Investigating the impact of oncology phase II trial design parameters on their ability to successfully screen new treatments

Nada Elbeltagi

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

The University of Leeds

School of Medicine and Health

June 2022

The candidate confirms that the work submitted is her own and that appropriate credit has been given where reference has been made to the work of others.

Acknowledgements

بسم الله الرحمن الرحيم

In the name of Allah, the most merciful, the most compassionate. All praise is first due to Allah, only by His help do actions come to fruition. I thank Allah for all the blessings he has given me, and for the strength and ability to complete this thesis. I have doubted myself many times along the way, but faith in You guided me till the very end.

To show true gratitude to Allah, I must also thank the people He placed in my life to help me on this journey. Firstly, I would like to thank my supervisors, Sarah Brown, Julia Brown, Nigel Stallard and Fiona Collinson for their continuous support and guidance; without them this research would not be what it is. Specifically, I want to thank Sarah for her expert advice and for her contagious energy that always encouraged me to keep going, Julia for always keeping me on track, Nigel for his breadth of knowledge and expertise which he shared with me when I could not see a way, and Fiona for her medical knowledge that guided me to know just where this research could be of use.

I would also like to thank my post-graduate tutors Rebecca Walwyn and Susanne Coleman for their advice and encouragement to keep going.

Last, but certainly not least, I would like to thank my family: thank you to my Mum to whom I owe all that I am; your prayers, support and encouragement has seen me through so much. I also thank my sisters, Hadir, Rewaa and Menna, and my brother Abdelrahman for listening to all my worries, even when I was becoming repetitive, and always encouraging me; for that I will be eternally grateful. I feel I would be amiss if I do not mention the newest addition to the family, Hadir’s son, Mustafa, who has brought so much joy and happiness to my heart. In dark times, his smile brought the light. I would also like to thank my Dad, who passed away knowing that this was the next journey in my life but never got to share it with me. Even in death you somehow still inspire me to be a better version of myself. Your memory and how proud this would have made you feel spurred me to keep going until the end.

Abstract

Introduction Phase III oncology trials have significantly high attrition rates, where many treatments fail to show efficacy over standard treatments. Design of phase II trials contribute to these inefficiencies and there is much debate regarding optimal phase II design. The effect of the relationship between phase II and III trial endpoints, randomisation, using one-stage or two stages and the operating characteristics of phase II trials in oncology, on the efficiency of the phase II and III process, are all investigated in this thesis.

Methods Evaluation of design parameters was based on simulating multiple phase II and III trials until a successful phase III trial is observed, assuming many treatments are available for testing. Phase II and III trials were conducted assuming the true effect of each treatment was drawn from a standard normal distribution. Phase III trials were assumed to be randomised, with a continuous primary endpoint, 80% power and 5% significance level. Specific design scenarios were considered. The effect of the correlation between the phase II and III trial endpoints was explored analytically, by ranging the variance of the true treatment effect, while randomisation, number of stages and operating characteristics of phase II trials were explored using simulations. The number of phase II and III patients required to lead to the first successful phase III trial was used to measure efficiency of design parameters.

Results For the scenarios considered, the number of patients required to lead to the first successful phase III trial decreased from 3200 to 1000 patients, on average, as the correlation between endpoints increased from 0 to 1. Randomised single-stage phase II trials required 730 patients to lead to the first successful phase III, while Jung's randomised two-stage design required 554. A'hern's exact single-arm single-stage design required 463 phase II and III patients while Simon's single-arm two-stage design required 438. The type I error, α, and power, $1-\beta$, significantly affected the efficiency of phase II trials. Less stringent $\alpha = 0.1, 0.15$ and 0.2 combined with powers $1-\beta = 0.4, 0.45, 0.5, 0.55, 0.6$ yielded 417 phase II and III patients on average, while stringent $\alpha = 0.01$ or 0.05 combined with any choice of power required 555 phase II and III patients to lead to the first successful phase III trial.

Discussion Understanding the impact of differing design parameters on the efficiency of phase II trials better equips us with the tools needed to improve their design. Based on the scenarios considered Simon's single-arm two-stage design with a less stringent type I error and small type II error yielded the greatest efficiency in phase II trials.

Table of Contents

List of Figures

List of Tables

Abbreviations

AO	Annals of oncology
BJC	British Journal of Oncology
CRUK	Cancer Research UK
CSD	Clinically significant Difference
FDA	U.S. Food and Drug Administration
JCO	Journal of Oncology
MAMS	Multi-arm multi-stage designs
NPV	Negative Predictive Value
OS	Overall Survival
PFS	Progression-Free Survival
PPV	Positive Predictive Value
RECIST	Response Evaluation Criteria in Solid Tumours
SLD	Sum of the Longest Diameter
TCV	Testing Confidence Value
TTG	Time to Growth

Chapter 1 Introduction

Upon discovering a potential new cancer drug, it must undergo several stages of testing before it can be widely used among patients. First, the activity and toxicity of the novel drug is tested in preclinical trials involving both experiments in the laboratory and on animals (1). If the drug shows promise it is then rigorously tested on humans in four phases of clinical trials (1). In phase I trials information about how the drug interacts with the body (pharmacokinetics), and how the body reacts to the drug (pharmacodynamics), is gathered from a limited number of patients (2).

Dosing schemes, based on the data from the preclinical stage, are adjusted in order to establish patients' tolerance to the novel therapy, and to determine appropriate dosing levels. Therapies that are deemed safe are then tested in phase II trials (2). Here, the preliminary signal of efficacy of the drug is assessed with a small number of patients (2). A pivotal decision is made at this stage: if the new therapy shows sufficient evidence of activity, then a phase III trial can be initiated, otherwise the process may be terminated.

Phase III trials proceed after successful phase II trials and usually require several hundreds, or even thousands, of patients and can take many years to complete (1). If the drug shows definitive evidence regarding the superiority of the novel therapy over the standard treatment, it is licensed and marketed and then further tested for long-term safety and efficacy in phase IV post-marketing trials (1).

Although, theoretically, this process seems linear and each phase has confined requirements, it is actually very flexible which may result in overlapping phases or combining two phases (3). The complexity of the drug development process also means that it requires a substantial amount of investment, in terms of money, patients and time. The cost of researching and discovering a single new drug is estimated to be £1.15 billion and takes an average of 12.5 years (4). Such high costs for finding a single efficacious drug are attributable to the resources used in trials that investigate futile treatments. While the failure of trials (i.e., objectives of the trial have not been met, therefore the therapy under investigation in that trial is terminated) may add to the knowledge about the treatments under investigation (5), the main purpose of the drug development process is to benefit patients by allowing them access to innovative and effective treatments. The current drug development process indicates that this

is not the case and it is of Kaitin's opinion that the process is slow, inefficient, risky and expensive (6).

In oncology, the attrition rates are particularly high compared to other specialties: only 5% of therapies entering clinical development are licensed, while 20% of therapies for cardiovascular disease are approved (7). The demand for more effective therapies, due to the increasing number of people developing cancer and often a poor prognostic outlook, has resulted in a lower clinically relevant threshold during preclinical oncology trials, which may explain the reason for the low therapy approval rates reported (8). However, this is not the only reason contributing to these poor statistics; other contributing factors include the design of trials (8) and the quality of reporting of the trials (9), which is often the trial's only representation available.

The largest risk in the drug development process is associated with the transition from a relatively small-scale phase II trial to a large confirmatory phase III trial, where attrition rates are particularly high. This is due to the large resource investment that is required for phase III. The sheer volume of the resources required in phase III pressurises the researcher to make the right decision at phase II (often referred to as the go/no-go decision), and therefore, phase II trials need to be designed in a manner that can adequately fulfil this objective. However, it is reported that in 2011 the failure rates in phase III oncology trials were between 50-60% (10) and this has not improved even recently where oncology phase III success rates are reported to be at 45% (11). This long-standing attrition of oncology phase III trials suggests that the phase II trials are not fulfilling their purpose of rejecting futile treatments early and identifying truly efficacious treatments for further study.

Given the large number of therapies available for investigation (10) and the need to approve new therapies in order to benefit patients, as efficiently as possible, in addition to the high attrition rates in oncology, it is of vital importance that improvements to the efficiency of the drug development process are made. Since a key element of the drug development process is the phase II trial, this research will focus on investigating the effect of its design parameters on its ability to lead to successful phase III trials.

1.1 The role and importance of phase II trials

Traditionally, phase II cancer trials accrue a small number of patients into a single-arm design and compare the effect of the experimental therapy, with historical control data (12). The effect of the treatment is usually quantified by a short-term outcome, such as tumour response rate, particularly when the

treatment's mechanism of action is thought to have cytotoxic effects (12). Historically, when the therapies available to treat cancer were limited, phase II trials that were designed in this way screened out ineffective treatments as quickly as possible, while limiting the number of patients that were exposed to the ineffective agents (12).

As mentioned above, phase II trials occur prior to a confirmatory phase III trial, assessing whether the novel therapy shows sufficient clinical efficacy for further investigation. Unlike phase III trials, phase II trials in cancer are designed to take a relatively short period of time, hence a short-term endpoint is utilised to demonstrate whether the experimental treatment is promising. Phase II trials assess the holistic potential benefit of an experimental treatment, therefore their role includes proof-of-concept of clinical improvement of the experimental treatment over the standard care and gathering information on dose levels and schedules. Occasionally, these roles are investigated separately: phase IIa trials aim to provide proof-of-concept and/or determine the dose range and phase IIb trials are designed to confirm the short-term clinical efficacy of the treatment. Throughout this research, the term 'phase II trial' is used to refer to a trial that involves a go/no-go decision regarding proceeding to phase III, where appropriate dose levels have already been found in prior trials.

In the current era of drug development, the number of novel therapies in cancer is ever increasing (13). Mandrekar and Sargent (14) state that the number of novel therapies are no longer limited, rather there are currently many potentially efficacious drugs to investigate, with restricted resources for their development and evaluation. As critical and vital as phase III trials are in their ability to determine the efficacy of novel and experimental treatments, testing all available treatments in phase III trials is not only difficult, due to the lack of available patients (13), but it is also inefficient, expensive and could take an extensive amount of time. As such, the role of the phase II trial, in the current era, is profound: it can streamline the treatments that proceed to phase III, by rejecting futile treatments, without unnecessarily (and unethically (15)) subjecting a large group of patients to them in phase III. Consequently, the role of phase II trials is important and necessary as it may lead to more efficient exploration of the treatments available in terms of resources and benefit to patients.

The current era has also seen the emergence of a myriad of different types of cancer treatments (13). Chemotherapy has been the gold standard treatment for cancer for many years (16), however, new technologies have allowed the emergence of novel therapies such as those with cytostatic mechanisms of action, immunotherapies and targeted therapies (13). Testing combinations of

these treatments, rather than single agents are more likely to result in a greater effect since each treatment targets a different pathway (13). As a result, the role of phase II trials is not limited to evaluating whether a single drug is active enough to proceed to phase III, rather which combination of treatment is the most effective (13). Therefore, the role of phase II trials is not only important, but also how it is designed: it is vital that it is appropriate and of value to the current era.

