.15)	0.6499	0.7786 (0.04)	30.61 (5.54)	-0.0007 (0.18)
	$b = 6$	1.688	0.0487	0.5001 (0.05)	0.16 (5.42)	0.0000 (0.15)	0.6417	0.7781 (0.04)	30.40 (5.54)	-0.0001 (0.18)
	$b = 9$	1.661	0.0529	0.4994 (0.05)	0.43 (5.43)	0.0017 (0.15)	0.6501	0.7777 (0.04)	30.27 (5.52)	-0.0004 (0.18)
	$b = 18$	1.684	0.0490	0.4999 (0.05)	0.11 (5.41)	0.0003 (0.15)	0.6570	0.7644 (0.04)	28.99 (5.54)	-0.0011 (0.18)
	$b = 36$	1.665	0.0516	0.5000 (0.05)	0.10 (5.41)	0.0017 (0.15)	0.7779	0.5865 (0.05)	9.57 (5.64)	0.0003 (0.15)
TS	$b = 1$	1.751	0.0496	0.4999 (0.11)	-0.1108 (5.44)	-0.0001 (0.17)	0.7425	0.6961 (0.11)	21.35 (6.14)	0.0302 (0.19)
	$b = 2$	1.739	0.0497	0.4997 (0.11)	-0.0431 (5.41)	-0.0016 (0.17)	0.7479	0.6934 (0.11)	21.27 (6.11)	0.0290 (0.19)
	$b = 6$	1.741	0.0513	0.4994 (0.11)	0.3098 (5.44)	-0.0001 (0.17)	0.7489	0.6825 (0.10)	19.88 (6.14)	0.0257 (0.18)
	$b = 9$	1.729	0.0499	0.5000 (0.10)	0.1311 (5.42)	0.0001 (0.17)	0.7547	0.6747 (0.10)	18.95 (6.13)	0.0229 (0.18)
	$b = 18$	1.722	0.0494	0.5008 (0.10)	0.4446 (5.42)	0.0013 (0.16)	0.7602	0.6509 (0.10)	16.40 (6.11)	0.0184 (0.17)
	$b = 36$	1.697	0.0507	0.4999 (0.08)	0.4332 (5.41)	0.0013 (0.16)	0.7726	0.6040 (0.10)	11.45 (6.07)	0.0095 (0.16)
CGI ($c = 2$)		1.887	0.0496	0.4871 (0.28)	0.19 (5.42)	0.0004 (0.24)	0.4298	0.8294 (0.11)	37.59 (6.19)	0.0340 (0.21)
MZR	$n^{\text{ER}} = 2$	1.794	0.0516	0.5005 (0.19)	0.2794 (5.41)	0.0002 (0.19)	0.7471	0.6569 (0.12)	17.15 (5.76)	0.0229 (0.17)
	$n^{\text{ER}} = 6$	1.780	0.0507	0.4998 (0.17)	0.3487 (5.43)	0.0001 (0.18)	0.7632	0.6414 (0.10)	15.47 (5.49)	0.0202 (0.16)
	$n^{\text{ER}} = 11$	1.751	0.0508	0.5001 (0.14)	-0.2534 (5.41)	0.0007 (0.17)	0.7755	0.6173 (0.08)	12.86 (5.29)	0.0155 (0.16)
Gwise	$n^{\text{ER}} = 2$	1.877	0.0495	0.4997 (0.13)	-0.08 (5.43)	0.0012 (0.18)	0.7193	0.4999 (0.13)	-0.11 (6.47)	-0.0013 (0.18)
	$n^{\text{ER}} = 6$	1.697	0.0482	0.5003 (0.06)	-0.08 (5.45)	-0.0000 (0.15)	0.7833	0.5005 (0.06)	0.02 (5.67)	-0.0013 (0.15)
	$n^{\text{ER}} = 11$	1.705	0.0492	0.4999 (0.06)	0.07 (5.41)	0.0007 (0.15)	0.7837	0.5000 (0.06)	-0.20 (5.66)	0.0000 (0.15)

Table 6.3.1: Comparison of performance measures for a two-armed trial using different designs when the variance is assumed unknown (with the exception of FLGI-known) and $T = 72$, averaged over 50,000 trial replications. Note that the true variance of the response is $\sigma_k^2 = 0.64^2$ for $k \in \{0, 1\}$.

	$\mu_0 = \mu_1 = 0.155$					$\mu_0 = 0.155,\ \mu_1 = 0.529$			
b	$t_{1-\alpha}$	α	$\mathbb{E}(p^*)$ (s.d.)	RelETO% (s.d.)	Bias (s.d.)	$1-\beta$	$\mathbb{E}(p^*)$ (s.d.)	RelETO% (s.d.)	Bias (s.d.)
(i) FLGI-known with $\sigma_0^2 = \sigma_1^2 = \frac{1}{2} \times 0.64^2$									
1	2.001	0.0509	0.4983 (0.27)	-0.04 (3.84)	0.0013 (0.26)	0.4722	0.9215 (0.07)	46.01 (4.34)	0.1430 (0.29)
2	1.977	0.0505	0.5010 (0.26)	0.17 (3.83)	-0.0007 (0.24)	0.5422	0.9142 (0.07)	45.22 (4.32)	0.1209 (0.26)
6	1.925	0.0525	0.4986 (0.24)	0.12 (3.85)	0.0009 (0.20)	0.6642	0.8952 (0.07)	43.29 (4.33)	0.0771 (0.21)
9	1.887	0.0518	0.5002 (0.23)	0.07 (3.82)	-0.0004 (0.18)	0.7241	0.8814 (0.07)	41.77 (4.30)	0.0545 (0.18)
18	1.798	0.0509	0.4999 (0.21)	0.14 (3.83)	0.0010 (0.15)	0.8406	0.8370 (0.07)	36.90 (4.29)	0.0224 (0.14)
36	1.698	0.0500	0.5006 (0.16)	-0.03 (3.83)	-0.0005 (0.12)	0.9341	0.7331 (0.06)	25.53 (4.19)	0.0050 (0.12)
(ii) FLGI with $\sigma_0^2 = \sigma_1^2 = \frac{1}{2} \times 0.64^2$									
1	2.280	0.0502	0.5014 (0.30)	-0.04 (3.84)	0.0017 (0.21)	0.4301	0.9177 (0.07)	45.60 (4.33)	0.0641 (0.20)
2	2.209	0.0484	0.4987 (0.29)	0.04 (3.83)	-0.0013 (0.20)	0.4713	0.9131 (0.07)	45.15 (4.32)	0.0602 (0.20)
6	2.132	0.0491	0.4987 (0.28)	0.01 (3.81)	-0.0017 (0.19)	0.5520	0.9008 (0.07)	43.88 (4.35)	0.0537 (0.19)
9	2.080	0.0505	0.4991 (0.27)	0.04 (3.84)	-0.0007 (0.18)	0.5958	0.8894 (0.07)	42.55 (4.33)	0.0427 (0.18)
18	1.897	0.0516	0.5004 (0.24)	0.04 (3.84)	-0.0004 (0.16)	0.7604	0.8466 (0.07)	37.95 (4.29)	0.0186 (0.15)
36	1.751	0.0501	0.5000 (0.19)	-0.22 (3.83)	-0.0002 (0.12)	0.9110	0.7401 (0.06)	26.33 (4.18)	0.0035 (0.12)
(iii) FLGI-known with $\sigma_0^2 = \sigma_1^2 = 2 \times 0.64^2$									
1	1.940	0.0505	0.4998 (0.37)	-0.04 (7.69)	0.0024 (0.81)	0.1517	0.8106 (0.27)	34.25 (10.54)	0.2656 (0.74)
2	1.924	0.0481	0.4997 (0.36)	0.13 (7.64)	-0.0015 (0.72)	0.1719	0.8116 (0.26)	34.29 (10.29)	0.2279 (0.66)
6	1.865	0.0484	0.4999 (0.33)	-0.28 (7.70)	-0.0008 (0.54)	0.2233	0.8018 (0.23)	33.06 (9.85)	0.1441 (0.48)
9	1.790	0.0529	0.4998 (0.31)	-0.21 (7.69)	-0.0009 (0.45)	0.2730	0.7935 (0.21)	32.21 (9.55)	0.1068 (0.40)
18	1.747	0.0514	0.4998 (0.27)	0.49 (7.70)	0.0006 (0.33)	0.3454	0.7593 (0.18)	28.34 (9.10)	0.0487 (0.30)
36	1.677	0.0497	0.5003 (0.20)	-0.03 (7.66)	0.0050 (0.25)	0.4532	0.6837 (0.14)	20.27 (8.53)	0.0108 (0.24)
(iv) FLGI with $\sigma_0^2 = \sigma_1^2 = 2 \times 0.64^2$									
1	2.386	0.0518	0.5013 (0.32)	-0.04 (7.69)	0.0009 (0.45)	0.1944	0.8117 (0.21)	34.37 (9.55)	0.1326 (0.41)
2	2.323	0.0501	0.4975 (0.31)	-0.21 (7.64)	-0.0037 (0.43)	0.2047	0.8088 (0.21)	33.57 (9.50)	0.1292 (0.4)
6	2.255	0.0502	0.5008 (0.29)	0.05 (7.68)	0.0000 (0.40)	0.2184	0.7945 (0.20)	32.22 (9.37)	0.1117 (0.38)
9	2.138	0.0525	0.5004 (0.28)	-0.07 (7.69)	0.0016 (0.38)	0.2413	0.7827 (0.19)	31.01 (9.27)	0.0936 (0.35)
18	1.888	0.0514	0.4974 (0.25)	-0.02 (7.63)	-0.0028 (0.32)	0.3346	0.7511 (0.17)	27.18 (8.97)	0.0517 (0.29)
36	1.753	0.0480	0.5002 (0.19)	0.13 (7.69)	-0.0012 (0.25)	0.4470	0.6748 (0.13)	19.11 (8.48)	0.0145 (0.24)

Table 6.3.2: Comparing the performance measures of FLGI-known, when the variance is *incorrectly* assumed to be 0.64^2, against those obtained from FLGI when the variance is assumed unknown (but with an initial estimate, $\widetilde{s}_{k,0}^2$, of 0.64^2). The true variance of the response is actually half or double 0.64^2, as indicated. These results are averaged over $50,000$ replications for a two-armed trial of size $T = 72$.

	$\mu_0 = \mu_1 = 0.155$					$\mu_0 = 0.155,\ \mu_1 = 0.529$			
b	$t_{1-\alpha}$	α	$\mathbb{E}(p^*)$ (s.d.)	RelETO% (s.d.)	Bias (s.d.)	$1-\beta$	$\mathbb{E}(p^*)$ (s.d.)	RelETO% (s.d.)	Bias (s.d.)
(v) FLGI-known with $\sigma_0^2 = 0.64^2$, $\sigma_1^2 = \frac{1}{2} \times 0.64^2$									
1	1.9900	0.0502	0.4284 (0.30)	0.03 (4.61)	0.1030 (0.38)	0.1828	0.9203 (0.09)	45.84 (4.32)	0.2522 (0.41)
2	1.9710	0.0487	0.4358 (0.29)	-0.35 (4.60)	0.0902 (0.35)	0.2200	0.9102 (0.09)	44.85 (4.32)	0.2038 (0.37)
6	1.9170	0.0487	0.4587 (0.27)	0.26 (4.63)	0.0583 (0.28)	0.3399	0.8857 (0.09)	42.06 (4.35)	0.1271 (0.28)
9	1.8730	0.0508	0.4666 (0.26)	-0.08 (4.65)	0.0477 (0.24)	0.4222	0.8703 (0.09)	40.49 (4.36)	0.0952 (0.24)
18	1.7890	0.0532	0.4836 (0.23)	0.09 (4.68)	0.0280 (0.19)	0.5893	0.8240 (0.09)	35.37 (4.43)	0.0427 (0.18)
36	1.7060	0.0533	0.4992 (0.18)	0.07 (4.70)	0.0131 (0.15)	0.7697	0.7237 (0.08)	24.42 (4.56)	0.0117 (0.15)
(vi) FLGI with $\sigma_0^2 = 0.64^2$, $\sigma_1^2 = \frac{1}{2} \times 0.64^2$									
1	2.425	0.0489	0.4822 (0.31)	-0.07 (4.72)	0.0341 (0.26)	0.2655	0.8897 (0.11)	42.51 (4.58)	0.1104 (0.26)
2	2.331	0.0524	0.4789 (0.30)	-0.07 (4.69)	0.0323 (0.25)	0.2969	0.8833 (0.11)	41.94 (4.59)	0.1044 (0.25)
6	2.266	0.0485	0.4752 (0.28)	0.05 (4.74)	0.0282 (0.24)	0.3431	0.8674 (0.11)	40.08 (4.62)	0.0927 (0.24)
9	2.185	0.0510	0.4722 (0.28)	-0.37 (4.72)	0.0263 (0.23)	0.3861	0.8562 (0.11)	38.81 (4.62)	0.0813 (0.23)
18	1.977	0.0489	0.4721 (0.25)	0.06 (4.75)	0.0197 (0.19)	0.5261	0.8158 (0.10)	34.69 (4.60)	0.0418 (0.19)
36	1.778	0.0511	0.4786 (0.19)	-0.04 (4.72)	0.0120 (0.15)	0.7429	0.7226 (0.09)	24.28 (4.60)	0.0111 (0.15)
(vii) FLGI-known with $\sigma_0^2 = 0.64^2$, $\sigma_1^2 = 2 \times 0.64^2$									
1	1.981	0.0489	0.5892 (0.34)	0.30 (6.45)	-0.1878 (0.65)	0.2669	0.7882 (0.29)	31.44 (11.25)	0.0885 (0.64)
2	1.941	0.0524	0.5747 (0.34)	0.33 (6.51)	-0.1584 (0.59)	0.3005	0.7958 (0.27)	32.39 (10.90)	0.0819 (0.55)
6	1.855	0.0519	0.5517 (0.31)	0.01 (6.53)	-0.1035 (0.45)	0.3741	0.8034 (0.23)	33.30 (10.18)	0.0586 (0.40)
9	1.821	0.0482	0.5422 (0.30)	-0.03 (6.56)	-0.0806 (0.38)	0.4119	0.7995 (0.21)	32.71 (9.84)	0.0454 (0.33)
18	1.735	0.0489	0.5194 (0.26)	-0.34 (6.59)	-0.0469 (0.28)	0.5105	0.7758 (0.17)	30.16 (9.15)	0.0197 (0.24)
36	1.665	0.0475	0.5007 (0.20)	-0.04 (6.63)	-0.0225 (0.22)	0.6098	0.6982 (0.12)	21.74 (8.20)	0.0004 (0.20)
(viii) FLGI with $\sigma_0^2 = 0.64^2$, $\sigma_1^2 = 2 \times 0.64^2$									
1	2.238	0.0517	0.5204 (0.32)	-0.05 (6.70)	-0.0480 (0.38)	0.2806	0.8522 (0.19)	38.78 (9.50)	0.0774 (0.34)
2	2.207	0.0484	0.5249 (0.31)	0.24 (6.65)	-0.0433 (0.37)	0.2885	0.8507 (0.18)	38.63 (9.40)	0.0710 (0.32)
6	2.095	0.0509	0.5290 (0.29)	-0.06 (6.69)	-0.0377 (0.34)	0.3353	0.8405 (0.17)	37.31 (9.28)	0.0637 (0.30)
9	2.039	0.0491	0.5297 (0.28)	-0.07 (6.71)	-0.0367 (0.33)	0.3554	0.8299 (0.16)	36.31 (9.14)	0.0501 (0.28)
18	1.839	0.0497	0.5284 (0.25)	0.05 (6.72)	-0.0296 (0.27)	0.4708	0.7943 (0.14)	32.01 (8.71)	0.0220 (0.23)
36	1.700	0.0489	0.5222 (0.19)	0.15 (6.70)	-0.0182 (0.21)	0.6050	0.7017 (0.11)	21.95 (8.05)	0.0030 (0.20)

Table 6.3.3: Continuation of Table 6.3.2, except now the true variances are *heterogeneous*, as indicated.

6.3.3 A Multi-Armed Trial

We now use the phase II cancer trial setting described in Karrison *et al.* (2007) as a case study. The primary endpoint is again the change in tumour size from baseline to eight weeks. Patients were randomly assigned to one of three treatment arms: 150

mg of erlotinib plus placebo; 150 mg of erlotinib plus 200 mg of sorafenib; or 150 mg of erlotinib plus 400 mg of sorafenib. We will refer to these as the control, low dose and high dose, respectively.

Based on data from previous trials, the log ratio of tumour sizes is assumed to have a mean of 0.05 for the control ($k = 0$), -0.07 for the low dose ($k = 1$) and -0.13 for the high dose ($k = 2$), with a common standard deviation of 0.346. To be consistent with our earlier assumption that larger responses are desirable, we instead consider tumour reduction. Therefore, we assume that $Y_{0,t} \sim N(-0.05, 0.346^2)$, $Y_{1,t} \sim N(0.07, 0.346^2)$ and $Y_{2,t} \sim N(0.13, 0.346^2)$. We simulate a trial of size $T = 120$, which should have at least 80% power using a one-sided test at $\alpha = 0.10$ when no correction for multiplicity is considered. In our simulations, we will ensure a one-sided test at the $\alpha = 0.10$ FWER level, and since we adjust for multiplicity, the power will fall slightly below 80%, illustrating the effect of correcting for multiplicity on power.