The introduction of cytostatic agents has meant that the same definition of activity can no longer apply to all types of cancer treatments. Activity of cytotoxic treatments is usually measured with response rate (12) and this measure may not be able to adequately identify the activity of cytostatic therapies. These novel treatments have shown that they improve survival despite low tumour response rates in patients, which has led some researchers to question the relevance of phase II trials in their entirety (17). It is clear though, that rather than dismiss phase II trials as a whole, there is a need to assess the applicability of traditional designs of phase II trials to the ever-changing drug development process in oncology.

The importance of the role that phase II trials has on the drug development process is highlighted by the INTACT1 phase III trial, which aimed to compare the effect of adding gefitinib to gemcitabine and cisplatin versus gemcitabine and cisplatin alone in patients with advanced or metastatic non-small-cell lung cancer. Gefininib on its own showed promise in phase II trials, however using it in combination with gemcitabine and cisplatin was not tested in a phase II trial. INTACT1 was initiated as a result of promising results of a phase I trial where this combination therapy showed “favourable tolerability” (18). INTACT1 recruited more than 1000 patients and failed to confirm the results found in its preceding phase I trial (18). Therefore, phase II trials with appropriate designs are an important screening tool to limit the number of ineffective therapies that are investigated in resource-intensive phase III trials.

The importance of phase II trials is highlighted further by the fact that on occasion they alone may be used to lead to regulatory approval when phase III trials are not feasible. An example of this was a phase II study comparing different doses of imatinib in a two-arm study conducted in advanced gastrointestinal stromal tumour patients (19). The novel therapy showed very promising results: more than 80% of patients responded to the treatment or their disease did not progress, while also the treatment was tolerated well by patients. As a consequence of having such encouraging results, the resource-intensive phase III trial was not initiated and patients were able to access a safe and efficacious new treatment. The trial, however, did recruit 147 more patients

after an interim analysis showed such promising results. This addition of such a small number of patients, relative to what is usually required in phase III, confirmed initial findings and as a result the resources saved in this trial may have been used elsewhere.

Differences in the role of phase II trials may exist depending on whether they are designed and analysed in pharmaceutical companies or in academic settings. The main difference arises in the decision-making process, namely, whether the trial is successful, i.e., has found an active treatment. As a result, the criteria which is used to deem a trial successful in both settings are different: in academia the criterion of success is pre-specified and is based on a particular aim that the trial is designed to fulfil. In pharmaceutical companies, the criteria of success of a phase II trial are not taken in isolation and may be based on many external factors, such as the company's objectives and funding. As such, pharmaceutical companies may have the capacity and funds to run multiple trials for a single disease area, which allows the rapid development of potentially beneficial treatments in the current era, where many treatments are available for testing.

1.2 Design options for phase II trials

Historically, single-arm designs were frequently the chosen method for phase II trials in oncology (12). They were typically designed to test the hypothesis that the novel treatment is promising if the response rate is equal to or above a certain threshold, usually at 20%, and a lack of promise, usually at or below 5% (12). This design was used frequently as it was believed that a treatment that had a 20% response rate would result in clinically meaningful outcomes in subsequent phase III trials measuring long-term time-to-event outcomes, such as overall survival (OS), which is typically used (12).

Phase II trials have since changed and many researchers have conducted reviews of phase II trials that highlight which phase II designs are widely used, common features between them, and how they have evolved. One such review by Mariani and Marubini (20) surveyed phase II trials published in 1997, only. They identified 308 phase II trials, of which more than 95% used a single-arm design with objective response as the primary endpoint; separating these two parameters, 98.7% of the trials were single-arm designs and 96.8% used objective response as the primary endpoint. They also found that only 58 of the 308 articles reported a statistically identifiable design, of which the two-stage, hypothesis testing methods were the most common. They also reported some of the key features of the single-arm designs they identified: 78.6% of the single-arm trials used chemotherapy as the experimental treatment, with 33.9%

taking up to two years and 30.8% taking longer to complete (35.3% of trials didn't report the duration). They also reported the sample sizes of the single-arm trials they surveyed: 27.1% of trials required up to 20 patients and an equal amount required 46 or more patients. The remaining single-arm trials used 21 to 41 patients.

Langrand-Escure et al. (9) carried out a more recent review which included phase II trials published between 2010 to 2015, with the same aim as the review presented by Mariani and Marubini (20). Comparisons between the findings in both these reviews highlight the evolution of the use of designs of phase II trials. Of the 557 trials identified, 56.6% of them were single-arm designs and 80.7% used response rate as the primary endpoint. These results show that the traditional design is still the most common choice for phase II trials, however, there is a clear reduction in the number of trials that implement it, particularly in the use of the single-arm design, as more researchers opt to conduct the phase II trial using a randomised approach (34.6%). This may be due to the fact that single-arm phase II trials use historical control data for comparisons and many problems arise from using them. These include their lack of reliability and the fact that the quality of the data available may be poor (21). In addition, randomised trials become more popular as the standard of care improves so that there are more potential control treatments. Furthermore, endpoints measured in historical data may be different from the endpoint measured in the current trial, therefore rendering the data incomparable (21).

Given the significant shifts in drug development, it is clear that researchers have recognised that the designs of phase II trials need to adapt in order to fulfil their objectives in the current era. Recognising that many methods have become available, Brown et al. (22) reviewed 122 articles describing new or adapted phase II designs and from them the researchers produced a structured framework to aid researchers during the design process of phase II oncology trials. The main thought process to identifying an appropriate phase II design included therapeutic considerations, which include the mechanism of action of the treatment in question, the phase II trial aim, the outcome of interest and how it will be measured, whether the trial should include randomisation, the design category of the trial (including whether it is a single-stage or two-stage design etc.) and practical considerations which may include availability or reliability of previous data and whether the trial should be terminated early for, either a lack of activity or evidence of a very active treatment. These encompass clinical and statistical considerations, highlighting the need for multi-disciplinary collaboration in the adequate design of phase II trials, allowing them to be efficient in the current era.

The choice of endpoint in phase II trials is an important consideration and has become so due to the emergence of targeted therapies. As previously mentioned, the most commonly used endpoint in phase II trials is tumour response rate. Variations for tumour response rate exist including standard criteria such as Response Evaluation Criteria in Solid Tumours (RECIST), which categorise tumour measurements of the lesion prior to and after receiving the treatment (23). However, response rate has come under question as it fails to capture the effect of cytostatic treatments. This is evidenced by the novel therapy sorafenib, which is used for patients with advanced renal cell carcinoma and hepatocellular carcinoma, which showed low response rates in phase II studies (24, 25), but subsequent phase III trials showed that it is clinically beneficial as it prolongs progression-free survival and overall survival (26, 27). Another drawback with response rate – a short-term outcome – is that it does not always reflect long-term efficacy in phase III trials (20), where the outcome of interest may be survival or quality of life of patients. Historically, this endpoint was chosen as it is a short-term outcome, only used to provide an indication of efficacy, rather than confirm it. Despite response rate's obvious drawbacks, researchers lean towards it due to the fact that it is standardised, easily applicable and yields early outcomes (28).

Alternatives to response rate have emerged, in order to accommodate cytostatic treatments, such as RECIST (version 1.1) (29), which incorporate stable disease as a positive outcome to the treatment. In addition, some researchers forego the categorisation of the change in tumour size before and after the treatment is administered, and use it as a continuous endpoint, which avoids the loss of data that may occur when the outcome is categorised. However, this endpoint is not commonly used and can be statistically intensive (28).

Due to the problems that exist with response rate as an appropriate endpoint, many researchers recommend the use of time-to-event outcomes in phase II trials (23, 30-32). The advantages of using overall survival (OS) in phase II trials is that it is an objective measure and is clinically meaningful, however, it can be argued that it is simply not appropriate to fulfil the purpose of phase II trials, as it requires a long follow-up period, and is ultimately the endpoint typically used in phase III trials (33). Progression-free survival (PFS) is more suited to phase II trials due to the fact that it typically requires a shorter follow-up period (in comparison to OS) and that it is more reflective of a longer-term clinical benefit (in comparison with response rate). However, PFS can be prone to biases particularly in unblinded trials and frequent tumour assessments are required during the follow-up period (34). Other endpoints are available such as

measures for quality of life, functional imaging and the use of biomarkers. However, these also have drawbacks including subjectivity, complex analyses, being time consuming and not being valid surrogates of efficacy (34).

The appropriateness of the design of phase II trials has also been questioned due to the change in the current era. Novel treatments often require the use of different endpoints, such as PFS, and the lack of historical data for these newer endpoints means that randomised phase II trials are required more often. However, this shift in design is not without reservation; the main concern against randomised designs is that it requires too many patients (35), particularly for the purpose of phase II trials which is to screen treatments for initial activity (as mentioned above). In addition, it is not as easy to implement and statistical analysis can be intense (35). Despite this, randomised phase II designs' main attraction is the fact that a valid and reliable control is present, so that robust conclusions can be made about the treatment effects. For this reason, some researchers argue that randomised designs can be the answer to increasing the efficiency of the drug development process in oncology (31).

In addition to randomisation, phase II trials can also be designed with a single-stage or two-stage or even multi-stage designs. Jung (36) discusses statistical issues with design and analysis of single arm two-stage designs. He states that these are the most common phase II trial designs, despite the vast availability of new designs (36). In two-stage designs, trials are conducted with an interim analysis separating the two stages. Critical values, or cut off boundaries, are obtained by pre-specifying the operating characteristics of the trial. The critical values are then used at each stage of the trial and are compared with the observed responses; the go/no-go decision made in these trials is based on the critical values. Jung states that the planned sample sizes at each stage are oftentimes not attained in the trial, due to practical challenges, such as lack of available patients, consequently the critical values become meaningless and cannot be used to make the pivotal decision at phase II (36).

In phase III trials, this problem rarely arises due to the fact that rigid protocols, with well-established methodologies, are created and adhered to, which include the statistical analysis of the trial which is conditional on the sample size. Phase III trial statistical designs and methodologies are generally agreed upon by researchers (37). A phase III trial usually compares an experimental treatment with standard care in a two-arm randomised design (37), although there are other design options for phase III trials, such as multi-arm multi-stage (MAMS) designs (38). Under the frequentist framework, the sample size of the phase III trial is obtained so that the trial is powered to detect a clinically significant difference at a pre-specified type I error rate. The type I error rate is also known

as the significance level and it refers to the probability of obtaining a false-positive result, i.e., recommending an ineffective treatment. In the event that this occurs, harm may befall the population, hence this error rate is restricted to a low level. A second error can also occur under the frequentist method, known as the type II error and this is the probability of failing to recommend a positive treatment i.e., false-negative. The power is the complement of the type II error. Conventionally, phase III trials have a pre-specified significance level of 0.05 and power set to 0.8 or 0.9 (39).

While phase III statistical methods are well-established, no consensus for phase II trial operating characteristics has been reached, due to differences in phase II trial purposes (proof-of-concept or dose-finding) and the fact that it is not a definitive, registrational trial, for which minimum standards exist. In Langrand-Escure et al.'s (9) review they reported that 37.7% of the 557 phase II trials they identified used 0.05 for the type I error value, while 5% used a type I error less than 0.05 and 27.7% used a value larger than 0.05. They also found that of the articles they reviewed 80% and 90% power were most commonly used, with 80% used more often (28.2% compared with 25.3%). Only 2.7% of articles used a power smaller than 80%, 9% used a power between 80% and 90% and 5.9% used a power larger than 90%. However, the striking thing revealed in Langrand-Escure et al.'s (9) review is that 28.9% of the articles did not report such an important statistical requirement in clinical trial designs.

The majority of clinical trial designs adopt the frequentist approach, where the probability of a particular event occurring is calculated, given the operating characteristics. The advantage of this approach is not only that it is well-known and familiar, but also that it is a rigorous and thorough process. However, some of its disadvantages include the fact that it is limiting: once a trial has been initiated, amendments to the design have to be rigorously thought out and thus may be difficult to implement (28).

An alternative framework to clinical trials is the Bayesian approach. This approach to designing trials involves prior information about the outcome measure to generate a prior distribution, which quantifies the uncertainty about the measure. The clinical trial is then conducted and the data is combined with the prior distribution to obtain a posterior distribution. Inferences can then be made based on the updated posterior distribution, including posterior probabilities, prediction intervals and credible intervals. One of the virtues of this approach is that it is flexible and data driven, so is suited to adaptive assignment of patients to therapies that prove to be more efficacious during the trial (28). Its usage can be particularly useful for rare diseases as prior information can help supplement the lack of data available in the study itself

(40). Despite the increase in Bayesian methodologies proposed there is a low-uptake of them in clinical trials (41). As such, throughout this research phase II trial design parameters that are investigated are based on the frequentist approach, as it is still the most widely used framework in clinical trials, particularly for common cancers (40)

The design options, highlighted in this section, by reviewing previous choices of phase II trial parameters, has emphasised that an updated review is needed. This is presented in Chapter 2 of this research and aims to answer the following question: how are phase II trials in oncology *currently* conducted? A detailed overview of this chapter is provided in Section 1.4.

1.3 The importance of adequately designed phase II trials

It is clear from Section 1.2 that a consensus on the optimal design of phase II trials has not been reached and it is still a subject of ongoing deliberation (35). The consequence of this is that phase II trials in oncology are conducted without strong evidence that they will lead to a positive outcome in phase III trials. Kola and Landis (42) state that the majority (60%) of novel agents, that showed promise in phase II trials, failed to translate this success in proceeding phase III trials. Maitland et al. (43) also revealed that phase II trials that evaluated combination therapies between 2001 to 2002 were successful 72% of the time, however, the chances that the proceeding phase III trial would result in a significant outcome was only 3.8%. Furthermore, Zia et al. (44) showed that the response rates in phase III trials were frequently found to be lower than their preceding phase II trial. There is a clear disconnect between the decision made at the end of phase II, and despite researcher's efforts to bridge the gap between the results of the two phases, there remains a need for guidance around the efficiency of phase II designs.

The go/no-go decision that occurs at the end of phase II trials hinges on an appropriately designed study. In order to investigate truly efficacious treatments in phase III trials, the preceding phase II trial must reject futile treatments, and accurately identify treatments that are likely to benefit patients. The design considerations of phase II trials include (but are not limited to) the aims of the trial, statistical design, randomisation and endpoints. Each of these require rigorous thought and need to be appropriately implemented in order to ensure the phase II trials allow informed decisions to be made, prior to embarking on a resource-intensive phase III trial. Doing so will ensure that patients have access to the best treatments available, therefore improving their chances of survival, while also ensuring that costs (monetary and ethical) are kept to a minimum.

1.4 Aims and structure of thesis

It is clear that phase II trials play a pivotal role in the drug development process in oncology. However, they are not fulfilling their purpose of being an adequate filter for phase III trials: recommending truly efficacious treatments and rejecting futile treatments as efficiently as possible. It is also clear that there are many designs available for phase II trials and researchers need to contemplate a number of clinical and statistical considerations for phase II trials in order to choose the most appropriate design. The aim of the research, presented in this thesis, is to aid researchers to make informed decisions about the design of phase II trials, by investigating and revealing the effects of oncology phase II design parameters on the efficiency of the drug development process as a whole. While each phase II trial is different and choosing a specific design parameter will depend on external factors, recommendations will be made to help guide these decisions, based on the findings of this research.

As previously mentioned, the design parameters of the phase II trials will be limited to those associated with the frequentist statistical framework, due to its popularity among researchers. With many design parameters available for phase II trials, the most commonly used ones were identified in a systematic review, which is presented in Chapter 2. The systematic review also aims to reveal how phase II trials are designed, and how the current designs and methods differ from previous years. The quality of reporting of phase II trials is also discussed. The phase II design parameters identified in the systematic review is the focus of future chapters in this thesis, where the aim is to reveal their effect on successfully screening new treatments.

With the design parameters selected from the systematic review, a literature review was conducted in Chapter 3 with the purpose of identifying a measure to use to evaluate the efficiency of phase II trials. The aim here was to find a measure that can be used to quantify the effect of a design parameter on the overall performance of the drug development process. Several criteria were applied in order to identify one measure which would be used to assess efficiency of phase II trials.

In Chapter 4, the overall methodology used to evaluate the effect of phase II design parameters on the drug development process is explained; here findings from the literature review are used to inform the assumptions made, to facilitate the investigations. While the main aim of the research is to evaluate the impact of design parameters of oncology phase II trials on their ability to successfully screen new treatments, specific research questions can be derived once the design parameters to study have been selected (Chapter 2) and what measure

to use to quantify their effect on oncology phase II trial efficiency has been identified (Chapter 3). Therefore, Chapter 4 will outline the specific research questions that will be reported in the remainder of the thesis. Hence, Chapters 5, 6 and 7 are dedicated to investigating the effect of the chosen three different design parameters on the performance of the drug development process. In addition, the methods employed to investigate the effect of the specific design parameters in each of the chapters (5, 6 and 7) are also explained and the results for each parameter are presented. Each chapter (5, 6 and 7) also included a discussion where the specific results to the corresponding chapter are interpreted and compared to previous literature. Key limitations and potential implications specific to each parameter are also presented in the corresponding chapter.

Finally, Chapter 8 summarises the findings of the research, while providing a critical evaluation of the assumptions and methods. In addition, the applicability of these methods and findings are also addressed, so that researchers are clear about the implications of the findings, presented in this research project, on future designs of phase II trials and, thus the drug development process as a whole.

Chapter 2 A systematic review exploring the design characteristics of phase II trials published in three leading oncology journals in 2015 and 2019

In this chapter, a systematic review is carried out with the aim of highlighting the design parameters currently used in oncology phase II trials. The findings from this chapter will help the selection of the design parameters that will be considered in this research.

2.1 Introduction

Phase II trials are typically designed to assess whether a novel therapy shows sufficient clinical efficacy for further investigation (45). A pivotal decision is made at this stage regarding the appropriateness and feasibility of proceeding to a resource-intensive phase III trial. The phase II-III decision point poses the highest financial risk due to the large resource investment required for phase III trials, as a result of increased patient numbers required and associated increases in time, and therefore cost. Subsequently, if the decision criteria, used in phase II trials, to ascertain whether to proceed to a phase III trial, is set too high or too low, the consequences can be wasteful: too high a threshold leads to potentially efficacious treatments being terminated, while too low a threshold means that too many inefficacious treatments proceed to phase III testing. Therefore, phase II trials are required to be designed in a manner that can most effectively inform this decision-making process.

In cancer drug development attrition rates are very high (46). Compared to other specialties, only 5% of therapies entering clinical development are ultimately licensed, compared to 20% of therapies entering clinical development for cardiovascular disease (7). Harrison (47) conducted a review of a total of 174 articles comprising of a mixture of phase II and III trials published during 2013 to 2015. They found that oncology trials failed 32% of the time, compared to only 7% in cardiovascular trials. More specifically, Kola and Landis (42) highlighted the failure rates of oncology phase III trials to be between 50-60%, which is significantly larger than other disease areas, such as cardiovascular disease where the failure rate is between 20-30%. These data suggest that oncology phase II trials require improved designs to better fulfil their purpose of rejecting futile treatments early and selecting truly efficacious treatments for further study. There is a significant demand for more effective therapies due to frequent poor outcomes for patients diagnosed with cancer, the increasing

number of people developing cancer and increasing understanding of cancer biology, leading to increased targets and the ability to define more discrete sub-populations within or across disease sites. This has contributed to a lower clinically relevant threshold during preclinical oncology trials, which may explain the reason for the low therapy approval rates reported (8). However, this is not the only reason contributing to these poor statistics; a key contributing factor is the design of these trials (8).

Understanding the key design characteristics of phase II trials is necessary to be able to explore their impact on the efficiency of the phase II to phase III pathway. Characteristics such as sample size, operating characteristics, endpoints, trial design and decision criteria all impact the ability to successfully screen treatments for appropriateness of continuing clinical development.

Previous literature has addressed the issue of the design, and quality of reporting, of phase II trials. In 2000, Mariani and Marubini (20) published a review with the aim of investigating which designs are frequently used in phase II trials. Their review was limited to phase II trials published during one year, specifically, 1997. They found that the statistical components of these trials were poorly reported, with only 61 out of 308 studies having an identifiable design. Grellety et al. (48) also conducted a review, focusing on highly ranked oncology journals, and found that even in journals with strict editorial policies, vital information about the designs was often omitted. Langrand-Escure et al. (9) reviewed the quality of reporting of phase II trials published in three high impact journals during 2010-2015, reporting also on the study characteristics of the trials. They concluded that the quality of reporting was poor and raised as topics for further investigation the evaluation of the design and methodological choices within phase II trials. This highlights the need to better understand the design characteristics of phase II oncology trials in order to evaluate the impact of these choices on the efficiency of decision making at phase II.

Recognising there are many phase II trial designs available, Brown et al. (22) reviewed methodological articles describing phase II designs and produced a structured framework to aid researchers during the design process of phase II oncology trials. Core components for design choice included type of endpoint and use of randomisation. In addition, in 2005 Lee and Feng (49) reviewed the use of randomised phase II trials, in comparison to the conventional single-arm design, and found that their use was increasing in cancer research. Given the shifts in drug development, where many novel treatments are available for testing and the emergence of novel types of therapies, researchers have recognised that the designs of phase II trials need to adapt to fulfil their objectives in the current era.