Results

Under the null, the only relevant difference amongst designs is the variability of resulting allocations, with the rules performing the best in terms of patient benefit being the most variable. Results under the alternative hypothesis are illustrated in Figure 6.3.1 and provided in full (for both $\mathcal{H}_0$ and $\mathcal{H}_1$) in Table 6.3.4. Figure 6.3.1 shows a star plot summarising the key features of each design (for blocks 1, 15, 40, and 60) where the most desirable values lie towards the *outer* edge of the star plot with the least favourable values towards the centre. We see that ER performs very well with respect to power, average bias and variability, but poorly with respect to patient benefit for all block sizes, whilst in contrast the FLGI design performs poorly with respect to power, average bias and variability but the best with respect to patient benefit. The CFLGI and TS design have values lying near to the outer edge of the star plot for all measures, thus showing that they perform well with respect to all of the performance

measures. Although CFLGI and TS have similar performances, they are not directly comparable as they attain different compromises between the competing objectives. Rather than having a flat probability protection for the control arm during the trial, the definition of the CFLGI rule could be adjusted in a similar way to TS and TP, which we expect would result in an advantage over TS in terms of patient benefit, especially for smaller trials with several arms.

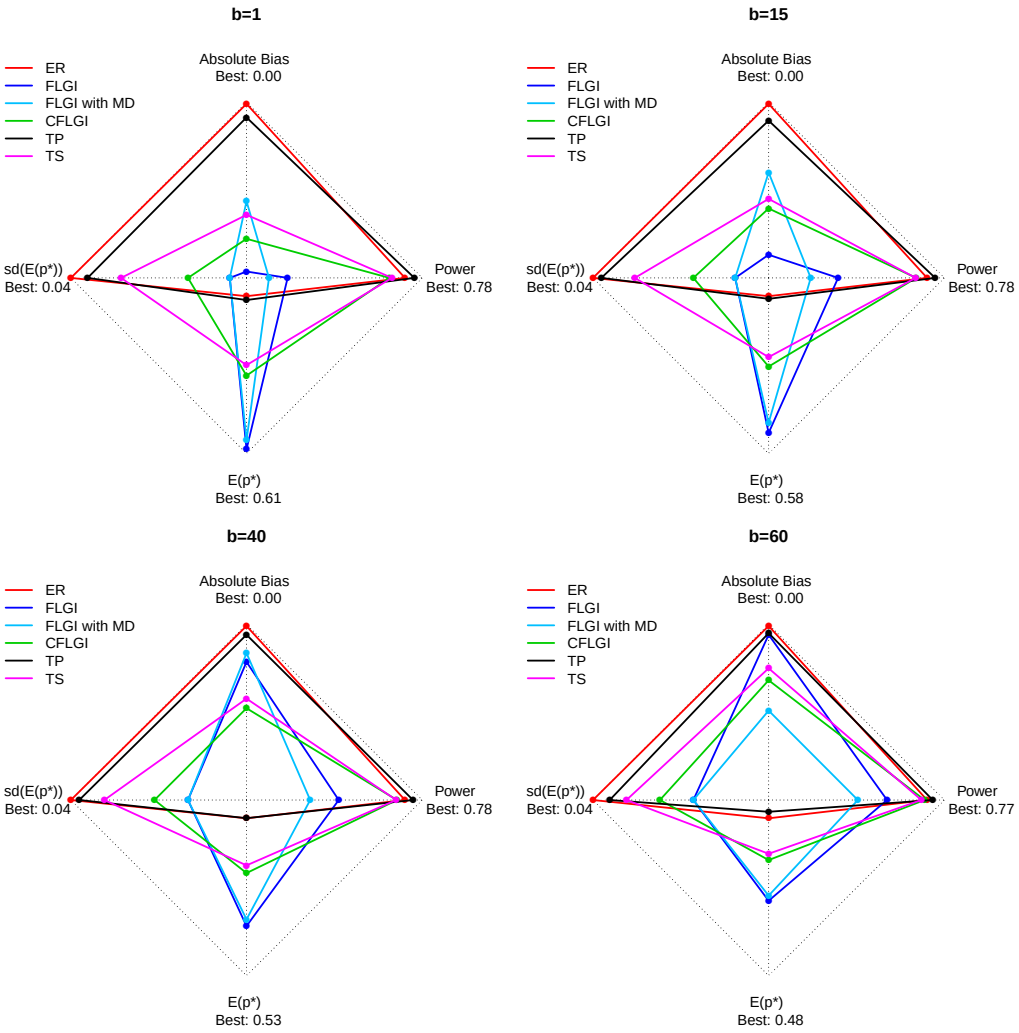


Figure 6.3.1: The trade-offs between the expected proportion of patients allocated to the superior arm, $\mathbb{E}(p^*)$, power, average absolute bias of the treatment effect estimator and variability of patient allocations for the different designs, including normal FLGI and normal FLGI with missing data (MD), for block sizes $b = (1, 15, 40, 60)$ in a three-armed trial of size $T = 120$ (assuming unknown variance).

		$\mu_0 = \mu_1 = \mu_2 = -0.05$					$\mu_0 = -0.05, \mu_1 = 0.07, \mu_2 = 0.13$			
Design		$t_{1-\alpha}$	α	$\mathbb{E}(p^*)$ (s.d.)	RelETO% (s.d.)	Bias (s.d.)	$1-\beta$	$\mathbb{E}(p^*)$ (s.d.)	RelETO% (s.d.)	Bias (s.d.)
ER	$b=1$	1.595	0.0997	0.3332 (0.04)	-0.28 (3.79)	-0.0007 (0.08)	0.7608	0.3333 (0.04)	-0.07 (3.87)	-0.0002 (0.08)
FLGI	$b=1$	1.731	0.0992	0.3331 (0.20)	0.03 (3.77)	0.0006 (0.12)	0.4936	0.6122 (0.23)	88.13 (4.34)	0.0212 (0.13)
	$b=2$	1.718	0.0994	0.3336 (0.20)	-0.09 (3.79)	0.0002 (0.12)	0.5121	0.6087 (0.22)	87.48 (4.35)	0.0209 (0.13)
	$b=4$	1.737	0.0979	0.3343 (0.19)	0.10 (3.79)	0.0001 (0.12)	0.5100	0.6024 (0.22)	86.04 (4.35)	0.0203 (0.13)
	$b=8$	1.741	0.0954	0.3323 (0.19)	-0.47 (3.79)	-0.0003 (0.11)	0.5235	0.5938 (0.21)	83.52 (4.35)	0.0198 (0.12)
	$b=15$	1.709	0.1018	0.3337 (0.18)	-0.03 (3.79)	0.0001 (0.11)	0.5582	0.5824 (0.21)	80.94 (4.34)	0.0191 (0.12)
	$b=20$	1.725	0.1001	0.3342 (0.18)	0.00 (3.78)	0.0004 (0.11)	0.5608	0.5716 (0.20)	77.82 (4.32)	0.0166 (0.11)
	$b=40$	1.662	0.1009	0.3342 (0.17)	-0.35 (3.80)	-0.0004 (0.10)	0.6100	0.5296 (0.18)	66.48 (4.26)	0.0047 (0.10)
	$b=60$	1.591	0.1009	0.3337 (0.15)	-0.05 (3.80)	0.0004 (0.09)	0.6697	0.4835 (0.16)	51.91 (4.16)	-0.0013 (0.09)
FLGI-HZ ($\gamma=2$)	$b=1$	1.603	0.0972	0.3333 (0.04)	0.37 (3.79)	-0.0001 (0.08)	0.6157	0.5137 (0.05)	67.11 (3.84)	-0.0001 (0.10)
	$b=2$	1.592	0.0996	0.3334 (0.04)	-0.41 (3.80)	-0.0002 (0.08)	0.6161	0.5137 (0.05)	67.19 (3.85)	-0.0005 (0.10)
	$b=4$	1.587	0.1009	0.3332 (0.04)	0.08 (3.79)	0.0000 (0.08)	0.6209	0.5134 (0.05)	66.96 (3.84)	0.0002 (0.10)
	$b=8$	1.600	0.0979	0.3331 (0.04)	0.79 (3.79)	-0.0002 (0.08)	0.6162	0.5133 (0.05)	67.03 (3.84)	-0.0004 (0.10)
	$b=15$	1.585	0.0994	0.3332 (0.04)	-0.18 (3.80)	-0.0006 (0.08)	0.6207	0.5131 (0.05)	66.86 (3.85)	-0.0002 (0.10)
	$b=20$	1.605	0.0991	0.3334 (0.04)	-0.32 (3.80)	-0.0001 (0.08)	0.6187	0.5115 (0.05)	66.74 (3.87)	-0.0001 (0.10)
	$b=40$	1.587	0.1014	0.3334 (0.04)	0.29 (3.80)	-0.0002 (0.08)	0.6839	0.4827 (0.06)	53.2 (3.93)	0.0003 (0.09)
	$b=60$	1.598	0.1000	0.3335 (0.05)	0.25 (3.79)	0.0004 (0.08)	0.7665	0.4240 (0.05)	19.34 (4.08)	0.0000 (0.08)
CFLGI	$b=1$	1.530	0.1012	0.2894 (0.16)	-0.01 (3.80)	-0.0236 (0.10)	0.7254	0.4780 (0.18)	30.41 (4.12)	-0.0171 (0.09)
	$b=2$	1.533	0.1009	0.2922 (0.16)	0.11 (3.78)	-0.0223 (0.10)	0.7272	0.4756 (0.17)	30.28 (4.12)	-0.0164 (0.09)
	$b=4$	1.520	0.1017	0.2949 (0.16)	-0.06 (3.79)	-0.0209 (0.10)	0.7324	0.4734 (0.17)	30.20 (4.09)	-0.0157 (0.09)
	$b=8$	1.522	0.1027	0.2960 (0.15)	-0.55 (3.80)	-0.0203 (0.10)	0.7366	0.4685 (0.17)	29.11 (4.07)	-0.0144 (0.09)
	$b=15$	1.530	0.0988	0.2976 (0.15)	0.04 (3.80)	-0.0190 (0.10)	0.7361	0.4615 (0.16)	27.68 (4.09)	-0.0133 (0.09)
	$b=20$	1.534	0.1020	0.2981 (0.15)	-0.53 (3.80)	-0.0180 (0.09)	0.7328	0.4554 (0.16)	26.79 (4.07)	-0.0130 (0.09)
	$b=40$	1.538	0.1006	0.3029 (0.14)	0.20 (3.78)	-0.0134 (0.09)	0.7425	0.4331 (0.14)	20.98 (4.03)	-0.0105 (0.08)
	$b=60$	1.540	0.1005	0.3077 (0.13)	-0.26 (3.79)	-0.0082 (0.08)	0.7552	0.4094 (0.12)	15.80 (4.02)	-0.0070 (0.08)
TP	$b=1$	1.582	0.1009	0.3116 (0.09)	0.11 (3.78)	-0.0127 (0.09)	0.7822	0.3403 (0.06)	-5.75 (3.82)	-0.0019 (0.08)
	$b=2$	1.580	0.0991	0.3119 (0.09)	0.08 (3.80)	-0.0128 (0.09)	0.7791	0.3402 (0.05)	-6.68 (3.82)	-0.0028 (0.08)
	$b=4$	1.571	0.0997	0.3121 (0.09)	0.05 (3.79)	-0.0128 (0.09)	0.7824	0.3400 (0.05)	-5.57 (3.81)	-0.0019 (0.08)
	$b=8$	1.579	0.0991	0.3130 (0.09)	-0.33 (3.78)	-0.0121 (0.09)	0.7820	0.3393 (0.05)	-6.07 (3.83)	-0.0016 (0.08)
	$b=15$	1.570	0.1009	0.3148 (0.08)	0.45 (3.78)	-0.0107 (0.09)	0.7793	0.3381 (0.05)	-6.31 (3.84)	-0.0023 (0.08)
	$b=20$	1.576	0.0982	0.3146 (0.08)	-0.34 (3.79)	-0.0098 (0.08)	0.7781	0.3370 (0.05)	-6.02 (3.82)	-0.0020 (0.08)
	$b=40$	1.567	0.1042	0.3174 (0.08)	-0.11 (3.80)	-0.0069 (0.08)	0.7792	0.3331 (0.05)	-6.68 (3.91)	-0.0013 (0.08)
	$b=60$	1.574	0.1021	0.3131 (0.07)	-0.35 (3.78)	-0.0051 (0.08)	0.7738	0.3214 (0.06)	-11.58 (3.95)	-0.0011 (0.08)
TS	$b=1$	1.629	0.1024	0.3331 (0.09)	0.09 (3.78)	0.0002 (0.09)	0.7313	0.4589 (0.10)	47.23 (4.08)	0.0141 (0.10)
	$b=2$	1.651	0.0985	0.3340 (0.09)	-0.10 (3.78)	0.0009 (0.09)	0.7223	0.4574 (0.10)	46.67 (4.07)	0.0132 (0.10)
	$b=4$	1.641	0.0986	0.3337 (0.09)	0.18 (3.80)	0.0006 (0.09)	0.7267	0.4557 (0.10)	46.10 (4.08)	0.0138 (0.10)
	$b=8$	1.620	0.1028	0.3336 (0.08)	0.10 (3.79)	0.0000 (0.09)	0.7366	0.4513 (0.10)	44.76 (4.07)	0.0124 (0.09)
	$b=15$	1.626	0.1012	0.3327 (0.08)	0.38 (3.78)	-0.0005 (0.09)	0.7349	0.4443 (0.09)	43.38 (4.07)	0.0121 (0.09)
	$b=20$	1.632	0.1006	0.3330 (0.08)	0.01 (3.78)	-0.0001 (0.09)	0.7344	0.4390 (0.09)	41.59 (4.06)	0.0110 (0.09)
	$b=40$	1.635	0.0980	0.3329 (0.07)	0.00 (3.79)	-0.0003 (0.08)	0.7410	0.4191 (0.08)	34.82 (4.05)	0.0093 (0.09)
	$b=60$	1.609	0.1016	0.3328 (0.07)	0.03 (3.78)	-0.0006 (0.08)	0.7478	0.3979 (0.08)	27.29 (4.02)	0.0055 (0.09)
Gwise	$n^{\text{ER}}=2$	1.778	0.0998	0.3343 (0.08)	0.50 (3.82)	0.0003 (0.09)	0.7101	0.3336 (0.08)	0.12 (4.12)	0.0002 (0.09)
	$n^{\text{ER}}=4$	1.620	0.1011	0.3334 (0.05)	-0.32 (3.78)	0.0001 (0.08)	0.7599	0.3328 (0.05)	-0.13 (3.89)	-0.0006 (0.08)
	$n^{\text{ER}}=8$	1.618	0.1010	0.3332 (0.04)	-0.09 (3.78)	-0.0001 (0.08)	0.7627	0.3332 (0.04)	-0.01 (3.9)	-0.0003 (0.08)

Table 6.3.4: Comparison of performance measures for a three-armed trial using different designs when the variance is assumed unknown and $T = 120$, averaged over 50,000 trial replications. Note that the true variance of the response is $\sigma_k^2 = 0.346^2$ for $k \in \{0, 1, 2\}$.

6.4 Dichotomisation: Patient Benefit and Efficiency Cost

Phase II cancer trials, such as the ones considered above, are traditionally conducted as single arm studies using a binary response rate as the primary endpoint, which is formed by splitting the underlying continuous data (change in tumour size) into two groups (success or failure of a treatment), that is, dichotomising. This dichotomisation is often based on the *Response Evaluation Criteria in Solid Tumors* (Eisenhauer *et al.*, 2009) which categorises the change in tumour size and number of lesions into four levels: complete response, partial response, stable disease, and progressive disease. A treatment is considered a success if patients experience either a partial or complete response (i.e. at least a 30% reduction in the total diameter of target lesions), and a failure otherwise. If new lesions appear, or non-target lesions grow beyond a certain percentage, this is also classed as a treatment failure.

Dichotomising continuous data is a widely adopted approach in clinical research. However, this comes at the cost of losing power as well as raising issues such as where exactly the dichotomisation cutpoint should be. For further implications, see Cohen (1983) and Maccallum *et al.* (2002). Within the literature, there is a strong focus on the loss of efficiency associated with dichotomising a continuous variable, but no mention of the cost to patients in the trial. Therefore, we will use the same two-armed example as in Section 6.3.2 to compare the performance, in terms of patient benefit measures, of the continuous FLGI to the binary FLGI proposed in Villar *et al.* (2015b). However, since the binary FLGI compares response rates, we increase the total sample size from $T = 72$ to 128, as this is the size required to detect an improvement from 20% to 40% with 80% power using a one-sided test at the $\alpha = 0.05$ level; a 77% increase on that required for the continuous case.