This chapter describes a systematic review to determine design characteristics of high-quality phase II trials in the current cancer clinical trial environment. Understanding how phase II trials are designed, and the key design parameters used, will enable future research to evaluate the impact of current design choices on the efficiency of phase II decision-making. This will ultimately allow researchers to make more informed choices when designing phase II oncology trials. Another objective of this systematic review is to explore how the designs of phase II trials have changed in recent years. The results of the review will feed into the methods for the research reported in the remainder of the thesis.

2.2 Methods

A systematic review was undertaken based on Mariani and Marubini's original systematic review published in 2000 (20), however a pragmatic streamlined approach was taken to capture a snapshot of the design characteristics of phase II trials published in high impact journals, reflecting the methods of Langrand-Escure et al. (9). High impact journals reviewed by Langrand-Escure et al. (9) were selected: Journal of Clinical Oncology, Annals of Oncology and British Journal of Cancer. This approach represents a focused review of high-quality trials that may be seen to reflect common practice in the design and conduct of oncology phase II trials. The aim of the review is to provide an overview of design characteristics, rather than to undertake an exhaustive review of the literature. Articles published in two years at the beginning and towards the end of the PhD (2015 and 2019) were therefore chosen, and the phase II oncology trial designs used, were compared.

2.2.1 Study selection

The studies included in the systematic review were identified through a literature search using the Ovid MEDLINE database and was last consulted 1st August 2021. Searches were performed for the years 2015 and 2019; the search strategies are presented in Table 2.1. The search was limited to articles that mentioned phase II trials, in different forms, in either the keywords, abstract or the title and that were focused on trials of antineoplastic agents. The search also included phase II/III trials, in order to include all trials with a go/no-go decision to be made.

Table 2.1 Left panel: search strategy used to identify published phase II cancer trials from three high impact journals. Right panel: key data extracted from articles

1. exp antineoplastic agents/ 2. Phase ii.tw.	1. Journal name

3. Phase 2.tw. 4. “phase ii/iii”.tw. 5. “phase 2/3”.tw. 6. 2 or 3 or 4 or 5 7. 1 and 6 8. limit 7 to English language 9. limit 8 to yr=”2015” or "2019" 10. "British journal of cancer".jn 11. "annals of oncology".jn. 12. "journal of clinical oncology".jn. 13. 10 or 11 or 12 14. 9 and 13	2. Statistical objectives (hypothesis testing/estimation/selection) 3. Disease site 4. Experimental treatment type 5. Recruitment duration 6. Follow-up duration 7. Number of primary endpoints 8. Type of primary endpoints 9. Randomisation incorporated 10. Randomisation ratio 11. Number of arms in the trial 12. Single-stage or two-stage design 13. Statistical design referenced 14. Bayesian or frequentist methods 15. Statistical test used 16. Target sample size 17. Type I error and sidedness 18. Type II error 19. Study result

The abstracts of the identified articles were reviewed for inclusion, with a selection of five of the papers that were included for full-text evaluation additionally reviewed independently by a second reviewer, for inclusion and data extraction. Articles describing pre-clinical, phase I, phase I/II or dose-finding trials were excluded. Phase II trials that were not evaluating an experimental treatment/combination for any cancer type were also excluded. These included trials evaluating prognostic factors, trials with the sole purpose of assessing the role of a biomarker, or trials evaluating the efficiency of diagnostic methods. In addition, updated results from previously published phase II trials were also excluded.

2.2.2 Data extraction and analysis

The following information was extracted from the full-text articles: journal name; statistical objectives of the study, defined as whether the trial tested a hypothesis, estimated a treatment effect or selected a treatment (statistical considerations reported were used to determine this); disease site; experimental treatment type; recruitment duration (months); follow-up duration (months); number of primary endpoints evaluated; type of primary endpoint; whether the design was randomised to either a control or experimental arm and number of arms in the trial; randomisation ratio used; whether the trial used a single-stage or two/multi-stage design; whether a specific statistical design was

referenced, and if so what design was used; whether the trial used Bayesian or frequentist methods; whether it was a comparative study; the test(s) used to compare the treatment; target sample size; type I and II error rates; whether biomarkers were assessed as a secondary objective; and the study result (Table 2.1).

The type of experimental treatment was classified as cytotoxic chemotherapy, immunotherapy, targeted therapy, hormonal therapy, radiation therapy, combination or other. When patients were given more than one treatment, only the experimental element was classified. In addition, if the experimental treatment was a combination of the same type of treatment, it was recorded as that type of treatment, rather than a combination; if the combination was of different treatment types, then a combination was recorded. Each treatment classification was extracted from three drug indexes (50-52) and was double-reviewed independently by myself and the clinical supervisor of this research (FC).

Trials were deemed to have specified a design if a single-arm or a non-comparative randomised trial referenced a specific design, or discussed the statistical test used to compare findings with historical data. In the case of comparative randomised designs, a design was considered specified if the test used to compare treatment arms was stated. In addition, the type I error requires two elements to be specified: the value at which it is set and whether it is one-tailed or two-tailed. If the article did not specify the number of tails, then it was recorded that they did not report it. As such, both elements were extracted from the articles, however, for consistency, the corresponding two-sided value is reported here.

Endpoints were categorised as response rate, safety, and time-to-event endpoints, such as progression-free survival, based on the articles reviewed. Time-to-event outcomes were often dichotomised to become a binary outcome depending on whether the event had occurred by a specified time; thus, they were extracted as dichotomised time-to-event outcomes.

Data were extracted and stored in Microsoft Excel. Rstudio 1.4 (53) was used to summarise the data using frequencies (percentage), means (standard deviation) and medians (interquartile range), where appropriate. Data were summarised overall and by year.

2.3 Results

For the year 2015, 97 articles were found, while 74 articles were identified for the year 2019. In 2015, 74 articles met the inclusion criteria, compared to 54

articles in 2019. Of the 74 articles selected from 2015, the majority were identified from Annals of Oncology (44.6%) and Journal of Clinical Oncology (44.6%). In 2019 the majority of papers identified were published in Journal of Clinical Oncology (57.4%). Figure 2.1 displays a detailed summary of the selection process and reasons for the exclusions.

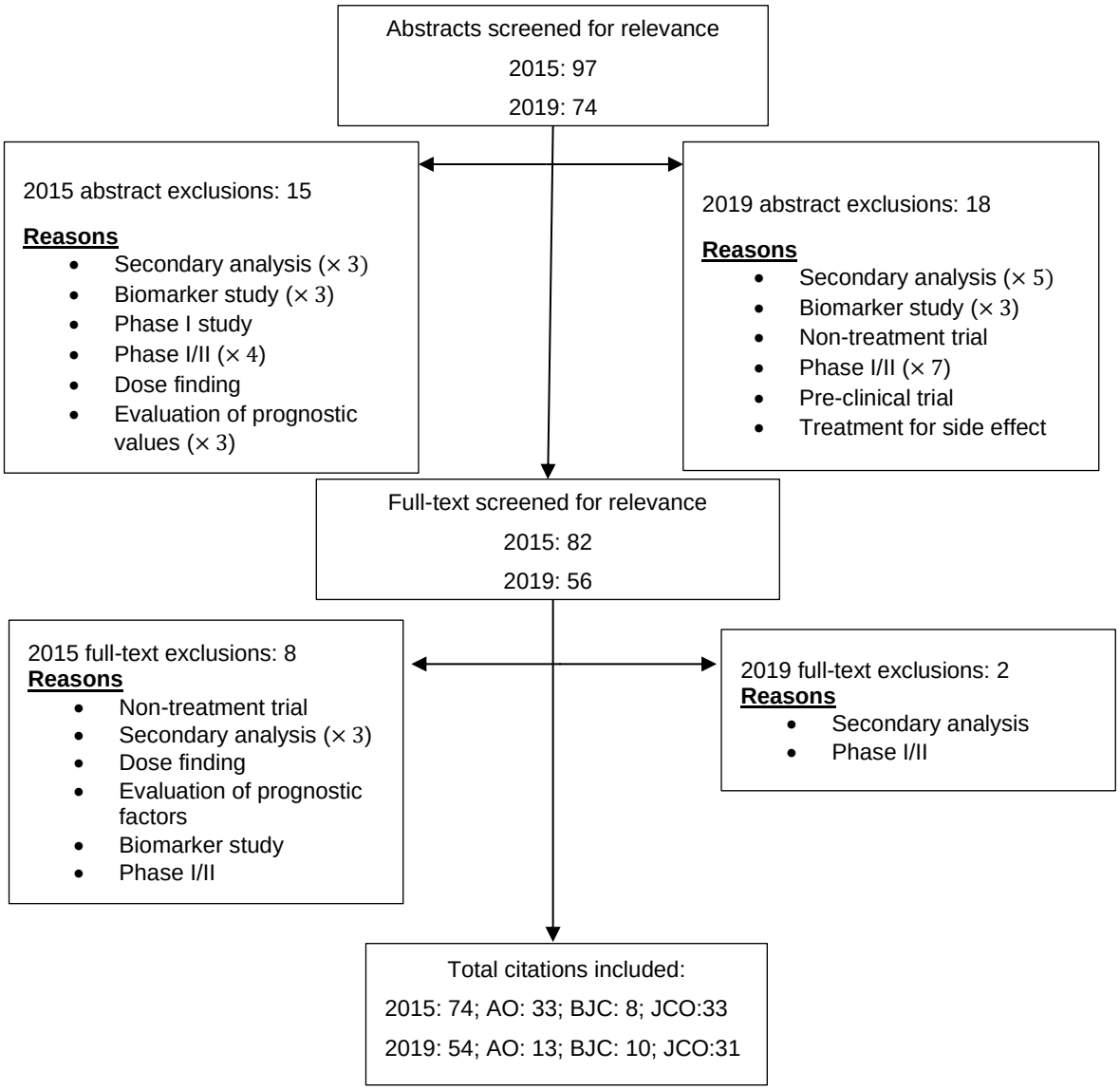


Figure 2.1 Flowchart summarising the selection of articles included in the review (AO: Annals of Oncology, BJC: British Journal of Cancer, JCO: Journal of Clinical Oncology)

Table 2.2 shows the characteristics of the phase II trials identified. The main statistical objective, which was derived from the statistical considerations and sample size calculation details, where otherwise not directly specified, was predominantly focused on hypothesis testing (overall 83.6%; 2015: 86.5%; 2019: 79.6%). Of note, in 2019 three (3.7%) papers used phase II trials to evaluate more than one main aim, namely estimation and hypothesis testing. They aimed to estimate a treatment effect, and thus used estimation methods to

derive their sample size, but also included cut-off decision criteria to proceed to phase III, in the form of hypothesis testing (54-56).

Table 2.2 Characteristics of the identified phase II trials published in 2015 and 2019 meeting inclusion criteria (* frequency (%) presented unless otherwise stated; ** median (IQR) presented)

Characteristics		Overall	Year	
		n (%)	2015 n(%)*	2019 n (%)*
Journal	Annals of Oncology	46 (35.9)	33 (44.6)	13 (24.1)
	British Journal of Cancer	18 (14.1)	8 (10.8)	10 (18.5)
	Journal of Clinical Oncology	64 (50.0)	33 (44.6)	31 (57.4)
Follow-up duration (months)	Mean(SD)	24(12-36) **	24(12-27) **	31.2(19.7)
	Not reported	78 (60.9)	46(62.2)	32(59.3)
Study aims	Treatment Selection	4 (3.1)	2(2.7)	2(3.7)
	Estimation	12 (9.4)	8(10.8)	4(7.4)
	Hypothesis testing	107 (83.6)	64(86.5)	43(79.6)
	More than one aim	3 (2.3)	0(0)	3(3.7)
	Not specified	2 (1.6)	0(0)	2(5.6)
Disease site	Brain	2 (1.6)	1(1.4)	2(3.7)
	Breast	18 (14.1)	9(12.2)	9(16.7)
	Gastrointestinal tract	24 (18.8)	16(21.6)	8(14.8)
	Gynaecology	6 (4.7)	2(2.7)	4(7.4)
	Head & neck	10 (7.8)	4(5.4)	6(11.1)
	Haematology	16 (12.5)	10(13.5)	6(11.1)
	Lung	22 (17.2)	14(18.9)	8(14.8)
	Many sites	3 (2.3)	1(1.4)	2(3.7)
	Sarcoma	5 (3.9)	3(4.1)	2(3.7)

	Skin	5 (3.9)	3(4.1)	2(3.7)
	Urology	13 (10.2)	9(12.1)	4(7.4)
	Other	3 (2.3)	2(2.7)	1(1.9)
Type of experimental treatment	Cytotoxic chemotherapy	33 (25.8)	23(31.1)	10(18.5)
	Hormone therapy	3 (2.3)	1(1.4)	2(3.7)
	Immunotherapy	11 (8.6)	1(1.4)	10(18.5)
	Targeted therapy	67 (52.3)	43(58.1)	24(44.4)
	Radiation therapy	3 (2.3)	0(0)	3(5.6)
	Combination	8 (6.3)	4(5.4)	4(7.4)
	Other	3 (2.3)	2(2.7)	1(1.9)
Duration of recruitment (months)	Mean(SD)	31 (20.8-41)**	27(18-28) **	37.3(19. 8)
	Not reported	12 (9.4)	7(9.5)	5(9.3)
Single-stage or multi-stage	Single-stage	91 (71.1)	54(73.0)	37(68.5)
	Multi-stage	37 (28.9)	20(27.1)	17(31.5)
Single-arm or randomised	Single-arm	58 (45.3)	33(44.6)	25(46.3)
	Randomised	67 (52.3)	41(55.4)	26(48.1)
	Multiple-arm non-randomised	3 (2.3)	0(0)	3(5.6)
Number of arms	1-arm	58 (45.3)	33(44.6)	25(46.3)
	2-arm	57 (44.5)	32(43.2)	25(46.3)
	3-arm	10 (7.8)	7(9.6)	3(5.6)
	4-arm	3 (2.3)	2(2.7)	1(1.8)
Multiple-arm trials	Comparative	59 (84.2)	36(87.8)	23(79.3)
	Non-comparative	11 (15.7)	5(12.2)	6(20.7)
Randomisation ratio	Equal	58 (82.9)	34(82.9)	22(84.6)
	Unequal	12 (17.1)	7(17.1)	4(15.3)
	Yes	115 (89.8)	65(87.8)	50(92.6)

Design specified	No	13 (10.2)	9(12.2)	4(7.4)
Number of primary endpoints	Unique	123 (96.1)	71(95.9)	52(96.3)
	Multiple	5 (3.9)	3(4.1)	2(3.7)
Type of Primary endpoint	Response rate	65 (50.8)	35(45.5)	30(52.6)
	Safety	5 (3.9)	1(1.3)	4(7.0)
	Time-to-event	30 (23.4)	22(28.6)	8(14.0)
	Dichotomised time-to-event	32 (25.0)	18(23.4)	14(24.6)
	Other	2 (1.6)	1(1.3)	1(1.8)
Target sample size met	Yes	79 (61.7)	47(63.5)	32(59.3)
	No	35 (27.3)	19(25.7)	16(29.6)
	Not assessable	14 (10.9)	8(10.8)	6(11.1)
Total target sample size	Median(IQR)	81(50-132.5)	80(51-150)	85 (50-120)
	10-50	30 (23.4)	17(23.0)	13(24.1)
	51-100	40 (31.3)	22(29.7)	18(33.3)
	101-150	25 (19.5)	13(17.6)	12(22.2)
	151+	19 (14.8)	14(18.9)	5(9.3)
	Not reported	14 (10.9)	8(10.8)	6(11.1)
Type I error value	<0.025	1 (0.8)	0(0)	1(1.9)
	0.025	19 (14.8)	7(9.5)	12(22.2)
	0.025< α <0.05	4 (3.1)	2(2.7)	2(3.7)
	0.05	31 (24.2)	22(29.7)	9(16.7)
	0.05< α <0.1	7 (5.5)	5(6.8)	2(3.7)
	0.1	32 (25.0)	14(18.9)	18(33.3)
	>0.1	3 (2.3)	3(4.1)	0(0)
	Not reported	16 (12.5)	11(14.9)	5(9.3)
	Not applicable	16 (12.5)	10(13.5)	6(11.1)
One or two tailed α	One-sided	35 (27.3)	23(31.1)	12(38.9)
	Two-sided	20 (15.6)	9(12.2)	11(20.4)
	Not reported	48 (37.5)	32(43.2)	16(29.6)

	Not applicable	16 (12.5)	10(13.5)	6(11.1)
Type II error value	<0.1	9 (7.0)	6(8.2)	3(5.6)
	0.1	35 (27.3)	21(28.4)	14(24.6)
	0.1<β<0.2	13 (10.2)	7(9.5)	6(11.1)
	0.2	34 (26.6)	16(21.6)	18(33.3)
	>0.2	2 (1.6)	1(1.4)	1(1.9)
	Not reported	20 (15.6)	13(17.6)	7(13.0)
	Not applicable	16 (12.5)	10(13.5)	6(11.1)
Biomarker analysis	Yes	35 (27.3)	19 (25.7)	16 (29.6)
	No	93 (72.7)	55 (74.3)	38 (70.4)
Statistical method	Frequentist	127 (99.2)	73(98.6)	54(100)
	Bayesian	1 (0.8)	1 (1.4)	0(0)
Study result	Positive	90 (70.3)	48(64.9)	42(77.8)
	Negative	36 (28.1)	24(32.4)	12(22.2)
	Inconclusive	2(1.6)	2(2.7)	0(0)

Overall, the most common types of experimental therapies evaluated were targeted therapy (52.3%) and cytotoxic chemotherapy (25.8%). The proportion of trials testing cytotoxic chemotherapies reduced from 31.1% in 2015 to 18.5% in 2019, reflected by the increase in the proportion of trials evaluating immunotherapies in 2019 (18.5% compared to 1.4% in 2015).

In both years, phase II trials utilised multiple-arm designs more frequently than single-arm designs. Overall, single-arm trials were used 45.3% of the time, while multiple arm designs (including randomised and non-randomised) were used 54.6% of trials. In 2015, all multiple-arm trials were randomised, however, in 2019, 26/29 (89.7%) were randomised and 3/29 (10.3%) were non-randomised. Multiple-arm non-randomised trials included multiple cohorts of patients stratified into different arms. Of the trials that were multi-arm, the majority were comparative in nature, however, there is a slight reduction in their use from 2015 (87.8%) to 2019 (79.3%). Multiple-arm non-comparative trials included trials with more than one arm, each assessed independently. The majority of randomised trials used equal randomisation ratios (84.6%). Master protocol designs were also used in four phase II trials investigated: one basket trial was identified in each year (57, 58) and there were two trials in 2019 that used a similar structure to a basket trial (59, 60), where patients were recruited

into different cohorts to answer different questions, however, they did not specifically refer to the design as a basket trial.

Of the 58 single-arm trials identified (2015: n=33, 2019: n=25), 31 referenced a statistical design (2015: n=17, 2019: n=14), and of the 67 randomised trials (2015: n=41, 2019: n=26), 7 referenced a design (2015: n=4, 2019: n=3). Table 2.3 shows the distribution of the referenced designs across the two years. Of the seven randomised trials that referenced a design two were comparative (61, 62), while the rest were non-comparative. Of the 38 articles that explicitly mentioned a design, Simon's design (63) was the most used across both years (n=23; 60.5%). Fleming's two-stage (64) and A'hern single-stage designs (65) were also frequently referenced. One trial employed a Bayesian design in 2015.

Table 2.3 statistical designs referenced in the trials with single arm or randomised designs

Referenced design	Overall (n (%))	2015 (n (%))	2019 (n (%))
Randomised (n)	7	4	3
Simon's minimax design	1 (14.3)	0	1 (33.3)
Simon's optimal design	1 (14.3)	1 (25.0)	0
Simon's design, not further specified	1 (14.3)	0	1 (33.3)
Bryant and Day design	1 (14.3)	0	1 (33.3)
A'hern/Fleming design	2 (28.6)	2 (50.0)	0
Fleming's two-stage design	1 (14.3)	1 (25.0)	0
Single-arm (n)	31	17	14
Simon's optimal design	9 (29.0)	4 (23.5)	5 (35.7)
Simon's minimax design	3 (9.7)	2 (11.8)	1 (7.1)
Simon's design, modified	1 (3.2)	1 (5.9)	0
Simon's design, not further specified	7 (22.6)	3 (17.6)	4 (28.6)
A'hern/Fleming design	7 (22.6)	6 (35.3)	1 (7.1)
Fleming's two-stage design	2 (6.4)	0	2 (14.2)
Group sequential design	1 (3.2)	0	1 (7.1)
Bayesian multi-stage design	1 (3.2)	1 (5.9)	0

Of all 128 articles identified 96.1% of them used one primary endpoint, while the remaining trials used two primary outcomes. The majority of trials in both years used response rate as their primary endpoint. However, in 2019 response rate

was used slightly more than in 2015 (45.5% vs 52.6%). In 2015, the second most frequent type of primary endpoint was time-to-event (28.6%); however, the usage of time-to-event outcomes was approximately half this (14%) in 2019. It is clear that the most common type of endpoint for both years is a binary outcome (response rate: 50.8%; safety: 3.9%; dichotomised time-to-event; 25.0%).

Overall, 61.7% trials met their target sample size, similar across both years. Median sample size was 81 (IQR 50-132.5), again similar across both years. In 2015, phase II trials used a wider range of target sample sizes than in 2019, with almost 20% of trials in 2015 reporting sample sizes of over 150. Strikingly, approximately 10% of the total identified trials did not report the target sample size, and this remained true across both years.