Figure 6.5.1 shows the efficiency costs of dichotomising a continuous endpoint. A

trial of size 128 achieves almost 100% power to detect the target treatment difference when using a continuous endpoint, as opposed to 80% power when using a binary one. Moreover, Figure 6.5.1 also illustrates that there is an important patient benefit cost of using a binary endpoint instead of a continuous one when using RAR. In particular, the normal FLGI (all versions) has not only a higher power level, but also a considerably higher expected proportion of patients on the best arm for every block size in a trial of size 128.

6.5 Imputing Complete Responses and Dropouts

The patient benefit cost associated with dichotomising requires an important practical consideration to be taken into account when interpreting it. To implement any response-adaptive design in practice, particularly in cancer trials like those used in this chapter, we need an online imputation method to account for patients who (a) die or dropout of the trial before the follow-up time, or (b) have a complete response (since this causes the log ratio to be undefined). Two approaches have been proposed to impute these cases in Karrison *et al.* (2007) and Jaki *et al.* (2013), a review of which is provided by Wason and Jaki (2016).

So far, we have assumed that *all* patients generate an observable response, which is clearly not realistic. Whereas deaths/dropouts and complete responses are easily imputed in the binary case, there is no obvious way of translating these outcomes into continuous variables. Building upon the solution in Karrison *et al.* (2007), where the best and worst possible outcomes are used to impute complete responses and deaths/dropouts, respectively, we instead randomise from the upper tail of the (theoretical) distribution under $\mathcal{H}_1$ if we observe a complete response, and from the lower tail of the null distribution to account for deaths or dropouts, regardless of which treatment the patient received. Thus, this approach allows for a response-adaptive

algorithm to be used by computing the missing values online as the trial progresses. Furthermore, choosing the missing values *randomly*, as opposed to using the same values every time, is perhaps a better reflection of reality or, at the very least, a reflection of the distributional assumptions made to determine the size of the study based on power considerations. Alternatively, we could estimate the best and worst possible outcomes based on the interim data observed after each block. However, in practice, if the deaths, dropouts or complete responses occur early on in the trial, there would be too few, or possibly no, observations available to accurately represent these values.

Figure 6.5.1 shows the results for the normal FLGI when we implement our online imputation method assuming that we observe a 4% rate of deaths or dropouts and a 1% rate of complete responses. This is illustrated under the assumption of both a known and unknown variance, labelled as FLGI-known with missing data (MD) and FLGI with MD, respectively. These rates are consistent with values reported in Karrison *et al.* (2007). Figure 6.5.1 shows that, as expected, this missing data assumption decreases both the efficiency and patient benefit advantages, relative to the FLGI with complete observations, for both the known and unknown variance cases. Nevertheless, the imputed continuous FLGI procedure continues to greatly outperform the binary FLGI with respect to both criteria. Figure 6.3.1 suggests that similar conclusions also apply for the multi-armed missing data case (see FLGI with MD).

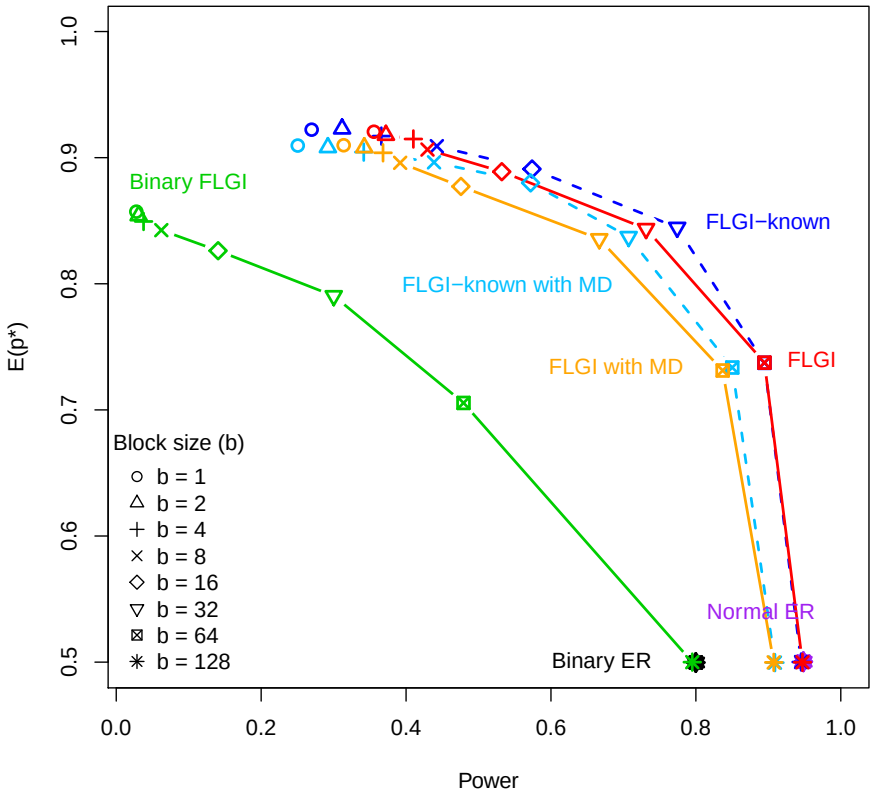


Figure 6.5.1: The trade-off between the expected proportion of patients allocated to the superior arm, $\mathbb{E}(p^*)$, and power for the: Binary ER, Normal ER, Binary FLGI, Normal FLGI and Normal FLGI with Missing Data (MD) imputed in an online fashion for block sizes $b = (1, 2, 4, 8, 16, 32, 64, 128)$ in a two-armed trial of size $T = 128$. The latter two designs are shown when assuming both an unknown variance and known (correct) variance (dashed line and labelled as FLGI-known).

6.6 Discussion

The RAR literature contains relatively few procedures for a continuous endpoint assumed to be normally distributed with unknown variance, fewer still that are defined for the multi-armed case and none that are forward-looking. We propose the first forward-looking RAR algorithm applicable to this case which is orientated towards an optimality criterion with respect to patient benefit.

In this chapter, we have shown that using a continuous endpoint instead of dichotomising can offer efficiency, but also patient benefit advantages, when combined with RAR. An implication of not dichotomising could be a lack of robustness to departures from the assumed response distribution. For example, if assuming responses are normally distributed but the observed data is non-normal, how much of an impact would this have on the performance measures of the proposed design, and would the aforementioned advantages over the dichotomisation approach persist? This forms an area of further work.

Implementing a RAR procedure, such as the FLGI, in the context of phase II cancer trials requires dealing with missing data from patients in an online fashion. The naïve imputation method suggested in this work, based on the method by Karrison *et al.* (2007), shows that there are still important benefits even if a low rate of missing observations is anticipated. Further work is needed to develop imputation methods that can be used in combination with RAR.

An important advantage of our proposed method is that it can be implemented without assuming a fixed, known, and common variance. In fact, the FLGI with unknown variance can learn about the variance simultaneously as it learns about the treatment means, and update the randomisation probabilities accordingly. Additionally, the method can incorporate covariates in the way suggested by Villar and Rosenberger (2018).

The motivation of our algorithm is in the setting of clinical trials, but it applies to sequential allocation problems more generally. Future research could consider the issue of estimation following the sequential tests used in combination with these novel designs, similar to work in Coad (1991a, 1994).

6.7 Appendix

6.7.1 The MABP and FLGI for Normally Distributed Endpoints

In this Appendix, we provide a more detailed description of the MABP for normally distributed endpoints, its solution by the Gittins index (GI) and an additional example of the Forward-Looking Gittins Index (FLGI) probabilities for normally distributed endpoints with a known variance.

Recall that the MABP in this case involves a multi-armed clinical trial that will test the effectiveness of K experimental treatments against a control treatment on a sample of T patients, with K and T fixed and known in advance. Patients are labelled by t ($t = 1, \ldots, T$) and treatments by k ($k = 0, \ldots, K$), where $k = 0$ denotes the control. The response of patient t allocated to arm k is a random variable denoted by $Y_{k,t}$ and assumed to follow a normal distribution $Y_{k,t} \sim N(\mu_k, \sigma_k^2)$. Without loss of generality, we also assume that a *larger* response is preferable and that σ_k^2 is known.

In order to derive the FLGI rule, we first need to obtain the GI for a normally distributed variable and the MABP associated with this trial design problem. For this purpose, we assume the following. (i) Each unknown parameter μ_k has a prior distribution $\pi_{k,0}$ at the start of the trial (before any observation has been made) which we take to be the normal prior $N\left(\mu_k^0, \frac{\sigma_k^2}{n_k^0}\right)$. Note that the form of the prior when both μ_k and σ_k^2 are unknown is provided below. (ii) Patients enter the trial one-by-one and responses are observed immediately after treatment. We will remove these assumptions when we formulate the FLGI rule. (iii) Only one treatment can be allocated per patient and we let $a_{k,t}^r$ be a binary indicator variable denoting whether patient $t + 1$ is assigned to treatment k for patient allocation rule r or not, given the information available on all treatments. (iv) Given the conjugacy of the prior and normally distributed responses, prior distributions are converted into normal posterior

distributions for each μ_k via Bayes' Theorem. After treating patient t, if $n_{k,t}$ responses from treatment k have been observed (each denoted by $y_{k,i}$ with $i \in \{1, \ldots, n_{k,t}\}$ and $n_{k,t} \leq t$), then the posterior distribution of μ_k at time t is $\pi_{k,t}(\mu_k \mid y_{k,1}, \ldots, y_{k,n_{k,t}}) \sim N\left(\frac{n_{k,t}\overline{y}_{k,t}+n_k^0\mu_k^0}{n_{k,t}+n_k^0}, \frac{\sigma_k^2}{n_{k,t}+n_k^0}\right)$ by Bayes' Theorem, where $\overline{y}_{k,t} = \frac{1}{n_{k,t}}\sum_{i=1}^{n_{k,t}} y_{k,i}$ is the sample mean and n_k^0 is the implicit sample size from the prior information (Spiegelhalter *et al.*, 2004, p. 62). The posterior distribution, $\pi_{k,t}$, can be identified by the parameters $\widetilde{y}_{k,t}$ (posterior mean) and $n_k^0 + n_{k,t}$, which we subsequently refer to as the *state* (of the bandit) (Gittins *et al.*, 2011). Note that when the variance is *unknown*, an additional parameter, $\widetilde{s}_{k,t}^2$, denoting the posterior variance of patient t on arm k, is required to identify $\pi_{k,t}$ and in this case, we need to specify a *joint* prior distribution for μ_k and σ_k^2 at the start of the trial. We take this to be the normal-inverse-gamma distribution (where the variance follows an inverse-gamma distribution and the mean, conditional on the variance, has a normal distribution). Consequently, the marginal prior distribution for μ_k has a Student's t-distribution. When we observe an outcome $y_{k,t+1}$ from patient $t+1$ on arm k, the state $(\widetilde{y}_{k,t}, \widetilde{s}_{k,t}, n_k^0 + n_{k,t})$ is updated as follows

$$\left(\frac{(n_k^0 + n_{k,t})\widetilde{y}_{k,t} + y_{k,t+1}}{n_k^0 + n_{k,t} + 1}, \left(\frac{\widetilde{s}_{k,t}^2(n_k^0 + n_{k,t} - 1)}{n_k^0 + n_{k,t}} + \frac{(y_{k,t+1} - \widetilde{y}_{k,t})^2}{n_k^0 + n_{k,t} + 1}\right)^{\frac{1}{2}}, n_k^0 + n_{k,t} + 1\right). \tag{6.7.1}$$

The MABP is to find a patient allocation rule r that attains the maximum expected patient response given the initial information about the treatments before the start of the trial. Mathematically, this is expressed as

$$\max_{r \in R} \mathbb{E}^r \left[\left(\sum_{t=0}^{T-1}\sum_{k=0}^{K} d^t \mathbb{E}[Y_{k,t} \mid \mathbf{x}_{k,t}] a_{k,t}^r\right) \,\middle|\, \widetilde{\mathbf{x}}_\mathbf{0}\right], \tag{6.7.2}$$

where $\mathbf{x}_{k,t} = (\overline{y}_{k,t}, n_{k,t}, \mu_k^0, n_k^0)$, $\widetilde{\mathbf{x}}_\mathbf{0} = \{\mathbf{x}_{k,0}\}_{k=0}^K$ is the initial joint state with all the prior parameters, R is the set of admissible allocation rules, $\mathbb{E}^r[\cdot]$ denotes expectation under allocation rule r, and $0 \leq d < 1$ is a discount factor. In MABPs, rewards are

geometrically discounted so that an infinite horizon can be considered, i.e. patient t's response yields a reward of $d^t Y_{k,t}$ for some k. In practice, a solution that depends on d, such as the GI, can be adapted to solve an undiscounted problem with a specific finite horizon, as explained in Edwards *et al.* (2017, Definition 6.6).

The exact solution to (6.7.2), obtained via dynamic programming, uses a backward induction algorithm which becomes computationally infeasible very quickly as T and K grow. The GI solution, first introduced by Gittins and Jones (1979), eliminates this computational infeasibility by ensuring that the optimal solution to (6.7.2) can be obtained by simply allocating every patient to the arm with the highest GI. Similarly to equation (6.2.1) for the unknown variance case, the GIs, $\mathcal{G}(\tilde{y}_{k,t}, \sigma_k, n_{k,t})$, for the known variance case in (6.7.2) can be expressed as

$$\mathcal{G}(\tilde{y}_{k,t}, \sigma_k, n_{k,t}) = \tilde{y}_{k,t} + \sigma_k \mathcal{G}(0, 1, n_k^0 + n_{k,t}, d), \tag{6.7.3}$$

where $\mathcal{G}(0, 1, n_k^0 + n_{k,t}, d)$ denotes the GI value of a standardised bandit problem with posterior mean 0, standard deviation 1, implicit sample size n_k^0, $n_{k,t}$ observations and discount factor d (Gittins *et al.*, 2011, Theorem 7.13). These were first computed in Jones (1975). Table 6.7.1 shows indices corresponding to the *unknown* variance case, as used in Sections 6.2–6.5, based on those presented in Gittins *et al.* (2011, Table 8.3).

We implement the solution in (6.7.3) at a very low computational cost by calculating the values of $\mathcal{G}(0, 1, n_k^0 + n_{k,t}, d)$ in advance and interpolating from the tables printed in Gittins *et al.* (2011, pp. 261–262). Details on how to compute these indices using value iteration can be found in Gittins *et al.* (2011, Chapters 7 and 8). Using (6.7.3) and the GI rule, we can compute the FLGI probabilities for normally distributed endpoints (with known variance) using equation (3) in Villar *et al.* (2015b). We now assume that instead of enrolling patients one-by-one, patients are enrolled in groups of size b over J stages, so that $J \times b = T$. Our response-adaptive rule will

sequentially randomise the next b patients among the $K+1$ treatments at stage j ($j = 1, \ldots, J$) given the data up to and including block $j-1$ according to what the GI rule would do.

Example

We now illustrate the rule's implementation using an example for the case of known variances. We calculate the FLGI probabilities using the simplest possible case of a two-arm trial testing a control treatment ($k = 0$) against an experimental treatment ($k = 1$) with a block of size two ($b = 2$) and a known, common variance of $\sigma_k^2 = \sigma^2 = 1$. We assume a prior of $\mu_k \sim N(0, 1)$ so that the initial state, $(\widetilde{y}_{k,0}, n_k^0)$, is $(0, 1)$ for both $k = \{0, 1\}$. Suppose further that both patients are allocated to the control treatment in the first block of the trial resulting in responses $y_{0,1} = 3.1$ and $y_{0,2} = -0.4$. The updated state after the first observation becomes $(\widetilde{y}_{0,1}, n_0^0 + n_{0,1}) = (1.55, 2)$ and after the second observation becomes $(\widetilde{y}_{0,2}, n_0^0 + n_{0,2}) = (\frac{0-0.4+3.1}{3}, 3) = (0.9, 3)$. Consequently, for the second block, the prior parameters for the control and experimental treatment respectively are $(0.9, 3)$ and $(0, 1)$, i.e. $\mu_0 \sim N(0.9, \frac{1}{3})$ and $\mu_1 \sim N(0, 1)$.

From equation (6.7.3), setting $d = 0.995$ and using Gittins *et al.* (2011, Table 8.1)[5], the GI for the control treatment is $\mathcal{G}_0(0.9, 1, 2) = 0.9 + 1 \times \mathcal{G}_0(0, 1, 3, 0.995) = 0.9 + \frac{0.20137}{3(1-0.995)^{\frac{1}{2}}} = 1.8493$. For the experimental treatment, we only have the information available from the initial state (since no observations have yet been observed on this arm). Thus, the corresponding GI for this arm is $\mathcal{G}_1(0, 1, 0) = 0 + \frac{0.12852}{(1-0.995)^{\frac{1}{2}}} = 1.8175$.