Given the number of trials that did not report the target sample size, it is not surprising that some trials also did not report the type I and II error rates, despite their importance in trial design and interpreting trial results; 12.5% of all trials identified did not report the type I error rate (α), while 15.6% of all trials did not report the type II error rate. In addition, 37.5% of all trials did not report whether a one-sided or two-sided significance level was used, however this decreased from 43.2% in 2015 to 29.6% in 2019. Furthermore, two trials (66, 67) were not statistically powered, rather the sample size was chosen based on practical considerations. In both 2015 and 2019, one-sided significance levels were the most commonly used (31.1% and 38.9%, respectively). Two-sided significance levels were only used in 12.2% of the trials in 2015 and increased to 20.4% in 2019. Upon converting type I error rates to two-sided values, the most common type I error used was 0.05 (29.7%) in 2015, while in 2019, the most common converted two-sided type I error increased to 0.1 (33.3%), indicating that more stringent type I error rates were used in 2015. A small type II error of 0.1 was also used most often in 2015 (28.4%), while in 2019 the most common type II error rate was 0.2 (33.3%).

The use of biomarker analyses, as a secondary objective, were limited in the phase II trials identified, with only 25.7% of trials published in 2015 including biomarker investigations; this did not increase by much in 2019 (29.6%). The use of Bayesian methods in the design of phase II trials was also extremely limited with only one article (68), out of the 128 articles identified, using a Bayesian design and another (69) designing interim analysis cut-offs using Bayesian methods, but with frequentist methods in the final analysis.

2.4 Discussions

Given the large number of therapies available for investigation (10), the need to approve new effective therapies as quickly as possible in order to benefit patients, and the high attrition rates of drugs in oncology, it is of vital importance that the efficiency of the drug development process is optimised. Since a key element of the drug development process is the phase II trial, understanding the decisions that are made by researchers when designing phase II trials reveals areas where their efficiency can be improved. The purpose of this systematic review was to identify current design parameters in use for phase II oncology trials, by assessing trials published in three high impact journals in two years. Simon's design remains commonly used, both response rate and dichotomised time-to-event outcomes are used as primary endpoints, and there are similar numbers of randomised and single-arm designs being used.

While the primary aim of this review was not to identify differences in phase II design choices between the years 2015 and 2019, key comparisons can still be made to reveal possible directions for future phase II designs. The choice of the significance levels used are larger, and the power of the trials are lower. Consequently, large phase II sample sizes have become less frequently used. The review also revealed shifts in the designs used in phase II trials, where A'hern's design has become less frequently used. It was also revealed that there have been slight changes to the type of endpoints used in phase II trials. While most trials still use response rate in recent years, there is a clear rise in the use of time-to-event outcomes, particularly when they are dichotomised. This change could be linked to the types of treatments assessed in the phase II trials: with less cytotoxic chemotherapies and more targeted therapies being tested, it is not surprising that trialists are opting to use time-to-event outcomes, and dichotomising them, so that they span a shorter period of time.

The extent to which comparisons with the reviews of Mariani and Marubini (20) and Langrand-Escure et al. (9) can be made is limited, acknowledging that these reviews had slightly different objectives, focussing on evaluating the incorporation of new designs and quality of reporting, and therefore addressed a few differing parameters. However, where comparisons can be made, these findings highlight the evolution of the use of phase II designs over the years. Mariani and Marubini (20) found 98.7% of phase II trials in their review were single arm, while Langrand-Escure et al. (9) showed that single-arm designs were still the majority (56.6%). This review revealed that randomised designs were more commonly used in recent years, as more researchers opt to conduct the phase II trial using a formal comparative approach. This may be because

single-arm phase II trials use historical control data for comparisons and many problems arise from using them. These include their lack of reliability and the fact that the quality of historical data available may be poor (70). In addition, endpoints measured in historical data may be different from the endpoint measured in the current trial, therefore rendering the data incomparable. This is further highlighted by the review conducted by Mariani and Marubini (20), where they found that 96.8% of trials, published in 1997, used objective response rate and 77.6% of them assessed cytotoxic chemotherapies. The emergence of novel treatment types has meant that other endpoints may be considered more appropriate.

Novel designs, such as basket trials and phase II trials with multiple cohorts assessed within one trial, were also identified. This may be one of the ways by which the efficiency of phase II trials improves, so that they are powered to extract as much information as possible to feed into the proceeding phase III trial, increasing its chances of finding a beneficial treatment (38). The improvement in efficiency is linked to better selection criteria of patients with disease harbouring most appropriate molecular signature. With increasing knowledge and understanding of cancer biology, it is being realised that many genomic aberrations occur across multiple histologies, and a specific molecular profile may be more predictive of drug sensitivity than histology alone (20).

As a consequence of investigating the design parameters used in phase II trials, the quality of reporting also emerged. Reporting of important statistical elements of the design was poor: the target sample size was omitted in approximately 10% of the trials reviewed. The type I and II error rates were also omitted frequently, however, the quality of reporting of phase II trials is improving. Langrand-Escure et al. (9) showed that type I and II error rates were not reported in approximately a quarter of phase II trials. However, in this review, these operating characteristics were included in most of the trials reviewed and, in 2019 they were very rarely omitted. In terms of specifying the statistical design used, Mariani and Marubini (20) reported that only 19.8% of trials reported this information. Again, an improvement in the reporting of the statistical design is evidenced here, where only 10% of trials omitted reporting statistical design parameters, however, the criteria used in this review to establish whether a trial specified the design differs from Mariani and Marubini (20). This is due to the differences in designs found in each period. In their review most trials published in 1997 used single-arm designs, hence they deemed trials with a referenced design as one that specified their methods. In the current era, and as confirmed here, the number of single-arm designs has given way to randomised designs. Therefore, it is no longer appropriate to

assess the specification of a design in the same way as Mariani and Marubini (20). The perspective used here was that if a trial specified the test used for the sample size calculation, then the trial was deemed to have specified the design.

This review also highlighted that more than a quarter of the phase II trials did not meet their target sample size due to poor recruitment. This highlights the need to design feasible phase II trials where the sample size is attainable. Decisions for other design parameters become irrelevant if the trial does not reach its planned completion. It is also clear that phase II trials, in this review, have a positive outcome about 70% of the time, indicating that they should proceed to phase III. The results from subsequent phase III trials were not evaluated here, however, clearly, if the number of robustly designed phase II trials that lead to phase III trials increases, more of them are likely to approve a larger number of novel treatments. It is, therefore, important that phase II trials are designed so that the answers they provide can increase the chances of success in phase III trials.

There are several limitations to this review. There were only three journals included in this review, all of which were of high quality (i.e., possess a high impact factor). The expectation is that this provides information about common practice for the design of phase II trials. However, it does mean that the findings cannot be extrapolated to phase II trials in other journals. This review was also limited to two years only. Despite this the aim of the review, which was to reveal how phase II trials are designed in the current era, was fulfilled and an overview of the designs was presented.

The inclusion criteria, used here, allowed phase II/III trials, however phase I/II trials were excluded. While this may be a limitation, the majority of phase I/II trials aim to establish a dose of treatment and provide estimates for the treatment effect, rather than typically making a decision as to whether or not to proceed to phase III. While standalone phase II trials are sometimes designed to estimate the treatment effect, this review highlights that this is only a minority. In addition, the aim of this review was to explore the choices used for designing a phase II trial that has potential to proceed directly to a phase III trial, hence their exclusion was justified.

In conclusion, it is clear from trials reported in this review that phase II trial sample sizes range between 10 to 150 patients, with type I error commonly set to two-sided 0.05 or 0.1, and type II error typically chosen to be 0.1 or 0.2. The review also identified that Simon’s two-stage design (63) and A’ Hern’s one-stage single-arm design (65) are commonly used, that randomisation and single-arm trials are equally popular choices of designs in phase II, with a binary

endpoint as the chosen outcome. Therefore, even as recently as 2019, it is clear that still there are a broad range of designs, size, endpoints, and error rates. In addition, there are still improvements to be made in terms of the quality of reporting and designing appropriate phase II trials for different settings. There are many options available, and it is important that researchers design phase II trials to fulfil their purpose: to minimise the risk of taking ineffective drugs to phase III or discounting a potentially effective drug at phase II. The parameters identified here are the main choices of oncology phase II trials and therefore will be used in future chapters where their effect on phase II efficiency will be explored.

Chapter 3 Literature Review: what is the best measure to quantify the efficiency of phase II trials?

In Chapter 1, I highlighted that phase II trials lead on to too many unsuccessful phase III trials that fail to identify an efficacious novel treatment. Amongst other factors, this high attrition in phase III trials, can be attributable to the design of phase II trials, which is the focus of this research.

In Chapter 2, I conducted a systematic review which identified how current phase II trials are carried out, in terms of their designs. I found that there is a broad range of designs, sample sizes and endpoints that are used in phase II trials with no clear consensus among researchers as to which parameters contribute to an efficient phase II design.

In this Chapter, I review the literature with the aim of identifying the most appropriate measure to quantify the efficiency of phase II trials.

3.1 Introduction

It is clear from the systematic review (Chapter 2) that current phase II trials are designed using a variety of different methods and statistical techniques. This gives rise to the main question that this research aims to address: what is the impact of using different design parameters on the efficiency of phase II trials and the phase II and III drug evaluation process? This question contains two elements that need to be clarified: Chapter 2 highlighted the design parameters that were selected for evaluation, namely, the choice of endpoint in phase II, the design, specifically whether randomisation and/or multiple stages are incorporated and the sample size of the phase II trials. The second element of the research question is the definition of efficiency of phase II trials. Defining how the efficiency of phase II trials is measured is vital as it is the basis of this thesis, from which conclusions may be drawn.

With different methods and perspectives available to assess the efficiency of phase II trials, the aim of this literature review is to critically evaluate the appropriateness of the measures. Upon doing so, the most appropriate measure for the efficiency of phase II trials will be used to evaluate the effect of different design parameters.

In this literature review, I identify articles describing measures of phase II trial efficiency and assess how applicable they may be to evaluating the performance of the phase II and III trial pathway, on the basis of several criteria.

An appropriate measure of efficiency in this setting needs to fulfil the following criteria:

1. The measure needs to capture the long-term benefits of an efficient phase II design with respect to the phase III outcome. By this I mean the measure needs to include the probability of phase III trial success (probability of whether treatment has been deemed efficacious) under a variety of different scenarios, e.g., when there is a treatment effect and when there is not, and also needs to include the probability of phase II success, under these scenarios. This is vital due to the purpose of phase II trials, which is to screen treatments to evaluate whether further testing in phase III is warranted.
2. The measure needs to be appropriate for testing the efficiency of a series of trials, assessing different treatments, rather than looking at an individual phase II trial with its subsequent phase III trial in isolation. This is important as there are many treatments available for testing yet limited resources to do so. This requirement also follows the sequential nature of the drug development pathway, where testing treatments is an ongoing process.
3. The measure needs to be flexible so that it is able to identify the efficiency of phase II trials under various underlying assumptions, independent of the phase III design choices. This means that an appropriate measure is one that is not confined by parameters other than those attributable to phase II, i.e., the design choice of the phase II are not limited by the design choices in phase III. This would allow the exploration of a wider variety of phase II design parameters.
4. The measure needs to be generalisable so that it can be used to measure the performance of phase II trials for a variety of scenarios and settings. This would ensure that the conclusions made here are not limited to a particular era time or place.

The measure that fulfils all these criteria will be the one chosen to evaluate the effect of the design parameters of phase II trials.

3.2 Methods

The main aim of the literature review is to explore the measures used by researchers to quantify the efficiency of phase II trials. Consequently, the inclusion criteria for this literature review were that articles had to incorporate a measure of efficiency for phase II trial designs. In addition, articles were limited to ones which look at go/no-go phase II trials, i.e., decision making trials, since

the aim of this research is to improve the decision-making process to go from phase II to phase III trials.

A keyword search was conducted using the Ovid MEDLINE database identifying articles published during 1946 to 26th July 2021 and included all the possible variations of terms for phase II and phase III trials, to fulfil the aim of the literature review. The search terms used can be found in Table 3.1.

Table 3.1 keyword terms and combinations used to find relevant data for the literature review

phase II.mp.
phase 2.mp.
pilot stud*.mp.
early-phase.mp.
optim*.mp.
Decision Making/ or Clinical Trials, Phase II as Topic/
phase III.mp.
phase 3.mp.
"go/no-go".mp.
decision point.mp.
1 or 2 or 3 or 4
7 or 8
6 or 9 or 10
11 and 13
5 and 14
12 and 13
15 and 16

*Key: [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]; *truncates word and finds variant word endings*

“Decision making” was included as a term in the search strategy and was mapped to the subheadings “clinical trials” and “phase II”. This limited the search to articles that discuss decision making in phase II clinical trials. “Decision point” and “go/no-go” were also included in the keyword search to expand the number of relevant articles. In addition, different forms of the word

“optimising” was combined with the other search terms to obtain the articles of potential interest. The Boolean operators “OR” and “AND” were used: OR was used to encompass all forms of similar words; AND was used to limit the search by combining two or more words together. Relevant articles were reviewed and from them the search was expanded by a method called pearl growing. This is when relevant articles are chosen from the reference list of articles found. While pearl growing was used to find articles from the relevant articles identified in the search, no further pearl growing was done from those pearl grown articles.

Data from the selected articles were extracted into Microsoft Excel. Findings were summarised narratively, specifically, the authors, year of publication, the aim of the phase II trial (used to whittle down papers if it was not based on a go/no-go trial), the aim/perspective from which the evaluation was taking place, whether the evaluation was considered using a single phase II trial or multiple phase II trials and the measure of efficiency of phase II trials. Articles that did not have a measure of efficiency were excluded, in addition to exploring the performance of a phase II trial that does not have a go/no-go aim.

3.3 Results

The search yielded 291 articles, of which only 55 were deemed relevant, regarding the aims of the literature review and the inclusion criteria highlighted above. These articles’ abstracts were reviewed and 18 were excluded. The remaining 37 full text articles were reviewed and as a result 15 articles were included, directly from the search, from which a further 10 articles were included by pearl growing. Therefore, a total of 25 articles were reviewed with the aim to identify the most appropriate measure that can be used to quantify the efficiency of phase II trials. The flow diagram of this process can be seen in Figure 3.1.

Differences in the literature, regarding the measure to quantify the efficiency of phase II trials, were present. These differences arose due to the choice of context the authors used. Specifically, some researchers explored the effects of design parameters on the efficiency of individual phase II trials, occurring once with a proceeding phase III trial. Others have looked at the efficiency of phase II trials under the context of a more holistic view, where multiple phase II trials are conducted, each testing a different treatment, and if successful lead to a phase III trial, with this process repeated several times. In this case, the efficiency of phase II trials is measured by the long-term performance of the whole drug development process. For each of these approaches, the measure of efficiency used was reviewed in order to identify the most appropriate measure to be used in this research.

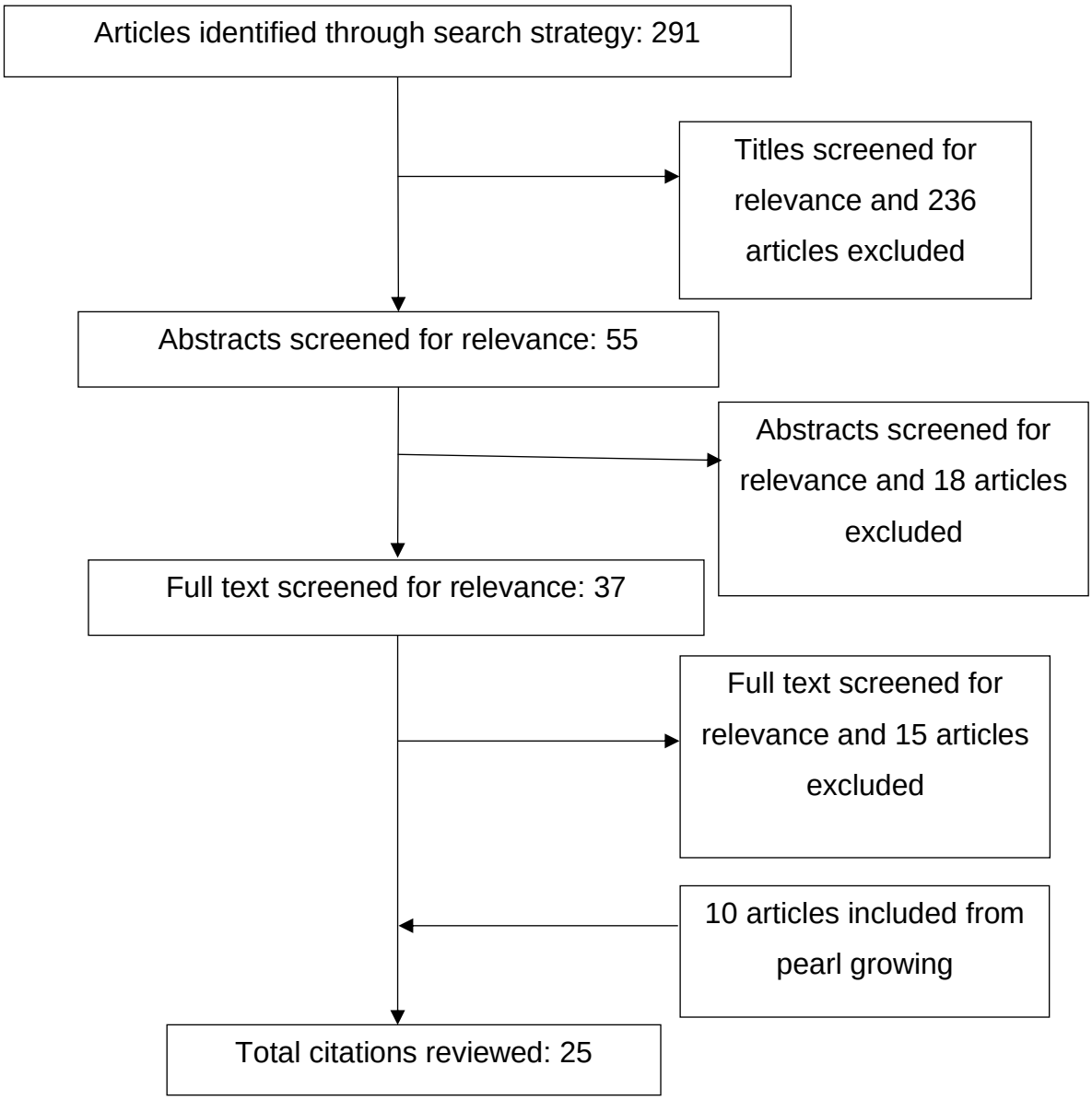


Figure 3.1 Flow diagram of the results of articles found, excluded, and included from the search strategy

The contexts considered are therefore individual and multiple phase II trials. Thus, the following sections review the measures of efficiency within these perspectives. Table 3.2 summarises the measures of efficiencies used by previous researchers.

Table 3.2 measures of efficiencies within the two contexts and broken down by the individual/ multiple perspective taken by the articles

Context	Authors	Measure of efficiency
Individual phase II trials	Sharma and Karrison et al. (71), Sharma and Gray et al.	Proportion of concordant resampled phase II conclusions with parent phase III conclusions

Context	Authors	Measure of efficiency
	(72), Fridlyand et al. (73), An and Han et al.(74)	
	Ayanlowo and Redden (75), Cellamare and Sambucini (76)	Number of patients required in phase II trials
	Chen et al. (77)	Proportion of incorrect decisions to proceed/terminate to a phase III trial, given that the experimental treatment is inefficacious/efficacious
	Taylor et al. (70), Tang et al.(78), Sambucini (79)	Probability of a phase II trial making an accurate conclusion when there is a treatment effect and when there is not
	Chen, Sun and Li (80)	Number of patients required in phase II and III trials
	Preussler et al. (81)	Utility function which takes into account the cost of the program and possible gains after successfully launching the product on the market
	Gottë et al. (82)	Probability of making the correct decision to proceed to phase III and the probability that phase III is successful
Multiple phase II trials	Pond and Abbasi (83)	Proportion of phase III trials conducted using promising agents.
	Marchenko et al. (84), Parke et al. (85)	Expected net present value
	Ding et al. (86), Yao et al. (87), Stallard (88), Leung and Wang (89)	Expected number of patients treated until the first promising treatment is identified in phase II
	Kirchner et al. (90), Hee and Stallard (91) and Keiser, Kirchner et al. (92)	Utility function, specifies cost and gains
	Stallard (39)	Total number of patients required to lead to a successful phase III result

3.3.1 Individual phase II trials

Many authors have investigated the efficiency of phase II trials using resampling methods. This is when patients are resampled from completed phase III trials into simulated phase II trials. The findings in the simulated phase II trials are compared to the conclusions of the phase III trials, from which they were resampled. The efficiency of phase II trials is then measured using the percentage of matching conclusions between the phase II and III trials.

Sharma and Karrison et al. (71) used this resampling method to compare different designs and endpoints used in phase II trials. They used two completed phase III trials; one was positive while the other had a negative outcome. They defined a positive phase III trial as one where the experimental treatment showed statistically significant superiority over the standard or control treatment. Patients were sampled from the experimental and control arm of the phase III trials into the respective arm in the randomised phase II trials. The measure of efficiency used was the percentage of positive phase II trials. To demonstrate efficiency, a high percentage was required when the phase II trials were sampled from the positive trial, while a low percentage of positive phase II trials was required when the patients were sampled from the negative phase III trial.