Given that the control treatment has the maximum GI, the first patient of the second block (i.e. patient 3) is allocated to the control treatment with probability 1 since there is only one optimal action possible at this point. If we denote the random outcome of this patient by $Y_{0,3}$, then the updated state for the control treatment

[5]Note that Gittins *et al.* (2011, Table 8.1) provides values of $(n_k^0 + n_{k,t})(1-d)^{\frac{1}{2}}\mathcal{G}(0, 1, n_k^0 + n_{k,t}, d)$.

is $(\widetilde{Y}_{0,3}, n_0^0 + n_{0,3}) = \left(\frac{0-0.4+3.1+Y_{0,3}}{4}, 4\right)$. Thus, the corresponding index for the control treatment can be expressed as a function of the random outcome from patient three as follows: $\mathcal{G}_0(\widetilde{Y}_{0,3}, 1, 3) = \frac{Y_{0,3}+2.7}{4} + \mathcal{G}_0(0, 1, 4, 0.995) = \frac{Y_{0,3}}{4} + 1.4669$, where $\mathcal{G}_0(0, 1, 4, 0.995) = \frac{0.22398}{4(1-0.995)^{\frac{1}{2}}} = 0.79189$.

For the experimental treatment, we have no new information and so the corresponding index remains unchanged at 1.8175. According to the GI rule, it will be optimal to allocate the control treatment to the second patient of the second block if and only if $\mathcal{G}_0(\widetilde{Y}_{0,3}, 1, 3) > \mathcal{G}_1(0, 1, 0) = 1.8175$, that is, if $Y_{0,3} > 1.4024$. Since $Y_{0,3} \sim N(0.9, 1)$, we expect this to happen with probability $\mathbb{P}(Y_{0,3} > 1.4024) = 0.3077$. If $Y_{0,3} < 1.4024$, which happens with probability 0.6923, then $\mathcal{G}_0(\widetilde{Y}_{0,3}, 1, 3) < \mathcal{G}_1(0, 1, 0)$ and the second patient of the second block is optimally allocated to the experimental treatment. Notice that if $Y_{0,3} = 1.4024$, then there is a tie in the index values and it is equally optimal to allocate any of the two treatments. Although theoretically we expect this to happen with probability 0 (since we are dealing with a continuous distribution), in practice this is possible and if it were to happen, we would simply randomise with probability 0.5. Hence, the probability of a patient receiving either the control or experimental treatment when using the normal FLGI procedure in this block is $\frac{1+1\times\mathbb{P}(Y_{0,3}>1.4024)}{2} = 0.6538$ and $\frac{0+1\times\mathbb{P}(Y_{0,3}<1.4024)}{2} = 0.3462$, respectively. Figure 6.7.1 illustrates how the FLGI probabilities for block two, given the data in block one, are computed via a probability tree.

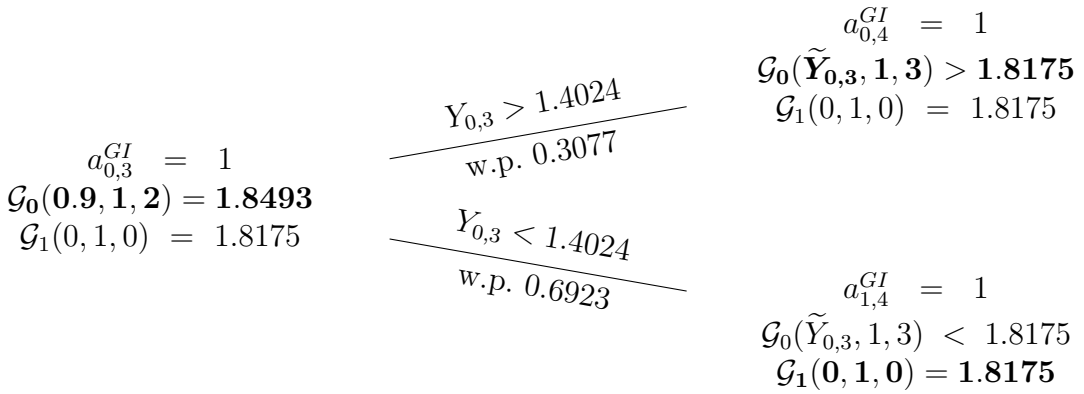


Figure 6.7.1: The FLGI rule and a probability tree of all trial histories using the GI rule when $K + 1 = 2$, $b = 2$, $d = 0.995$ and the state at the start of the second block, $(\widetilde{y}_{k,2}, n_k^0 + n_{k,2})$, is $(0.9, 3)$ for arm $k = 0$ and $(0, 1)$ for arm $k = 1$. Bold text indicates the allocated treatment under the GI rule $\{a_{k,t}^{GI}\}$. (Note that for simplicity of the illustration we have omitted the branch corresponding to the case when $Y_{0,3} = 1.4024$ since $\mathbb{P}(Y_{0,3} = 1.4024) = 0$).

d / $n_k^0 + n_{k,t}$	0.5	0.6	0.7	0.8	0.9	0.95	0.99	0.995
2	0.23984	1.04741	1.55545	2.81630	5.16921	10.14092	39.33433	65.58475
3	0.15620	0.21476	0.29804	0.43425	0.73571	1.16561	3.10200	4.60490
4	0.09486	0.13001	0.17914	0.25664	0.41606	0.61934	1.34279	1.81263
5	0.07058	0.09673	0.13323	0.19047	0.30608	0.44776	0.90524	1.17299
6	0.05679	0.07791	0.10742	0.15369	0.24666	0.35900	0.70542	0.89632
7	0.04779	0.06564	0.09061	0.12983	0.20866	0.30352	0.59010	0.74336
8	0.04135	0.05685	0.07858	0.11278	0.18165	0.26451	0.51233	0.64259
9	0.03649	0.05021	0.06948	0.09988	0.16128	0.23525	0.45557	0.57012
10	0.03268	0.04500	0.06234	0.08974	0.14527	0.21234	0.41187	0.51498
20	0.01611	0.02228	0.03106	0.04515	0.07444	0.11090	0.22299	0.28120
30	0.01072	0.01485	0.02076	0.03032	0.05049	0.07615	0.15786	0.20137
40	0.00804	0.01115	0.01560	0.02285	0.03829	0.05821	0.12347	0.15903
50	0.00643	0.00892	0.01250	0.01834	0.03086	0.04719	0.10189	0.13229
60	0.00536	0.00744	0.01043	0.01532	0.02586	0.03971	0.08697	0.11368
70	0.00459	0.00638	0.00895	0.01316	0.02225	0.03429	0.07599	0.09991
80	0.00402	0.00558	0.00784	0.01153	0.01953	0.03018	0.06755	0.08927
90	0.00357	0.00496	0.00697	0.01026	0.01741	0.02696	0.06084	0.08077
100	0.00321	0.00447	0.00627	0.00924	0.01570	0.02436	0.05538	0.07381
200	0.00161	0.00224	0.00314	0.00464	0.00793	0.01242	0.02944	0.04024
300	0.00107	0.00149	0.00210	0.00310	0.00531	0.00834	0.02015	0.02790
400	0.00080	0.00112	0.00157	0.00233	0.00399	0.00628	0.01534	0.02142
500	0.00064	0.00090	0.00126	0.00186	0.00319	0.00504	0.01239	0.01740
600	0.00054	0.00075	0.00105	0.00155	0.00266	0.00421	0.01040	0.01466
700	0.00046	0.00064	0.00090	0.00133	0.00228	0.00361	0.00896	0.01268
800	0.00040	0.00056	0.00079	0.00116	0.00200	0.00316	0.00787	0.01117
900	0.00036	0.00050	0.00070	0.00104	0.00178	0.00281	0.00702	0.00999
1000	0.00032	0.00045	0.00063	0.00093	0.00160	0.00253	0.00634	0.00903

Table 6.7.1: Gittins indices for a normal reward process with unknown variance where d and $n_k^0 + n_{k,t}$ denote the discount factor and total amount of information, respectively. These values are based on those reported in Gittins *et al.* (2011, Table 8.3).

6.7.2 Effect of Discount Factor on FLGI Performance

A practical consideration for our design is the choice of discount factor, d. We recommend choosing d to be close to that obtained when applying the formula suggested by Wang (1991b), namely, $d = 1 - 1/T$, where T is the trial size. Here, we discuss the implications of *not* following this recommendation on the performance of the FLGI (with known variance) by presenting results corresponding to $d = 0,\ 0.5$ and 0.99 in Table 6.7.2. Note that the results for $d = 0.995$ (the discount factor used throughout) are shown in (i) of Table 6.7.3. When $d = 0$, the design is analogous to a fully myopic policy which treats every patient as if they are the last one in the trial. In contrast, the closer d is to 1, the greater the influence that potential responses from future participants have on allocation decisions made earlier in the trial, that is, the more "forward looking" the design will be. Thus, we expect the patient benefit measures to increase with d (up to a limit determined by the actual trial size), as illustrated in Table 6.7.2. In particular, Table 6.7.2 shows that as d increases from 0 to 0.995 for $b = 1$, $\mathbb{E}(p^*)$ increases by 0.164, which is equivalent to 11 more patients receiving the superior arm, and the relative *ETO* increases by 17.77%. As a result of the greater imbalance between the treatment arms for larger d, the bias of the treatment effect estimator (under $\mathcal{H}_1$) is also increased.

Interestingly, for smaller d, we observe that the patient benefit measures increase (up to around $b = 9$) followed by a decrease. This is due to an interaction between the discount factor and block size, whereby the increase in block size counteracts the myopic effect of a small d by forcing learning and consequently improving patient benefit. However, as the block size continues to grow, the effect of the design becoming more balanced supersedes the effect of the discount factor, causing the patient benefit to now reduce. Therefore, when choosing d, it is important to consider which block size will be used.

In terms of the power of the design, it increases somewhat with the size of d as

illustrated by Table 6.7.2 which shows that the power exhibited for the FLGI when $d = 0$ and $b = 1$ is 0.213 compared to 0.229 when $d = 0.995$. This makes sense because increasing d from a value that is much smaller than its recommendation for a fixed T reduces the myopic nature of the rule, meaning it will explore more of the arms (thus increasing power) and make better choices (also increasing patient benefit).

The discount factor also affects the variability of the allocations, which decreases considerably with the value of d under both $\mathcal{H}_0$ and $\mathcal{H}_1$. For example, Table 6.7.2 shows that under $\mathcal{H}_1$, the standard deviation (s.d.) of p^* when $d = 0$ and $b = 1$ is 0.43, which is 2.7 times larger than the corresponding s.d. when $d = 0.995$. Given that allocations under index-based designs (and response-adaptive designs more generally) can already be very variable, it does not make sense to choose a discount factor which exacerbates this even further.

A further practical drawback of using a discount factor that is too small is that it will increase the likelihood of the design allocating all patients to only one of the treatments (due to an under exploration). The number of times this occurred out of the 50,000 trial realisations is reported in the "Discarded" column of Table 6.7.2. For example, when $d = 0$ and $b = 1$, more than half of the 50,000 trial realisations under $\mathcal{H}_1$ (namely 25,621) resulted in this extreme allocation. Therefore, for the purpose of calculating the test statistic (and hence power) and bias values in these cases, we randomly sampled an observation from the distribution corresponding to the missing arm instead. In contrast, when $d = 0.995$, this problem did not occur in any of the 50,000 trial realisations (and similarly when $d = 0.99$).

Note that all of the aforementioned differences are most pronounced for smaller block sizes (which is when the design is most adaptive) since as the block size grows and the FLGI design becomes more balanced, the respective performance measures eventually converge, irrespective of the value of d.

Overall, provided that d is near to the recommendation suggested by Wang (1991b),

the performance of the FLGI will be similar — as illustrated by the results for $d = 0.99$ (Table 6.7.2) and $d = 0.995$ (Table 6.7.3(i)). However, choosing d to be too small in relation to T can alter the behaviour of the design significantly. Moreover, if we were to use this design in a rare disease context, where we envisage it would be best suited, d should be chosen to be large enough so that we account for all of the patient outcomes in the adaptations and hence ensure patient benefit for all.

	$\mu_0 = \mu_1 = 0.155$						$\mu_0 = 0.155, \mu_1 = 0.529$				
b	$z_{1-\alpha}$	α	$\mathbb{E}(p^*)$ (s.d.)	RelETO% (s.d.)	Bias (s.d.)	Discarded	$1-\beta$	$\mathbb{E}(p^*)$ (s.d.)	RelETO% (s.d.)	Bias (s.d.)	Discarded
$d = 0$ (Myopic)											
1	1.827	0.050	0.498 (0.48)	-0.14 (5.45)	0.00 (0.69)	14707	0.213	0.718 (0.43)	23.92 (12.49)	0.08 (0.68)	25621
2	1.829	0.049	0.497 (0.46)	0.02 (5.42)	0.00 (0.60)	7970	0.222	0.761 (0.39)	28.52 (11.53)	0.06 (0.60)	13789
4	1.799	0.050	0.501 (0.43)	0.12 (5.49)	-0.00 (0.50)	2211	0.266	0.810 (0.32)	33.87 (10.13)	0.05 (0.48)	3889
6	1.776	0.050	0.499 (0.41)	-0.24 (5.44)	-0.00 (0.42)	598	0.312	0.832 (0.27)	36.44 (9.03)	0.05 (0.39)	1093
9	1.752	0.050	0.501 (0.38)	0.05 (5.41)	-0.00 (0.33)	94	0.383	0.840 (0.22)	37.12 (8.06)	0.04 (0.31)	174
12	1.720	0.052	0.502 (0.35)	0.46 (5.43)	-0.00 (0.29)	15	0.448	0.837 (0.19)	36.87 (7.44)	0.03 (0.26)	28
18	1.708	0.050	0.499 (0.31)	-0.07 (5.45)	0.00 (0.24)	2	0.533	0.814 (0.15)	34.39 (6.80)	0.02 (0.22)	3
36	1.676	0.050	0.499 (0.21)	0.11 (5.40)	0.00 (0.18)	0	0.694	0.720 (0.10)	24.23 (6.05)	0.00 (0.17)	0
$d = 0.5$											
1	1.819	0.050	0.497 (0.46)	0.02 (5.42)	0.00 (0.66)	4894	0.228	0.772 (0.39)	29.64 (11.38)	0.07 (0.66)	18359
2	1.818	0.049	0.501 (0.44)	0.03 (5.45)	-0.00 (0.58)	2637	0.243	0.804 (0.35)	33.38 (10.54)	0.07 (0.58)	9769
4	1.770	0.054	0.497 (0.42)	-0.15 (5.47)	0.00 (0.49)	669	0.287	0.832 (0.29)	36.20 (9.45)	0.07 (0.46)	2696
6	1.765	0.052	0.499 (0.40)	-0.25 (5.45)	-0.00 (0.41)	192	0.329	0.844 (0.25)	37.72 (8.60)	0.06 (0.38)	763
9	1.757	0.049	0.501 (0.37)	0.14 (5.45)	-0.00 (0.33)	24	0.391	0.844 (0.21)	37.41 (7.80)	0.05 (0.30)	101
12	1.749	0.048	0.500 (0.34)	-0.05 (5.39)	-0.00 (0.29)	3	0.441	0.838 (0.18)	37.13 (7.26)	0.04 (0.26)	25
18	1.707	0.050	0.503 (0.30)	-0.17 (5.43)	-0.00 (0.24)	0	0.543	0.816 (0.15)	34.62 (6.70)	0.03 (0.22)	0
36	1.677	0.049	0.501 (0.21)	-0.16 (5.43)	-0.00 (0.18)	0	0.693	0.720 (0.10)	24.04 (6.09)	0.01 (0.17)	0
$d = 0.99$											
1	1.981	0.050	0.498 (0.35)	0.08 (5.43)	0.00 (0.53)	0	0.209	0.882 (0.19)	41.94 (7.38)	0.22 (0.47)	0
2	1.944	0.051	0.502 (0.34)	0.08 (5.44)	-0.00 (0.47)	0	0.255	0.879 (0.17)	41.44 (7.20)	0.19 (0.42)	0
4	1.907	0.051	0.501 (0.32)	-0.31 (5.45)	-0.00 (0.41)	0	0.313	0.871 (0.16)	40.62 (6.91)	0.14 (0.36)	0
6	1.885	0.049	0.500 (0.31)	-0.20 (5.41)	-0.00 (0.36)	0	0.349	0.865 (0.15)	39.95 (6.73)	0.11 (0.31)	0
9	1.850	0.049	0.501 (0.29)	0.36 (5.44)	-0.00 (0.31)	0	0.409	0.851 (0.14)	38.38 (6.64)	0.08 (0.26)	0
12	1.816	0.051	0.499 (0.28)	0.16 (5.42)	0.00 (0.27)	0	0.467	0.839 (0.13)	37.11 (6.44)	0.06 (0.23)	0
18	1.758	0.050	0.499 (0.25)	-0.10 (5.44)	0.00 (0.23)	0	0.556	0.811 (0.12)	33.89 (6.31)	0.04 (0.20)	0
36	1.684	0.051	0.499 (0.19)	-0.03 (5.44)	0.00 (0.18)	0	0.710	0.716 (0.10)	23.45 (6.02)	0.01 (0.17)	0

Table 6.7.2: The effect of altering the discount factor, d, on the performance of the FLGI for a two-armed trial when $\sigma_k^2 = 0.64^2$ is assumed known and $T = 72$, averaged over 50,000 trial replications. NB The "Discarded" column reports the number of trials that resulted in an extreme allocation with all patients being allocated to only one arm.

6.7.3 Effect of Prior Information on FLGI Performance

In this Appendix, we investigate how sensitive the FLGI is to the choice of prior on the location parameter μ_k when the variance is assumed known. Ultimately, the choice of prior on μ_k determines which GI we start the allocation rule with. The minimum amount of information we can assume, *a priori*, in order to initiate the GI policy is $n_k^0 = 1$ (known variance case) and $n_k^0 = 2$ (unknown variance case) since the GI is undefined for $n_k^0 = 0$. This gives rise to a normal prior with large variance (see Figure 6.7.2) which can be used as a so-called 'non-informative' prior (Spiegelhalter *et al.*, 2004, p. 62). All of the results presented thus far correspond to this 'non-informative' prior so that we can report the effects on patient response and other relevant statistical properties of the FLGI alone, without the influence of additional prior information. However, we now turn our attention to using different priors in conjunction with the FLGI. We use the results for the 'non-informative' prior (in (i) of Table 6.7.3) as a reference, and therefore refer to it as a *reference* prior from hereon (in keeping with the terminology used in Spiegelhalter *et al.* (1994, 2004), for example).

Taking the two-armed example from Section 6.3.2 (but now assuming known variance), we follow the suggestion provided in Spiegelhalter *et al.* (2004, Chapter 5) and consider two archetypal priors on μ_1, namely, the *sceptical* and *enthusiastic* prior (with the reference prior on μ_0).

The sceptical prior attempts to formalise the belief that large treatment differences are unlikely. In particular, the sceptical normal prior on μ_1 is centred around the (null hypothesis) value of 0.155 with only a small probability, say 5%, that the true value exceeds the alternative hypothesis value of 0.529, i.e. $\mathbb{P}(\mu_1 > 0.529) = 0.05$. This corresponds to a prior distribution of $\mu_1 \sim N\left(0.155, \frac{0.64^2}{n_1^0}\right)$, which has the following property

$$-0.64 \times \frac{z_{0.05}}{\sqrt{n_1^0}} = 0.529 - 0.155, \tag{6.7.4}$$

where n_1^0 is the implicit (prior) sample size and $z_{0.05} = -1.645$ is the fifth percentile of the standard normal distribution. Rearranging equation (6.7.4) gives $n_1^0 \approx 8$. Intuitively, this is equivalent to having eight patients' worth of information (with null mean) available at the start of the trial, that is, approximately 11% of the trial sample size expressing scepticism and showing no treatment difference. The performance measures of our design when starting with this prior on the experimental arm are shown in (ii) of Table 6.7.3 for all block sizes, b.

The enthusiastic prior, on the other hand, is centred on the alternative hypothesis value of 0.529 (with the same variance as the sceptical prior) and specifies that there is little evidence of no treatment effect a priori, i.e. there is a 5% chance of observing a value less than the null mean of 0.155. This corresponds to the following normal prior distribution $\mu_1 \sim N\left(0.529, \frac{0.64^2}{8}\right)$, which is equivalent to having already observed eight 'enthusiastic' responses before the start of the trial. The corresponding results when starting with this prior on the experimental arm are displayed in (iii) of Table 6.7.3.

			$\mu_0 = \mu_1 = 0.155$			$\mu_0 = 0.155, \mu_1 = 0.529$			
b	$z_{1-\alpha}$	α	$\mathbb{E}(p^*)$ (s.d.)	RelETO% (s.d.)	Bias (s.d.)	$1-\beta$	$\mathbb{E}(p^*)$ (s.d.)	RelETO% (s.d.)	Bias (s.d.)
(i) Reference ($n_0^0 = 1$) vs. Reference ($n_1^0 = 1$)									
1	1.991	0.053	0.500 (0.33)	-0.06 (5.42)	-0.00 (0.49)	0.229	0.882 (0.16)	41.69 (7.01)	0.22 (0.45)
2	1.969	0.051	0.500 (0.32)	0.34 (5.42)	-0.00 (0.44)	0.270	0.878 (0.16)	41.27 (6.85)	0.19 (0.41)
4	1.949	0.046	0.502 (0.30)	0.32 (5.43)	-0.00 (0.38)	0.313	0.870 (0.14)	40.30 (6.65)	0.15 (0.35)
6	1.911	0.048	0.499 (0.29)	-0.03 (5.45)	0.00 (0.34)	0.360	0.861 (0.14)	39.38 (6.59)	0.11 (0.30)
9	1.864	0.049	0.498 (0.28)	-0.15 (5.43)	0.00 (0.30)	0.423	0.848 (0.13)	38.14 (6.53)	0.08 (0.26)
12	1.825	0.050	0.502 (0.26)	0.11 (5.42)	-0.00 (0.27)	0.478	0.834 (0.13)	36.68 (6.41)	0.06 (0.23)
18	1.766	0.051	0.502 (0.24)	0.14 (5.44)	-0.00 (0.23)	0.565	0.807 (0.12)	33.72 (6.30)	0.04 (0.20)
36	1.682	0.050	0.500 (0.19)	0.26 (5.45)	0.00 (0.17)	0.712	0.714 (0.09)	23.25 (5.98)	0.01 (0.17)
(ii) Reference ($n_0^0 = 1$) vs. Sceptical ($n_1^0 = 8$)									
1	2.004	0.051	0.470 (0.32)	0.25 (5.44)	0.07 (0.43)	0.427	0.844 (0.18)	37.72 (6.51)	0.22 (0.41)
2	1.953	0.050	0.479 (0.31)	-0.37 (5.43)	0.05 (0.38)	0.491	0.833 (0.17)	36.22 (6.46)	0.17 (0.34)
4	1.913	0.050	0.484 (0.29)	-0.21 (5.43)	0.04 (0.33)	0.538	0.820 (0.17)	34.96 (6.38)	0.13 (0.29)
6	1.886	0.049	0.491 (0.28)	-0.49 (5.43)	0.03 (0.31)	0.565	0.811 (0.16)	33.92 (6.44)	0.10 (0.26)
9	1.856	0.051	0.494 (0.27)	-0.26 (5.45)	0.02 (0.28)	0.590	0.800 (0.16)	32.92 (6.40)	0.08 (0.24)
12	1.844	0.049	0.497 (0.26)	0.27 (5.43)	0.02 (0.25)	0.597	0.788 (0.15)	31.44 (6.41)	0.07 (0.22)
18	1.792	0.051	0.499 (0.24)	-0.20 (5.42)	0.02 (0.22)	0.650	0.771 (0.14)	29.61 (6.32)	0.05 (0.20)
36	1.715	0.052	0.483 (0.18)	0.24 (5.43)	0.01 (0.18)	0.710	0.717 (0.12)	23.69 (6.09)	0.02 (0.17)
(iii) Reference ($n_0^0 = 1$) vs. Enthusiastic ($n_1^0 = 8$)									
1	1.964	0.047	0.315 (0.24)	-0.00 (5.41)	0.14 (0.37)	0.176	0.920 (0.08)	45.86 (5.79)	0.24 (0.41)
2	1.908	0.052	0.323 (0.24)	-0.44 (5.45)	0.13 (0.34)	0.234	0.911 (0.08)	44.80 (5.83)	0.20 (0.36)
4	1.892	0.049	0.329 (0.23)	-0.02 (5.42)	0.11 (0.31)	0.275	0.902 (0.09)	43.88 (5.82)	0.17 (0.33)
6	1.872	0.050	0.332 (0.22)	0.24 (5.41)	0.10 (0.30)	0.313	0.896 (0.09)	43.38 (5.86)	0.15 (0.31)
9	1.864	0.048	0.337 (0.22)	0.15 (5.39)	0.09 (0.28)	0.348	0.887 (0.09)	42.32 (5.86)	0.13 (0.29)
12	1.834	0.052	0.338 (0.21)	-0.10 (5.39)	0.08 (0.27)	0.388	0.878 (0.09)	41.37 (5.87)	0.12 (0.28)
18	1.804	0.054	0.338 (0.20)	-0.19 (5.45)	0.07 (0.25)	0.444	0.865 (0.10)	39.83 (5.95)	0.10 (0.25)
36	1.761	0.050	0.324 (0.17)	0.08 (5.41)	0.05 (0.21)	0.515	0.836 (0.11)	36.62 (6.07)	0.05 (0.21)

Table 6.7.3: The effect of using archetypal priors on the performance of the FLGI for a two-armed trial when $\sigma_k^2 = 0.64^2$ is assumed known, $T = 72$ and $d = 0.995$, averaged over 50,000 trial replications.

Conclusions

The main conclusions to draw from these experiments are that when using the FLGI in the known variance case with a *sceptical* prior on the experimental arm, the power

of the design increases whilst the patient benefit measures decrease relative to the corresponding results when starting with the reference prior. This is what we would expect to observe because the sceptical prior implies that there is a 0.95 probability that μ_1 lies below 0.529 (as depicted in Figure 6.7.2) which is incorrect under $\mathcal{H}_1$, and as such it provides the FLGI algorithm with a 'false start'. Thus, it takes longer for the design to correctly identify the best arm, resulting in fewer patients allocated to the superior arm but a larger power due to less imbalance.

In contrast, when starting with an *enthusiastic* prior on the experimental arm, the reverse happens (as shown in (iii) of Table 6.7.3); the power decreases whilst the patient benefit measures increase (relative to starting with the reference prior). Again, this is not surprising because the enthusiastic prior specifies that the most likely value of μ_1 is 0.529 (as illustrated in Figure 6.7.2). Under $\mathcal{H}_1$, this is correct and so it gives the algorithm a 'head start' in the right direction meaning it identifies the superior arm quicker. Thus, less allocations are made to the control arm resulting in more imbalance and hence reduced power. Under $\mathcal{H}_0$, however, this prior specification on the experimental arm is incorrect and so the FLGI incorrectly allocates fewer patients to the control arm, as observed in (iii) of Table 6.7.3 (where the control arm is taken to be the 'superior' arm under $\mathcal{H}_0$). This explains why only $\approx 33\%$ of patients in the trial are allocated to the control arm for all block sizes under $\mathcal{H}_0$. Fewer observations on the control arm leads to an underestimation of $\hat{\mu}_0$ and consequently the treatment effect estimator under $\mathcal{H}_0$ exhibits bias. It is also worth noting that the variability in the allocations decreases when using the enthusiastic prior (relative to the reference prior) since, under $\mathcal{H}_1$, the observed data and prior information match which reduces the uncertainty of the allocations.

Overall, our recommendation is to be very cautious when incorporating prior information into bandit-based designs such as the FLGI because it influences the speed at which the design updates and favours an arm (depending on how informative the

prior is). Since these designs are so dynamic anyway, there is not as much to gain from using prior information as there may be with less responsive designs. If the prior specification is correct, then the incoming data will further enhance the effect of the prior and the design will favour the superior arm sooner, whereas if the prior is misspecified, the bandit may spend more time in the exploration phase or degenerate to allocating all patients to one arm. However, it is likely that the incoming data during the trial will eventually dilute the effect of the misspecified prior. How long the design takes to correct for the misspecification depends on the value of n_k^0; the greater its value, the more influence the prior will have. Therefore, if one wishes to use prior information in conjunction with the FLGI, we suggest setting a small value for n_k^0.

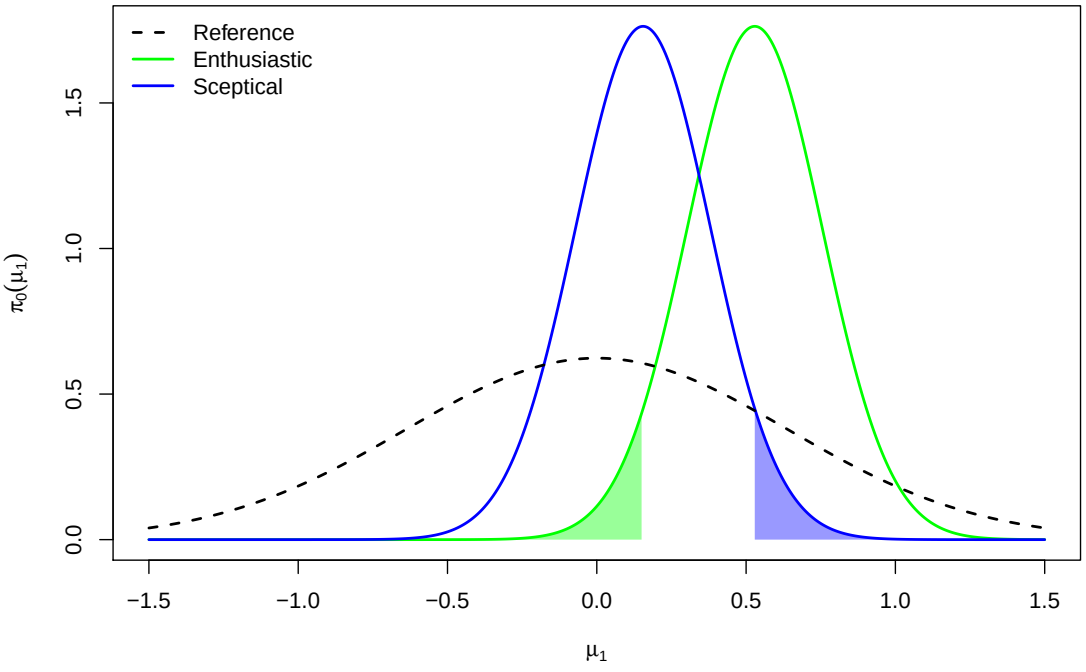


Figure 6.7.2: Sceptical and enthusiastic prior densities with the reference prior depicted in black. The sceptics' probability that the true mean response is greater than 0.529 (the alternative value) is 0.05, shown by the blue shaded region. The enthusiasts' probability that the true mean response is less than 0.155 (the null value) is also 0.05, shown by the green shaded region.

Chapter 7

Conclusions and Further Work

7.1 Summary and Contributions

This thesis connects two lines of work, namely, bandit methodology and clinical trial design. In particular, we have proposed a range of clinical trial designs which are based upon solutions to the multi-armed bandit problem (MABP) and thus share the same principal goal of maximising patient benefit within the trial. However, each design is intended to address a different issue that has been suggested as preventing such bandit-based designs from being implemented in practice. Below, we briefly outline each design in turn and highlight their main contributions. The main points are also summarised succinctly in Table 7.2.1.

7.1.1 Chapter 3, CRDP

In Chapter 3, we proposed the constrained randomised dynamic programming (CRDP) design, so called because we introduced: (i) a constraint to force a minimum number of patients on each arm, and (ii) randomisation into an otherwise deterministic design based on the optimal dynamic programming (DP) solution. Several performance measures of the proposed design were evaluated and compared to alternative designs

through extensive simulation studies using a recently published trial as motivation. For simplicity, a two-armed trial with binary endpoints and immediate responses was considered. Simulation results for the proposed design showed that: (i) the percentage of patients allocated to the superior arm is much higher than in the traditional fixed randomised design; (ii) relative to the optimal DP design, the power is largely improved upon and (iii) it exhibits only a very small bias and mean squared error of the treatment effect estimator.

7.1.2 Chapter 4, FCRDP

CRDP, as with most response-adaptive designs, hinges on the limiting assumption of patient responses being available before allocation of the next patient. This is one of the greatest challenges, both for clinical trial practice (Rosenberger *et al.*, 2012) and the bandit literature (Caro and Yoo, 2010). Therefore, in Chapter 4, not only do we study the impact of delayed responses on CRDP, but we take it one step further by extending the design for the fixed delay case (e.g. constant arrivals and fixed response time). This design is referred to as FCRDP. Simulation results revealed that CRDP continues to offer patient benefit (albeit less than in the immediate response case) even when the information during each adaptation is reduced due to the delay. It is therefore relatively robust to delayed responses. Nevertheless, implementation of FCRDP in the same scenarios showed that there are worthwhile patient benefit gains to be made, with minimal impact on the corresponding power, bias and mean squared error, by utilising the pipeline data in the updates of the allocation probabilities.

7.1.3 Chapter 5, RCRDP

Having a fixed number of patients in the pipeline is only representative of a small number of trials. Therefore, in Chapter 5, we extend the CRDP design to the most general case so that it can be applied to a greater variety of trials which encounter

a *random* number of patients in the pipeline (including those with random arrivals or random response times, or both). This aptly takes the initialism RCRDP. When implemented for a hypothetical trial with exponential inter-arrival times and a fixed follow-up time, RCRDP was shown to perform very similarly to FCRDP with respect to all performance measures.

Conclusion of DP-Based Designs

Overall, CRDP and its delayed variants (with suggested degree of randomisation $p = 0.9$ and constraining $\ell = 0.15n$) were found to strike a balance between the two conflicting objectives of patient benefit (individual ethics) and power (collective ethics), which is indicative of a "good clinical trial design" (Lee *et al.*, 2010) and "properly chosen RAR method" (Du *et al.*, 2015). By adjusting the constraint and/or degree of randomisation, CRDP provides a continuum of designs with DP and fixed randomisation at the extremes. As such, the design can be tailored to suit the individual objectives of the trial and attain the most appropriate balance. This greatly increases the prospects of a DP-based design being implemented in clinical trial practice.

One of the main practical limitations of the aforementioned DP-based designs is their associated computational expense (Jiang *et al.*, 2013) which grows exponentially with the patient horizon (Villar *et al.*, 2015a; Ahuja and Birge, 2019). However, this is not as prohibitive as it is commonly perceived to be in the literature (Jacko, 2019b, Section 7.1) which emphasises the importance of collaboration between disciplines. In particular, the recent survey by Jacko (2019b) demonstrates that a computer with 32GB RAM is able to optimally solve the two-armed Bernoulli bandit problem up to a trial size of 1440 or 4440, depending on whether storage of the optimal allocation policy is or is not required, respectively (see also Jacko, 2019a, for details of implementation).

The augmentation of the state space to include the delay parameters in the FCRDP and RCRDP designs increases the computational complexity even further. Consequently, such designs can only be applied to relatively small-scale trials. Nevertheless, for rare disease settings, which is where we anticipate that our proposed designs are most pertinent, this does not pose a serious problem. For example, a review by Bell and Smith (2014) found that 67% of rare disease trials had 0–50 patients, 19% had between 51–100 patients and only 14% had more than 100 patients (with just 1% over 500 patients). Therefore, even with the additional complexity caused by the delay, our designs maintain computational feasibility for the most commonly encountered sample sizes in rare disease trials. Moreover, the ideas from Jacko (2019a) could be applicable and useful for developing a code that can solve the delayed model for larger trials than those considered in this thesis.

7.1.4 Chapter 6, FLGI

Motivated by the fact that the "RAR literature dealing with continuous outcomes is much smaller and less developed" (Hu and Rosenberger, 2006; Flournoy *et al.*, 2013) and "the generalisation of [RAR] features to the multiple arm setting has been less explored" (Viele *et al.*, 2020), we propose a RAR design for multi-armed trials with continuous outcomes that are assumed to be normally distributed with unknown, non-homogeneous variances. This design is based on the Gittins index (GI) solution to the MABP, which we refer to as FLGI (the "FL-" indicative of its forward-looking nature). Compared to the DP solution, this approach is much simpler to communicate and be understood by all parties involved in the drug development process, including trial stakeholders, participants, etc. (Pallmann *et al.*, 2018). Moreover, it avoids the computational burden associated with DP-based designs, thus can easily be implemented in multi-armed trials, for example.

In contrast to previous chapters, we implement the GI-based design in a group

sequential setting. This ameliorates the logistical difficulties of having to update the randomisation probabilities after each patient has been observed (Chappell and Karrison, 2006), which is another main reason cited for the limited uptake of RAR designs (Karrison *et al.*, 2003; Wason *et al.*, 2019).

We illustrate the proposed procedure by simulations in the context of phase II cancer trials, and compare its performance against a variety of existing designs. Results show that there are efficiency and patient benefit gains of using RAR designs, such as FLGI, with a continuous endpoint instead of artificially dichotomising to form a binary one. These gains persist even if an anticipated low rate of missing data is imputed online using an approach suggested in this chapter. The effect of varying the prior information, as well as the discount factor, on the performance of FLGI is also evaluated. Additionally, we demonstrate that protecting allocation to the control arm continues to substantially improve patient benefit, whilst achieving similar power to the traditional FR, in multi-armed trials with normal outcomes.[1]

7.1.5 Areas Covered

Collectively, we have covered a broad range of topics from DP policies to index policies, two-armed trials to multi-armed trials, binary endpoints to continuous endpoints, sequential designs to group-sequential designs, 'non-informative' priors to informative priors, as well as the problem of missing data through either delayed responses or loss to follow-up. These are central to the development and application of bandit-based designs to clinical practice. However, not all of these issues have been covered by any one design which leaves scope for many natural extensions. In particular: the generalisation of the DP-based designs to endpoints other than binary, the evaluation of the DP-based designs in multi-armed settings, the application of the DP-based designs to a group-sequential setting, the incorporation of prior information into the

[1]Different control allocations for RAR designs in a binary response, multi-armed setting have recently been compared in Viele *et al.* (2020).

DP-based designs (by eliciting expert opinion or using historical data from a related trial, for example, see Hampson *et al.*, 2014), the incorporation of delayed responses into the GI-based design, the extension of FLGI to other endpoints (e.g. exponential), etc. Note that such practicalities can be taken into account by: (i) evaluating the original design via a simulation study which allows for the practicality of interest (as in Section 4.2 when exploring the impact of delays on CRDP, for example), or (ii) extending the modelling framework to incorporate the required practicality (as in Chapter 5 when extending the CRDP model to incorporate delayed responses, for example).

Moreover, there are several remaining challenges that we have not addressed in this thesis, some of which have already been highlighted within the relevant chapters so we do not repeat them here. We therefore conclude this thesis by suggesting some general ideas for further work which are applicable to all of the proposed designs.

7.2 Areas of Further Work

7.2.1 Joint Efficacy/Toxicity Outcome

Throughout this thesis, we have restricted attention to the efficacy outcome as an indication of whether the treatment has been successful or not. However, in practice, treatment toxicities should be monitored concurrently (Lee *et al.*, 2012). For example, what if a treatment is efficacious yet causes an adverse reaction in the patient; should this treatment still be considered a success? This motivates the joint evaluation of both efficacy and toxicity, particularly for trials testing treatments with a high risk of severe adverse side effects such as oncology trials. An example of an RAR design based on a joint efficacy/toxicity outcome is proposed by Ji and Bekele (2009). Adapting the bandit-based allocation rules proposed in this thesis to reflect both efficacy and toxicity outcomes remains an open problem.

7.2.2 Multiple Outcomes

Furthermore, as stated in Kaibel and Biemann (2019), response-adaptive MAB approaches, such as those developed in this thesis, also have the potential to determine randomisation probabilities for diseases based on multiple outcomes of interest. Within the multi-objective MAB literature, the linear scalarised function (Eichfelder, 2008), which transforms the vector of multiple outcomes into a single outcome, is a popular approach because of its simplicity (Yahyaa and Manderick, 2015). This would also allow for the outcome variables to be weighted according to their relevance, for example. Whether this is applicable to the clinical trial setting is an area to be explored.

7.2.3 Alternative Objective Functions

The standard bandit objective, which we have considered throughout this thesis, is to maximise the expected total reward (i.e. treatment effectiveness) over the time horizon. However, in a clinical trial context, this may not necessarily be the most desirable option since "controlling multiple properties of a design may not be easily achieved through a single utility function" (Zhang *et al.*, 2019). For example, a treatment which works best on average may also exhibit considerable variability, thus causing adverse side effects for some patients. In this case, a treatment which is less effective on average, but has a smaller variability so that its behaviour is more consistent amongst patients, may be preferred. Therefore, it may be better to explore alternative objective functions, such as the mean-variance model (Markowitz, 1952). The mean-variance bandit problem focuses on the problem of selecting the arm which effectively trades off its expected reward with its variability. Two algorithms have been proposed by Sani *et al.* (2012) to solve the mean-variance bandit problem and it would be interesting to evaluate their performance in the clinical trial setting.

Other objectives that could be incorporated are economical (e.g. the cost of treat-

ment, see Pertile *et al.* (2014) and Chick *et al.* (2017)), quality of life measures such as invasiveness (e.g. implications of surgical intervention versus a vaccination), or duration (e.g. is it better to undergo surgery once or receive life-long medication?).

7.2.4 Incorporation of Covariates

Our proposed designs were formulated under the assumption that patients allocated to the same treatment will have the same expected response. However, in practice this may be unreasonable if there are certain covariates (such as age and gender) which influence their response (Zhang *et al.*, 2007), for example. In this case, it may not be appropriate to use responses from all preceding patients to determine the current patient's randomisation probability since only a particular subset of the available responses may be clinically relevant. This motivates the use of covariate-adjusted RAR which generalises RAR to include a patient's covariate profile (Rosenberger and Lachin, 2016). More specifically, this means that the randomisation probabilities will not only depend on the history of patient responses, but also on the covariate information of previous patients and the current patient. Hence, different randomisation probabilities will be used for different patient subgroups (Meurer *et al.*, 2012). This contributes to the personalisation of treatment allocation during the trial, with more patients receiving the treatment most appropriate for them (Qiao *et al.*, 2019). As increasing numbers of biomarkers are being identified, particularly in cancer research, personalised medicine is gaining considerable attention, and hence there is a growing need for novel trial designs which can make use of this additional information. Consequently, how to incorporate covariate information into the proposed designs forms a particularly topical and promising area of further research.

7.2.5 Accounting for Patient Drift

Continuing with the theme of how to adequately reflect patient heterogeneity within the proposed designs raises the question of how to implement such designs in the presence of underlying time trends caused by a systematic change in patient characteristics during the trial. The possibility of so-called patient drift is another major criticism of using RAR (see Rosenberger *et al.*, 2012, Section 4.3) since if this is not taken into consideration, the parameter estimates may be biased which would erroneously imbalance the treatment allocation and inflate the type I error (e.g. Thall *et al.*, 2015). Sequential tests accounting for linear time trends have been proposed and investigated in Coad (1991a,b), and the effect of time trends on several response-adaptive rules has been examined in Coad (1992).

Time trends are more likely to occur in trials with a long duration, such as rare disease trials in which the recruitment period typically extends over a very long time (Villar *et al.*, 2018). Since the rare disease setting is where the proposed bandit-based designs are deemed to be most applicable, it would be useful to first investigate the impact of time trends on these designs before suggesting how this could be accounted for. This could also combine with the previous area of further research by including a particular time trend as a covariate (see e.g. Rosenberger *et al.*, 2001a).

7.2.6 Adding/Dropping Arms

"MABs offer the flexibility to add further treatments easily at any point in time" (Kaibel and Biemann, 2019). Therefore, further extensions could also explore the performance of bandit-based designs when additional arms are added to the trial. This would be relatively straightforward to implement for the GI-based design since the GI is independent of the number of arms (Villar *et al.*, 2015a). Similarly, the incorporation of early stopping rules (for efficacy or futility) (see e.g. Du *et al.*, 2015; Lee *et al.*, 2010) and optimal stopping times (see e.g. Chick *et al.*, 2017) could potentially

be investigated.

7.2.7 Dose-Finding Trials

Another direction for future research is applying and extending the concepts covered in this thesis to the context of dose-finding trials, in which the primary goal is to find the maximum tolerated dose (MTD) of a treatment. In dose-finding trials, the different arms represent different dose levels of a treatment, and hence the arms are now correlated. This induces additional computational complexity into the associated MABP (relative to the classic MABP with independent arms) which is prohibitive in most practical situations. However, since dose-finding trials are typically small in size, the additional computational complexity may be manageable in this context. A framework for dose-finding trials using the theory of bandit problems was suggested by Leung and Wang (2002), and generalised to include multiple outcomes and early termination by Fan and Wang (2006). More recently, the dose-finding problem has also been posed as a MABP in Kano *et al.* (2019) and Aziz *et al.* (2019). In this context, the trade-off is between finding the MTD and treating as many patients as possible with the MTD (whilst avoiding allocation to toxic doses).

It is hoped that the ideas raised in this thesis will prompt further development of the proposed designs and encourage more collaboration between researchers and practitioners.

Proposed design	Solution method	Simulation setting	Main limitations	Main issues addressed
CRDP Constrained randomised dynamic programming	Dynamic programming	Two arms, binary endpoint, sequential	Immediate responses, computationally intensive, logistically difficult	Lack of randomisation, insufficient power, biased estimates
FCRDP CRDP adjusted for *fixed* pipeline	Dynamic programming	Two arms binary endpoint, sequential	Sequential arrivals & fixed response time, computationally intensive, logistically difficult	Impact of delay, extension to fixed delays
RCRDP CRDP adjusted for *random* pipeline	Dynamic programming	Two arms, binary endpoint, sequential	Computationally intensive, logistically difficult	Extension to random delays
(C)FLGI (Controlled) Forward-looking Gittins index	Gittins index	Two & multiple arms, normal endpoint with unknown variance, group sequential	Immediate responses	Computational complexity, lack of randomisation, continuous endpoints, dichotomisation, missing data, effect of prior info.

Table 7.2.1: Overview of proposed designs.

Bibliography

Ahuja, V. and Birge, J. R. (2016). Response-adaptive designs for clinical trials: Simultaneous learning from multiple patients. *European Journal of Operational Research*, **248**(2), 619–633.

Ahuja, V. and Birge, J. R. (2019). An approximation approach for response adaptive clinical trial design. Technical report, SMU Cox School of Business Research Paper No. 18-26.

Akech, S. O., Karisa, J., Nakamya, P., Boga, M., and Maitland, K. (2010). Phase II trial of isotonic fluid resuscitation in Kenyan children with severe malnutrition and hypovolaemia. *BMC Pediatrics*, **10**(1), 1.

Amaral, J. A. F. P. (1985). *Aspects of Optimal Sequential Resource Allocation*. D.Phil. thesis. University of Oxford.

Amberson, J. B., McMahon, B. T., and Pinner, M. A. (1931). Clinical trial of sanocrysin in pulmonary tuberculosis. *American Review of Tuberculosis*, **24**(4), 401–435.

Antognini, A. B. and Giovagnoli, A. (2015). *Adaptive Designs for Sequential Treatment Allocation*. Chapman and Hall/CRC.

Armitage, P. (1985). The search for optimality in clinical trials. *International Statistical Review/Revue Internationale de Statistique*, **53**(1), 15–24.

Atkinson, A. C. and Biswas, A. (2014). *Randomised Response-Adaptive Designs in Clinical Trials*. CRC Press.

Aziz, M., Kaufmann, E., and Riviere, M.-K. (2019). On multi-armed bandit designs for phase I clinical trials. *arXiv preprint arXiv:1903.07082*.

Bai, Z. D., Hu, F., and Rosenberger, W. F. (2002). Asymptotic properties of adaptive designs for clinical trials with delayed response. *The Annals of Statistics*, **30**(1), 122–139.

Bandyopadhyay, U. and Biswas, A. (1996). Delayed response in randomized play-the-winner rule: A decision theoretic outlook. *Calcutta Statistical Association Bulletin*, **46**(1-2), 69–88.

Barker, A. D., Sigman, C. C., Kelloff, G. J., Hylton, N. M., Berry, D. A., and Esserman, L. J. (2009). I-SPY 2: An adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clinical Pharmacology & Therapeutics*, **86**(1), 97–100.

Bartlett, R. H., Roloff, D. W., Cornell, R. G., Andrews, A. F., Dillon, P. W., and Zwischenberger, J. B. (1985). Extacorporeal circulation in neonatal respiratory failure: A prospective randomized study. *Pediatrics*, **76**(4), 479–487.

Bather, J. (1980). Randomised allocation of treatments in sequential trials. *Advances in Applied Probability*, **12**(1), 174–182.

Bather, J. A. (1981). Randomized allocation of treatments in sequential experiments (with discussion). *Journal of the Royal Statistical Society: Series B (Methodological)*, **43**(3), 265–283.

Bauer, P., Koenig, F., Brannath, W., and Posch, M. (2010). Selection and bias — Two hostile brothers. *Statistics in Medicine*, **29**(1), 1–13.

Bell, S. A. and Smith, C. T. (2014). A comparison of interventional clinical trials in rare versus non-rare diseases: An analysis of ClinicalTrials.gov. *Orphanet Journal of Rare Diseases*, **9**(1), 9–170.

Bellman, R. (1956). A problem in the sequential design of experiments. *Sankhyā: The Indian Journal of Statistics*, **16**(3/4), 221–229.

Bellman, R. (1957). *Dynamic Programming*. Princeton University Press, Princeton, New Jersey.

Bellman, R. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton University Press.

Berger, V. W. (2015). A note on response-adaptive randomization. *Contemporary Clinical Trials*, **40**, 240.

Berry, D. A. (1976). The application of two-armed bandit strategies to clinical trials. Technical Report 265, University of Minnesota.

Berry, D. A. (1978). Modified two-armed bandit strategies for certain clinical trials. *Journal of the American Statistical Association*, **73**(362), 339–345.

Berry, D. A. (2011). Adaptive clinical trials: The promise and the caution. *Journal of Clinical Oncology*, **29**(6), 606–609.

Berry, D. A. and Eick, S. G. (1995). Adaptive assignment versus balanced randomization in clinical trials: A decision analysis. *Statistics in Medicine*, **14**(3), 231–246.

Berry, D. A. and Fristedt, B. (1985). *Bandit Problems: Sequential Allocation of Experiments*. Monographs on Statistics and Applied Probability. Chapman & Hall.

Berry, D. A. and Stangl, D. K. (1996). Bayesian methods in health-related research. In D. A. Berry and D. K. Stangl, editors, *Bayesian Biostatistics*, pages 3–66. Marcel Dekker.

Berry, S. M., Carlin, B. P., Lee, J. J., and Müller, P. (2011). *Bayesian Adaptive Methods for Clinical Trials*. Chapman & Hall/CRC Biostatistics Series. CRC press.

Biswas, A. (1999). Delayed response in randomized play-the-winner rule revisited. *Communications in Statistics - Simulation and Computation*, **28**(3), 715–731.

Biswas, A. and Bhattacharya, R. (2009). Optimal response-adaptive designs for normal responses. *Biometrical Journal*, **51**(1), 193–202.

Biswas, A. and Bhattacharya, R. (2016). Response-adaptive designs for continuous treatment responses in phase III clinical trials: A review. *Statistical Methods in Medical Research*, **25**(1), 81–100.

Biswas, A. and Coad, D. S. (2005). A general multi-treatment adaptive design for multivariate responses. *Sequential Analysis*, **24**(2), 139–158.

Biswas, A., Bhattacharya, R., and Zhang, L. (2007). Optimal response-adaptive designs for continuous responses in phase III trials. *Biometrical Journal*, **49**(6), 928–940.

Biswas, A., Bandyopadhyay, U., and Bhattacharya, R. (2008). Response-adaptive designs in phase III clinical trials. In A. Biswas, S. Datta, J. P. Fine, and M. R. Segal, editors, *Statistical Advances in the Biomedical Sciences: Clinical Trials, Epidemiology, Survival Analysis, and Bioinformatics*, Wiley Series in Probability and Statistics, pages 22–53. John Wiley & Sons, Inc.

Biswas, S., Liu, D. D., Lee, J. J., and Berry, D. A. (2009). Bayesian clinical trials at the University of Texas M. D. Anderson Cancer Center. *Clinical Trials*, **6**(3), 205–216.

Blackwell, D. and Hodges, J. L. (1957). Design for the control of selection bias. *Annals of Mathematical Statistics*, **28**(2), 449–460.

Bowden, J. and Glimm, E. (2008). Unbiased estimation of selected treatment means in two-stage trials. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, **50**(4), 515–527.

Bowden, J. and Trippa, L. (2017). Unbiased estimation for response adaptive clinical trials. *Statistical Methods in Medical Research*, **26**(5), 2376–2388.

Caro, F. and Gallien, J. (2007). Dynamic assortment with demand learning for seasonal consumer goods. *Management Science*, **53**(2), 276–292.

Caro, F. and Yoo, O. S. (2010). Indexability of bandit problems with response delays. *Probability in the Engineering and Informational Sciences*, **24**(3), 349374.

Carreras, M. and Brannath, W. (2013). Shrinkage estimation in two-stage adaptive designs with midtrial treatment selection. *Statistics in Medicine*, **32**(10), 1677–1690.

Chapelle, O. (2014). Modeling delayed feedback in display advertising. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 1097–1105, New York, NY, USA. ACM.

Chappell, R. and Karrison, T. (2006). Letter to the editor (in response to "Continuous Bayesian adaptive randomization based on event times with covariates"). *Statistics in Medicine*, **26**, 3050–3052.

Cheng, Y. and Berry, D. A. (2007). Optimal adaptive randomized designs for clinical trials. *Biometrika*, **94**(3), 673–689.

Cheung, Y. K., Inoue, L. Y. T., Wathen, J. K., and Thall, P. F. (2006). Continuous Bayesian adaptive randomization based on event times with covariates. *Statistics in Medicine*, **25**(1), 55–70.

Chick, S., Forster, M., and Pertile, P. (2017). A Bayesian decision theoretic model of sequential experimentation with delayed response. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **79**(5), 1439–1462.

Chow, S.-C. and Chang, M. (2012). *Adaptive Design Methods in Clinical Trials*. Chapman & Hall/CRC Biostatistics Series, second edition.

Chow, S.-C. and Liu, J.-P. (2014). *Design and Analysis of Clinical Trials: Concepts and Methodologies*. Wiley series in probability and statistics. John Wiley & Sons, Hoboken, N.J., third edition.

Coad, D. S. (1991a). Sequential estimation with data-dependent allocation and time trends. *Sequential Analysis*, **10**(1-2), 91–97.

Coad, D. S. (1991b). Sequential tests for an unstable response variable. *Biometrika*, **78**(1), 113–121.

Coad, D. S. (1992). A comparative study of some data-dependent allocation rules for Bernoulli data. *Journal of Statistical Computation and Simulation*, **40**(3–4), 219–231.

Coad, D. S. (1994). Estimation following sequential tests involving data-dependent treatment allocation. *Statistica Sinica*, **4**, 693–700.

Coad, D. S. (1995). Sequential allocation involving several treatments. In N. Flournoy and W. F. Rosenberger, editors, *Adaptive Designs*, volume 25 of *Lecture Notes – Monograph Series*, pages 95–109. Institute of Mathematical Statistics, Hayward, CA.

Coad, D. S. (2008). Response adaptive randomization. In R. B. D'Agostino, L. Sullivan, and J. Massaro, editors, *Wiley Encyclopedia of Clinical Trials*, pages 1–7. John Wiley & Sons, Inc.

Coad, D. S. and Ivanova, A. (2001). Bias calculations for adaptive urn designs. *Sequential Analysis*, **20**(3), 91–116.

Coad, D. S. and Rosenberger, W. F. (1999). A comparison of the randomized play-the-winner rule and the triangular test for clinical trials with binary responses. *Statistics in Medicine*, **18**(7), 761–769.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, **7**(3), 249–253.

Cohen, J. D., McClure, S. M., and Angela, J. Y. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **362**(1481), 933–942.

Du, Y., Wang, X., and Lee, J. J. (2015). Simulation study for evaluating the performance of response-adaptive randomization. *Contemporary Clinical Trials*, **40**, 15–25.

Edwards, J., Fearnhead, P., and Glazebrook, K. (2017). On the identification and mitigation of weaknesses in the knowledge gradient policy for multi-armed bandits. *Probability in the Engineering and Informational Sciences*, **31**(2), 239–263.

Efron, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika*, **58**(3), 403–417.

Ehsan, N. and Liu, M. (2004). On the optimality of an index policy for bandwidth allocation with delayed state observation and differentiated services. In *Proceedings IEEE INFOCOM*, volume 3, pages 1974–1983.

Eichfelder, G. (2008). *Adaptive Scalarization Methods in Multiobjective Optimization*. Vector optimization. Springer-Verlag Berlin Heidelberg.

Eick, S. G. (1988a). Gittins procedures for bandits with delayed responses. *Journal of the Royal Statistical Society: Series B (Methodological)*, **50**(1), 125–132.

Eick, S. G. (1988b). The two-armed bandit with delayed responses. *The Annals of Statistics*, **16**(1), 254–264.

Eisele, J. R. (1994). The doubly adaptive biased coin design for sequential clinical trials. *Journal of Statistical Planning and Inference*, **38**(2), 249–261.

Eisenhauer, E. A., Therasse, P., Bogaerts, J., Schwartz, L. H., Sargent, D., Ford, R., Dancey, J., Arbuck, S., Gwyther, S., Mooney, M., *et al.* (2009). New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *European Journal of Cancer*, **45**(2), 228–247.

Facey, K. M. (1992). A sequential procedure for a phase II efficacy trial in hypercholesterolemia. *Controlled Clinical Trials*, **13**(2), 122–133.

Fan, S. K. and Wang, Y.-G. (2006). Decision-theoretic designs for dose-finding clinical trials with multiple outcomes. *Statistics in Medicine*, **25**(10), 1699–1714.

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain*, **33**, 503–513.

Flournoy, N., Haines, L. M., and Rosenberger, W. F. (2013). A graphical comparison of response-adaptive randomization procedures. *Statistics in Biopharmaceutical Research*, **5**(2), 126–141.

Freedman, B. (1987). Equipoise and the ethics of clinical research. *New England Journal of Medicine*, **317**(3), 141–145.

Gittins, J., Glazebrook, K., and Weber, R. (2011). *Multi-armed Bandit Allocation Indices*. John Wiley & Sons, second edition.

Gittins, J. C. (1979). Bandit processes and dynamic allocation indices (with discussion). *Journal of the Royal Statistical Society: Series B (Methodological)*, **41**(2), 148–177.

Gittins, J. C. and Jones, D. M. (1974). A dynamic allocation index for the sequential design of experiments. *Progress in Statistics*, **9**, 241–266.

Gittins, J. C. and Jones, D. M. (1979). A dynamic allocation index for the discounted multiarmed bandit problem. *Biometrika*, **66**(3), 561–565.

Glazebrook, K. D. (1978). On the optimal allocation of two or more treatments in a controlled clinical trial. *Biometrika*, **65**(2), 335–340.

Glazebrook, K. D. (1980). On randomized dynamic allocation indices for the sequential design of experiments. *Journal of the Royal Statistical Society: Series B (Methodological)*, **42**(3), 342–346.

Gu, X., Chen, N., Wei, C., Liu, S., Papadimitrakopoulou, V. A., Herbst, R. S., and Lee, J. J. (2016). Bayesian two-stage biomarker-based adaptive design for targeted therapy development. *Statistics in Biosciences*, **8**(1), 99–128.

Gwise, T. E., Zhou, J., and Hu, F. (2011). An optimal response adaptive biased coin design with k heteroscedastic treatments. *Journal of Statistical Planning and Inference*, **141**(1), 235–242.

Hall, N. S. (2007). R.A. Fisher and his advocacy of randomization. *Journal of the History of Biology*, **40**(2), 295–325.

Hampson, L. V. and Jennison, C. (2013). Group sequential tests for delayed responses (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **75**(1), 3–54.

Hampson, L. V., Whitehead, J., Eleftheriou, D., and Brogan, P. (2014). Bayesian methods for the design and interpretation of clinical trials in very rare diseases. *Statistics in Medicine*, **33**(24), 4186–4201.

Hardwick, J., Oehmke, R., and Stout, Q. F. (2006). New adaptive designs for delayed response models. *Journal of Statistical Planning and Inference*, **136**(6), 1940–1955.

Hardwick, J. P. (1995). A modified bandit as an approach to ethical allocation in clinical trials. In N. Flournoy and W. F. Rosenberger, editors, *Adaptive Designs*, volume 25 of *Lecture Notes – Monograph Series*, pages 65–87. Institute of Mathematical Statistics, Hayward, CA.

Hardwick, J. P. and Stout, Q. F. (1991). Bandit strategies for ethical sequential allocation. *Computing Science and Statistics*, **23**, 421–424.

Hey, S. P. and Kimmelman, J. (2015). Are outcome-adaptive allocation trials ethical? *Clinical Trials*, **12**(2), 102–106.

Hu, F. (2012). Statistical issues in trial design and personalized medicine. *Clinical Investigation*, **2**(2), 121–124.

Hu, F. and Rosenberger, W. F. (2003). Optimality, variability, power: Evaluating response-adaptive randomization procedures for treatment comparisons. *Journal of the American Statistical Association*, **98**(463), 671–678.

Hu, F. and Rosenberger, W. F. (2006). *The Theory of Response-Adaptive Randomization in Clinical Trials*. John Wiley & Sons.

Hu, F. and Zhang, L.-X. (2004). Asymptotic properties of doubly adaptive biased coin designs for multitreatment clinical trials. *The Annals of Statistics*, **32**(1), 268–301.

Hu, F., Rosenberger, W. F., and Zhang, L.-X. (2006). Asymptotically best response-adaptive randomization procedures. *Journal of Statistical Planning and Inference*, **136**(6), 1911–1922.

Hu, F., Zhang, L.-X., Cheung, S. H., and Chan, W. S. (2008). Doubly adaptive biased coin designs with delayed responses. *Canadian Journal of Statistics*, **36**(4), 541–559.

Hu, F., Zhang, L.-X., and He, X. (2009a). Efficient randomized-adaptive designs. *The Annals of Statistics*, **37**(5A), 2543–2560.

Hu, F., Zhang, L.-X., and He, X. (2009b). Efficient randomized-adaptive designs. *The Annals of Statistics*, **37**(5A), 2543–2560.

Ivanova, A. (2003). A play-the-winner-type urn design with reduced variability. *Metrika*, **58**, 1–13.

Ivanova, A. and Rosenberger, W. F. (2000). A comparison of urn designs for randomized clinical trials of $k > 2$ treatments. *Journal of Biopharmaceutical Statistics*, **10**(1), 93–107.

Jacko, P. (2019a). BinaryBandit: An efficient Julia package for optimization and evaluation of the finite-horizon bandit problem with binary responses. Working Paper 4, Lancaster University Management School.

Jacko, P. (2019b). The finite-horizon two-armed bandit problem with binary responses: A multidisciplinary survey of the history, state of the art, and myths. Working paper, Lancaster University Management School.

Jaki, T., André, V., Su, T., and Whitehead, J. (2013). Designing exploratory cancer trials using change in tumour size as primary endpoint. *Statistics in Medicine*, **32**(15), 2544–2554.

Jameson, G. J. O. (2016). The incomplete gamma functions. *The Mathematical Gazette*, **100**(548), 298–306.

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

Jeon, Y. and Hu, F. (2010). Optimal adaptive designs for binary response trials with three treatments. *Statistics in Biopharmaceutical Research*, **2**(3), 310–318.

Ji, Y. and Bekele, B. N. (2009). Adaptive randomization for multiarm comparative clinical trials based on joint efficacy/toxicity outcomes. *Biometrics*, **65**(3), 876–884.

Jiang, F., Jack Lee, J., and Müller, P. (2013). A Bayesian decision-theoretic sequential response-adaptive randomization design. *Statistics in Medicine*, **32**(12), 1975–1994.

Jones, D. M. (1970). *Search Procedures for Industrial Chemical Research*. Master's thesis, U.C.W. Aberystwyth.

Jones, D. M. (1975). *Search Procedures for Industrial Chemical Research*. Ph.D. thesis, University of Cambridge.

Kaibel, C. and Biemann, T. (2019). Rethinking the gold standard with multi-armed bandits: Machine learning allocation algorithms for experiments. *Organizational Research Methods*, pages 1–26.

Kano, H., Honda, J., Sakamaki, K., Matsuura, K., Nakamura, A., and Sugiyama, M. (2019). Good arm identification via bandit feedback. *Machine Learning*, **108**(5), 721–745.

Karrison, T. G., Huo, D., and Chappell, R. (2003). A group sequential, response-adaptive design for randomized clinical trials. *Controlled Clinical Trials*, **24**(5), 506–522.

Karrison, T. G., Maitland, M. L., Stadler, W. M., and Ratain, M. J. (2007). Design of phase II cancer trials using a continuous endpoint of change in tumor size: application to a study of sorafenib and erlotinib in non–small-cell lung cancer. *Journal of the National Cancer Institute*, **99**(19), 1455–1461.

Kateri, M. (2014). *Contingency Table Analysis: Methods and Implementation Using R*. Statistics for Industry and Technology. Birkhäuser.

Kelly, F. P. (1981). Multi-armed bandits with discount factor near one: The Bernoulli case. *The Annals of Statistics*, **9**(5), 987–1001.

Kim, E. S., Herbst, R. S., Wistuba, I. I., Lee, J. J., Blumenschein, G. R., Tsao, A., Stewart, D. J., Hicks, M. E., Erasmus, J., Gupta, S., Alden, C. M., Liu, S., Tang, X., Khuri, F. R., Tran, H. T., Johnson, B. E., Heymach, J. V., Mao, L., Fossella, F., Kies, M. S., Papadimitrakopoulou, V., Davis, S. E., Lippman, S. M., and Hong, W. K. (2011). The BATTLE trial: Personalizing therapy for lung cancer. *Cancer Discovery*, **1**(1), 44–53.

Korn, E. L. and Freidlin, B. (2011). Outcome-adaptive randomization: Is it useful? *Journal of Clinical Oncology*, **29**(6), 771.

Kuleshov, V. and Precup, D. (2000). Algorithms for the multi-armed bandit problem. *Journal of Machine Learning Research*, **1**, 1–32.

Langenberg, P. and Srinivasan, R. (1981). On the Colton model for clinical trials with delayed observations — Normally-distributed responses. *Biometrics*, **37**(1), 143–148.

Langenberg, P. and Srinivasan, R. (1982). On the Colton model for clinical trials with delayed observations — Dichotomous responses. *Biometrical Journal*, **24**(3), 287–296.

Lattimore, T. and Szepesvári, C. (2019). *Bandit Algorithms*. Cambridge University Press. Forthcoming.

Lavin, P. T. (1981). An alternative model for the evaluation of antitumor activity. *Cancer Clinical Trials*, **4**(4), 451–457.

Lee, J. J., Gu, X., and Liu, S. (2010). Bayesian adaptive randomization designs for targeted agent development. *Clinical Trials*, **7**(5), 584–596.

Lee, J. J., Chen, N., and Yin, G. (2012). Worth adapting? Revisiting the usefulness of outcome-adaptive randomization. *Clinical Cancer Research*, **18**(17), 4498–4507.

Lee, K. M. and Wason, J. (2019). Design of experiments for a confirmatory trial of precision medicine. *Journal of Statistical Planning and Inference*, **199**, 179 – 187.

Lellouch, J. and Schwartz, D. (1971). L'essai therapeutique: Eethique individuelle ou ethique collective? *Revue de l'Institut International de Statistique/Review of the International Statistical Institute*, **39**(2), 127–136.

Leung, D. H.-Y. and Wang, Y.-G. (2002). An extension of the continual reassessment method using decision theory. *Statistics in Medicine*, **21**(1), 51–63.

Lipsky, A. M. and Lewis, R. J. (2013). Response-adaptive decision-theoretic trial design: Operating characteristics and ethics. *Statistics in Medicine*, **32**(21), 3752–3765.

London, A. J. (2018). Learning health systems, clinical equipoise and the ethics of response adaptive randomisation. *Journal of Medical Ethics*, **44**(6), 409–415.

Maccallum, R. C., Zhang, S., Preacher, K. J., and Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, **7**(1), 19–40.

Magirr, D. (2011). Block response-adaptive randomization in clinical trials with binary endpoints. *Pharmaceutical Statistics*, **10**(4), 341–346.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, **7**(1), 77–91.

McEvoy, B., Haidar, D., Tehranisa, J., and Meurer, W. J. (2016). Optimizing communication of emergency response adaptive randomization clinical trials to potential participants. *bioRxiv*.

Melfi, V. F. and Page, C. (2000). Estimation after adaptive allocation. *Journal of Statistical Planning and Inference*, **87**(2), 353–363.

Meurer, W. J., Lewis, R. J., and Berry, D. A. (2012). Adaptive clinical trials: A partial remedy for the therapeutic misconception? *Journal of the American Medical Association*, **307**(22), 2377–2378.

Nowacki, A. S., Zhao, W., and Palesch, Y. Y. (2017). A surrogate-primary replacement algorithm for response-adaptive randomization in stroke clinical trials. *Statistical Methods in Medical Research*, **26**(3), 1078–1092.

Pallmann, P., Bedding, A. W., Choodari-Oskooei, B., Dimairo, M., Flight, L., Hampson, L. V., Holmes, J., Mander, A. P., Sydes, M. R., Villar, S. S., Wason, J. M. S., Weir, C. J., Wheeler, G. M., Yap, C., and Jaki, T. (2018). Adaptive designs in clinical trials: Why use them, and how to run and report them. *BMC Medicine*, **16**(1), 29.

Palmer, C. R. and Rosenberger, W. F. (1999). Ethics and practice: Alternative designs for phase III randomized clinical trials. *Controlled Clinical Trials*, **20**(2), 172 – 186.

Papadimitrakopoulou, V., Lee, J. J., Wistuba, I. I., Tsao, A. S., Fossella, F. V., Kalhor, N., Gupta, S., Byers, L. A., Izzo, J. G., Gettinger, S. N., Goldberg, S. B.,

Tang, X., Miller, V. A., Skoulidis, F., Gibbons, D. L., Shen, L., Wei, C., Diao, L., Peng, S. A., Wang, J., Tam, A. L., Coombes, K. R., Koo, J. S., Mauro, D. J., Rubin, E. H., Heymach, J. V., Hong, W. K., and Herbst, R. S. (2016). The BATTLE-2 study: A biomarker-integrated targeted therapy study in previously treated patients with advanced non-small-cell lung cancer. *Journal of Clinical Oncology*, **34**(30), 3638.

Peace, K. E. and Chen, D.-G. D. (2010). *Clinical Trial Methodology*. Chapman & Hall/CRC Biostatistics Series.

Perchet, V., Rigollet, P., Chassang, S., and Snowberg, E. (2016). Batched bandit problems. *The Annals of Statistics*, **44**(2), 660–681.

Pertile, P., Forster, M., and Torre, D. L. (2014). Optimal Bayesian sequential sampling rules for the economic evaluation of health technologies. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **177**(2), 419–438.

Pocock, S. J. (1983). *Clinical Trials: A Practical Approach*. John Wiley & Sons.

Powell, W. B. (2011). *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. John Wiley & Sons, second edition.

Pullman, D. and Wang, X. (2001). Adaptive designs, informed consent, and the ethics of research. *Controlled Clinical Trials*, **22**(3), 203–210.

Puterman, M. L. (2014). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.

Qiao, W., Ning, J., and Huang, X. (2019). A clinical trial design with covariate-adjusted response-adaptive randomization using superiority confidence of treatments. *Statistics in Biopharmaceutical Research*, **11**(4), 336–347.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, **58**(5), 527–535.

Robertson, D. S. (2016). *Correcting for Selection Bias in the Analysis of Multi-stage Trials*. Ph.D. thesis, University of Cambridge.

Robertson, D. S. and Glimm, E. (2019). Conditionally unbiased estimation in the normal setting with unknown variances. *Communications in Statistics - Theory and Methods*, **48**(3), 616–627.

Robertson, D. S. and Wason, J. M. S. (2019). Familywise error control in multi-armed response-adaptive trials. *Biometrics*, **75**(3), 885–894.

Robinson, D. (1983). A comparison of sequential treatment allocation rules. *Biometrika*, **70**(2), 492–495.

Rosenberger, W. F. (1996). New directions in adaptive designs. *Statistical Science*, **11**(2), 137–149.

Rosenberger, W. F. (1999). Randomized play-the-winner clinical trials: Review and recommendations. *Controlled Clinical Trials*, **20**(4), 328–342.

Rosenberger, W. F. and Hu, F. (2004). Maximizing power and minimizing treatment failures in clinical trials. *Clinical Trials*, **1**(2), 141–147.

Rosenberger, W. F. and Lachin, J. (2016). *Randomization in Clinical Trials: Theory and Practice*. John Wiley & Sons, second edition.

Rosenberger, W. F. and Lachin, J. M. (1993). The use of response-adaptive designs in clinical trials. *Controlled Clinical Trials*, **14**(6), 471–484.

Rosenberger, W. F. and Seshaiyer, P. (1997). Adaptive survival trials. *Journal of Biopharmaceutical Statistics*, **7**(4), 617–624.

Rosenberger, W. F., Vidyashankar, A., and Agarwal, D. K. (2001a). Covariate-adjusted response-adaptive designs for binary response. *Journal of Biopharmaceutical Statistics*, **11**(4), 227–236.

Rosenberger, W. F., Stallard, N., Ivanova, A., Harper, C. N., and Ricks, M. L. (2001b). Optimal adaptive designs for binary response trials. *Biometrics*, **57**(3), 909–913.

Rosenberger, W. F., Sverdlov, O., and Hu, F. (2012). Adaptive randomization for clinical trials. *Journal of Biopharmaceutical Statistics*, **22**(4), 719–736.

Rosenberger, W. F., Uschner, D., and Wang, Y. (2019). Randomization: The forgotten component of the randomized clinical trial. *Statistics in Medicine*, **38**(1), 1–12.

Routledge, R. (2005). Fisher's exact test. In P. Armitage and T. Colton, editors, *Encyclopedia of Biostatistics*. American Cancer Society.

Royston, P., Altman, D. G., and Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine*, **25**(1), 127–141.

Sani, A., Lazaric, A., and Munos, R. (2012). Risk-aversion in multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 3275–3283.

Simon, R. (1977). Adaptive treatment assignment methods and clinical trials. *Biometrics*, **33**(4), 743–749.

Smith, A. L. and Villar, S. S. (2018). Bayesian adaptive bandit-based designs using the Gittins index for multi-armed trials with normally distributed endpoints. *Journal of Applied Statistics*, **45**(6), 1052–1076.

Spiegelhalter, D. J., Freedman, L. S., and Parmar, M. K. B. (1994). Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **157**(3), 357–387.

Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. John Wiley & Sons.

Stallard, N. and Rosenberger, W. F. (2002). Exact group-sequential designs for clinical trials with randomized play-the-winner allocation. *Statistics in Medicine*, **21**(4), 467–480.

Sully, B. G., Julious, S. A., and Nicholl, J. (2013). A reinvestigation of recruitment to randomised, controlled, multicenter trials: A review of trials funded by two UK funding agencies. *Trials*, **14**(1), 166.

Sutton, R. S. and Barto, A. G. (2017). *Reinforcement Learning: An Introduction*. MIT press, Cambridge, second edition. Forthcoming.

Sverdlov, O., editor (2015). *Modern Adaptive Randomized Clinical Trials: Statistical and Practical Aspects*. Chapman & Hall/CRC Biostatistics Series. CRC Press.

Sverdlov, O., Ryeznik, Y., and Wong, W. K. (2012). Doubly adaptive biased coin designs for balancing competing objectives in time-to-event trials. *Statistics and Its Interface*, **5**(4), 401–413.

Tamura, R. N., Faries, D. E., Andersen, J. S., and Heiligenstein, J. H. (1994). A case study of an adaptive clinical trial in the treatment of out-patients with depressive disorder. *Journal of the American Statistical Association*, **89**(427), 768–776.

Tehranisa, J. S. and Meurer, W. J. (2014). Can response-adaptive randomization increase participation in acute stroke trials? *Stroke*, **45**(7), 2131–2133.

Medical Research Council (1948). Streptomycin treatment of pulmonary tuberculosis: A Medical Research Council investigation. *British Medical Journal*, **2**(4582), 769–782.

U.S. Food and Drug Administration (2004). Innovation/stagnation: Challenge and opportunity on the critical path to new medical products. http://wayback.archive-it.org/7993/20180125032208/https://www.fda.gov/ScienceResearch/SpecialTopics/CriticalPathInitiative/CriticalPathOpportunitiesReports/ucm077262.htm.

U.S. Food and Drug Administration (2006). Critical path opportunities report. http://wayback.archive-it.org/7993/20180125142845/https://www.fda.gov/downloads/ScienceResearch/SpecialTopics/CriticalPathInitiative/CriticalPathOpportunitiesReports/UCM077254.pdf.

U.S. Food and Drug Administration (2018). Adaptive designs for clinical trials of drugs and biologics: Draft guidance for industry. https://www.fda.gov/media/78495/download.

Thall, P., Fox, P., and Wathen, J. (2015). Statistical controversies in clinical research: Scientific and ethical problems with adaptive randomization in comparative clinical trials. *Annals of Oncology*, **26**(8), 1621–1628.

Thall, P. F. and Wathen, J. K. (2007). Practical Bayesian adaptive randomisation in clinical trials. *European Journal of Cancer*, **43**(5), 859–866.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, **25**(3/4), 285–294.

Trippa, L., Lee, E. Q., Wen, P. Y., Batchelor, T. T., Cloughesy, T., Parmigiani, G., and Alexander, B. M. (2012). Bayesian adaptive randomized trial design for patients with recurrent glioblastoma. *Journal of Clinical Oncology*, **30**(26), 3258.

Tymofyeyev, Y., Rosenberger, W. F., and Hu, F. (2007). Implementing optimal allocation in sequential binary response experiments. *Journal of the American Statistical Association*, **102**(477), 224–234.

Upton, G. J. G. and Lee, R. D. (1981). Discussion of 'Randomized allocation of treatments in sequential experiments' (by J.A. Bather). *Journal of the Royal Statistical Society: Series B (Methodological)*, **43**(3), 291.

Vernade, C., Cappé, O., and Perchet, V. (2017). Stochastic bandit models for delayed conversions. In *Conference on Uncertainty in Artificial Intelligence*.

Viele, K., Broglio, K., McGlothlin, A., and Saville, B. R. (2020). Comparison of methods for control allocation in multiple arm studies using response adaptive randomization. *Clinical Trials*, **17**(1), 52–60.

Villar, S. S. (2018). Bandit strategies evaluated in the context of clinical trials in rare life-threatening diseases. *Probability in the Engineering and Informational Science*, **32**(2), 229–245.

Villar, S. S. and Rosenberger, W. F. (2018). Covariate-adjusted response-adaptive randomization for multi-arm clinical trials using a modified forward looking Gittins index rule. *Biometrics*, **74**(1), 49–57.

Villar, S. S., Bowden, J., and Wason, J. (2015a). Multi-armed bandit models for the optimal design of clinical trials: Benefits and challenges. *Statistical Science*, **30**(2), 199–215.

Villar, S. S., Bowden, J., and Wason, J. (2015b). Response-adaptive randomisation for multi-arm clinical trials using the forward looking Gittins index rule. *Biometrics*, **71**(4), 969–978.

Villar, S. S., Bowden, J., and Wason, J. (2018). Response-adaptive designs for binary responses: How to offer patient benefit while being robust to time trends? *Pharmaceutical Statistics*, **17**(2), 182–197.

Wang, X. (2000). A bandit process with delayed responses. *Statistics & Probability Letters*, **48**(3), 303–307.

Wang, X. (2002). Asymptotic properties of bandit processes with geometric responses. *Statistics & Probability Letters*, **60**(2), 211–217.

Wang, Y.-G. (1991a). Gittins indices and constrained allocation in clinical trials. *Biometrika*, **78**(1), 101–111.

Wang, Y.-G. (1991b). Sequential allocation in clinical trials. *Communications in Statistics - Theory and Methods*, **20**(3), 791–805.

Wason, J. and Trippa, L. (2014). A comparison of Bayesian adaptive randomization and multi-stage designs for multi-arm clinical trials. *Statistics in Medicine*, **33**(13), 2206–2221.

Wason, J. M. and Jaki, T. (2016). A review of statistical designs for improving the efficiency of phase II studies in oncology. *Statistical Methods in Medical Research*, **25**(3), 1010–1021.

Wason, J. M. and Jaki, T. (2018). Multi-arm multi-stage trials can improve the efficiency of finding effective treatments for stroke: a case study. *BMC Cardiovascular Disorders*, **18**(1).

Wason, J. M., Mander, A. P., and Eisen, T. G. (2011). Reducing sample sizes in two-stage phase II cancer trials by using continuous tumour shrinkage end-points. *European Journal of Cancer*, **47**(7), 983 – 989.

Wason, J. M. S., Abraham, J. E., Baird, R. D., Gournaris, I., Vallier, A.-L., Brenton, J. D., Earl, H. M., and Mander, A. P. (2015). A Bayesian adaptive design for biomarker trials with linked treatments. *British Journal of Cancer*, **113**(5), 699–705.

Wason, J. M. S., Brocklehurst, P., and Yap, C. (2019). When to keep it simple — Adaptive designs are not always useful. *BMC Medicine*, **17**(152), 1–7.

Weber, R. and Weiss, G. (1990). On an index policy for restless bandits. *Journal of Applied Probability*, **27**(3), 637–648.

Wei, L. J. (1988). Exact two-sample permutation tests based on the randomized play-the-winner rule. *Biometrika*, **75**(3), 603–606.

Wei, L. J. and Durham, S. (1978). The randomized play-the-winner rule in medical trials. *Journal of the American Statistical Association*, **73**(364), 840–843.

Whitehead, J. (1993). Application of sequential methods to a phase III clinical trial in stroke. *Drug Information Journal*, **27**(3), 733–740.

Whittle, P. (1980). Multi-armed bandits and the Gittins index. *Journal of the Royal Statistical Society: Series B (Methodological)*, **42**(2), 143–149.

Whittle, P. (1982). *Optimization Over Time: Dynamic Programming and Stochastic Control*, volume 1 of *Wiley Series in Probability and Statistics - Applied Probability and Statistics Section*. John Wiley & Sons, New York.

Whittle, P. (1988). Restless bandits: Activity allocation in a changing world. In J. Gani, editor, *A Celebration of Applied Probability*, volume 25A of *Journal of Applied Probability*, pages 287–298. Applied Probability Trust.

Williamson, S. F., Jacko, P., Villar, S. S., and Jaki, T. (2017). A Bayesian adaptive design for clinical trials in rare diseases. *Computational Statistics and Data Analysis*, **113**, 136 – 153.

Xu, J. and Yin, G. (2014). Two-stage adaptive randomization for delayed response in clinical trials. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **63**(4), 559–578.

Yahyaa, S. Q. and Manderick, B. (2015). Thompson sampling for multi-objective multi-armed bandits problem. In *Proceedings: 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 47–52. Presses universitaires de Louvain.

Yin, G. (2012). *Clinical Trial Design: Bayesian and Frequentist Adaptive Methods*. John Wiley & Sons, Hoboken, New Jersey.

Zelen, M. (1969). Play the winner rule and the controlled clinical trial. *Journal of the American Statistical Association*, **64**(325), 131–146.

Zhang, L. and Rosenberger, W. F. (2006). Response-adaptive randomization for clinical trials with continuous outcomes. *Biometrics*, **62**(2), 562–569.

Zhang, L. and Rosenberger, W. F. (2007). Response-adaptive randomization for survival trials: The parametric approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **56**(2), 153–165.

Zhang, L.-X., Hu, F., Cheung, S. H., and Chan, W. S. (2007). Asymptotic properties of covariate-adjusted response-adaptive designs. *The Annals of Statistics*, **35**(3), 1166–1182.

Zhang, L.-X., Hu, F., Cheung, S. H., and Chan, W. S. (2011). Immigrated urn models — Theoretical properties and applications. *The Annals of Statistics*, **39**(1), 643–671.

Zhang, Y., Trippa, L., and Parmigiani, G. (2019). Frequentist operating characteristics of Bayesian optimal designs via simulation. *Statistics in Medicine*, **38**(21), 4026–4039.

Zhou, X., Liu, S., Kim, E. S., Herbst, R. S., and Lee, J. J. (2008). Bayesian adaptive design for targeted therapy development in lung cancer — A step toward personalized medicine. *Clinical Trials*, **5**(3), 181–193.

Zhu, H. and Hu, F. (2009). Implementing optimal allocation for sequential continuous responses with multiple treatments. *Journal of Statistical Planning and Inference*, **139**(7), 2420–2430.