Another study which assesses the performance of phase II trials through resampling methods was conducted by Sharma and Gray et al. (72). They assessed the performance of different endpoints in terms of their ability to correctly decide to proceed to a phase III trial. They resampled 5000 phase II trials from data from one completed phase III trial in metastatic colorectal cancer. Since the phase III trial was positive, the measure of efficiency was the proportion of positive phase II trials out of the total number of simulated trials. One of the limitations of this study was that they only included one phase III trial, therefore the results were not generalisable, therefore limited to that particular phase III design and the disease area. As such, conclusions cannot be extrapolated to negative phase III trials.

To make their results more generalisable, Fridlyand et al. (73) resampled patients from six phase III trials spanning several disease areas. The studies were a mixture of phases II and III trials and included five trials with a positive outcome and one with a negative outcome. They aimed to compare progression-free survival with percentage change in tumour burden as efficient endpoints for phase II trials. The measure of efficiency was the estimated probability of a correct decision to start a phase III trial, by calculating the number of successful phase II trials out of the 2000 replications of randomised

phase II trials. Even though, they included data from six different trials, they only included one negative trial, which does not reflect the current failure rates of phase III trials in oncology, where 50-60% of phase III trials do not recommend their experimental treatment (10).

Another study, conducted by An and Han et al. (74), used resampling methods comparing the endpoints of phase II trials which included absolute changes in tumour measurements, relative changes in tumour size and tumour response as defined by RECIST. The authors resampled 90 patients into 2000 randomised phase II trials from four completed phase III trials. Two of the phase III trials were positive while the others were negative. The authors assessed the endpoints' ability to inform a correct decision to start a phase III trial by calculating the positive predictive value (PPV) and negative predictive value (NPV). These were defined as the probability that a positive or negative phase III trial is yielded given that the preceding phase II trial was also positive or negative, respectively. An and Han et al. (74) recommend the use of PPV and NPV as they allow the direct measurement of the probability of a positive and negative phase III trial given a positive or negative phase II trial. However, unlike the measures of success used in the other articles (71-73), PPV and NPV require the false-positive rate and the false negative rates, which cannot be estimated from the data. Consequently, the authors treated these as parameters with a range of specified values, and they obtained a range of PPV and NPV values, instead of a single value, by assuming that a certain proportion of treatments would be positive and the remainder negative.

While resampling methods have the advantage of including phase III trial conclusions in the assessment of phase II trial efficiency, allowing the assessment of long-term benefits of the phase II trial, there is also a clear disadvantage. Namely, it confines the resampled phase II trials to include patients with the same type of cancer as in the completed phase III trials, which consequently means that the type of endpoint and relationship between the endpoints are fixed to that particular disease, and therefore, the evaluations are not generalisable. Another problem it poses is the choice of phase III trial to use: if a positive trial is used then a phase II design that gives more positive results will inevitably perform better, whereas it would perform worse if a negative phase III trial was chosen.

Several authors proposed new designs and methods to carry out phase II trials. They evaluated their efficiency by comparing them to some standard phase II designs. The measure of efficiency used by these authors vary depending on the aim of the new design they proposed. Since the aim of this literature review is to identify an appropriate measure of efficiency of phase II trials to use within

my research, critical evaluation of the proposed designs is outside the scope of this research.

Ayanlowo and Redden (75) discuss the efficiency of phase II trials in terms of the number of patients entered into the trial when there is an efficacious treatment and when there is not. They discuss that, in both situations, phase II trials should recruit a restricted number of patients to see the benefit quickly or limit the time patients are exposed to a futile treatment. They argue that current designs do not allow the termination of a futile treatment early enough, and their solution to remedy this is to utilise stochastic curtailment in the phase II designs. This allows researchers to include unplanned interim analyses at unspecified times during the trial. The measure of efficiency they used, to assess their method, was the number of patients recruited into the phase II trial. However, this does not incorporate the conclusions of the phase III trial, and therefore the long-term benefits of a successful phase II trial cannot be assessed.

Cellamare and Sambucini (76) propose a two-arm two-stage design based on a Bayesian predictive approach. The purpose of their design is to ensure a large probability of obtaining substantial posterior evidence that the experimental treatment is efficacious given that it is. They compare their proposed design with Jung's two arm two-stage design (93), in terms of the expected number of patients for the phase II trial. The more patients required the less efficient the design. This is similar to Ayanlowo and Redden's (75) measure and also has the same drawback: it does not include phase III conclusions.

Chen et al. (77) attempted to improve the transitional decision made from phase II trials to phase III by proposing a new statistical decision rule. The decision rule is based on the p-value from hypothesis testing and their new testing confidence value (TCV), that depends on the uncertainty associated with the specified null hypothesis. A simulation study was used to compare the traditional decision rule with their proposed rule. The measure of efficiency was the proportion of incorrect decisions to proceed to a phase III trial, given that the experimental treatment is inefficacious, and the proportion of the incorrect decisions to terminate the development of the drug, given that it is efficacious. While this measure of efficiency is adaptable to different scenarios and can therefore be applied to a number of evaluations, it does not consider the outcome in phase III trials, however this could be incorporated in the measure, if the design of the phase III trial is known.

In the articles reviewed, some authors aimed to evaluate the efficiency of phase II trials with different established design choices. Since the aim of this research is the same, I will evaluate authors design choices in addition to identifying the

measure they used to quantify phase II trial efficiency. The design parameters that have been explored frequently in the literature are whether the phase II trial should be randomised or a single-arm trial, what endpoints should be used and the sample size of phase II trials.

Randomisation in phase II trials was explored by Taylor et al. (70). They compared the efficiency of one-arm phase II trials versus randomised phase II trials, taking into account the historical rate uncertainty (occurs when running a single-arm phase II trial), and ranging the true treatment effects. One of the weaknesses of this study was that they only looked at one type of randomised design and one type of single-arm design. In addition, their statistical definition of significant success of a phase II trial was if the experimental treatment was superior to the control, regardless of the magnitude of the difference between treatment arms, i.e., even if the difference was 0.1 the trial would be deemed successful. The measure of efficiency Taylor et al. (70) used to compare randomised and single-arm designs was the probability of launching a phase III trial when the novel therapy was truly more efficacious than the standard treatment and when it was not. This measure allows the exploration of phase II trial efficiency under different assumptions; however, its drawback is that it does not capture phase III conclusions, therefore there is no way of knowing that a successful phase II will lead on to a successful phase III.

Tang et al. (78) also encourage the use of randomised phase II trials based on simulated error rates of randomised and single-arm designs from individual patient data from a colorectal phase III study and statistical models. The statistical models they created incorporated random and systematic variation in the historical control data. They found that variability in historical control rates and outcome drifts in patient populations over time can result in inaccurate false-positive and false-negative error rates in single-arm designs, however, these factors have little effect on the rates in randomised designs. The measure of efficiency used by Tang et al. (78) is the probability of a phase II trial making an accurate conclusion when there is a treatment effect and when there is not. This is the same as Taylor et al.'s (70) measure and has the same drawbacks.

Sambucini (79) also used the same measure of efficiency as Taylor et al. (70) and Tang et al. (78) when comparing single-arm and randomised phase II designs. However, his explorations were carried out with a Bayesian framework. They assume that the probability of success for the experimental and standard treatments are regarded as random variables, under this framework, as opposed to fixed parameters, under the frequentist framework. Similar to Taylor et al. (70), they also conclude that randomised and single-arm phase II trials are both appropriate in certain situations: when the historical data is correctly

estimated single-arm phase II trials are preferred. If this is not the case, a randomised phase II trial is preferred.

Moving away from the designs of phase II trials to the endpoints of phase II trials, Chen, Sun and Li (80) evaluated early efficacy endpoints for phase II trials. They propose that the choice of endpoint should be determined by evaluating the ratio of benefit of running a trial to its cost. The benefit of running a trial was calculated as the probability of a positive outcome at the end of the phase III, given that the treatment is efficacious. This is contrasted with the cost of running a phase II and a phase III trial, to get a successful phase III trial, and the number of patients required in each. Including the cost of trials is complex and often requires estimates, rather than actual projections, of the cost of a trial. They are also difficult to estimate under different settings and differ between academic and pharmaceutical trials, and even national and international trials. These estimates can quickly become outdated. For these reasons, this measure of efficiency can be difficult to implement or interpret, and may not be generalisable. To remedy this, Chen, Sun and Li (80) use a simplified version of this measure where the benefit term is ignored, and they assumed that the only cost of running a phase II and a subsequent phase III trial is driven by their sample size, i.e., the number of patients required in the phase II and III trials.

Optimising the sample size of phase II trials has been the aim of many articles identified in this literature review. Preussler et al. (81) focus their sample size optimisation on instances where a phase II trial is proceeded by multiple phase III trials assessing the same treatment. They state the regulatory authorities usually require statistical significance in two or more phase III trials. The measure of efficiency they used to optimise the sample size of phase II trials was a utility function, which takes into account the cost of the program and possible gains after successfully launching the product on the market. The inaccuracy of estimating costs and gains of running a trial and the ever-changing costs of trials, renders this measure inappropriate to use.

Gottë et al. (82) also presented an approach in which the phase II sample size was addressed. Their proposed approach involved a decision rule informing the transition to go from phase II to phase III studies. The decision boundaries were selected such that the phase II sample size is minimised given that the probability of making the correct decision to proceed to phase III and the probability that phase III is successful, when the treatment is efficacious. This measure has the advantage of incorporating phase II and III trial conclusions but does not allow the long-term efficiency to be established.

3.3.2 Multiple phase II trials

An alternative approach that authors have used, is optimising, or assessing, phase II trial designs in the context of running the phase II to phase III program and repeating it over a long period of time. Under this context, the measure of efficiency of phase II trials is different and these are reviewed in this section. Articles in this section evaluated different design options for phase II trials. These included evaluating statistical designs of phase II trials and their sample size.

Pond and Abbasi (83) compared randomised with single-arm phase II trials, assuming that a series of phase II trials will be conducted in a particular population over a fixed period of time. They assumed that 1000 patients are available only for the phase II trials. The measure of efficiency they used to compare the phase II designs was the proportion of phase III trials conducted using promising agents. They calculated this measure using the number of correctly successful phase III trials divided by the total number of phase III trials conducted. This measure has many advantages: it includes the outcome of phase III trials; therefore, it captures the long-term benefit of a successful phase II trial; it allows the exploration of different designs and assumptions and is also generalisable to many different scenarios. However, since the drug development process is ongoing, this measure is inappropriate as it takes a snapshot of the efficiency of phase II trials, when no more resources are left to be used. In Pond and Abbasi's (83) case they fix the number of patients available for phase II trials.

Marchenko et al. (84) compared different phase II designs, such as two-arm phase II trials or including interim analyses in the design or conducting Bayesian phase II trials. These designs were compared, mainly using expected net present value which takes into account the total revenue minus the cost of phase II trial minus the cost of phase III. Parke et al. (85) builds on the work done by Marchenko et al. (84) to further compare other phase II designs in combination with subsequent phase III designs in a whole development program. They utilised the same measure to compare the efficiency of the designs. While this measure's strength is that it takes into account the phase III findings, the inaccuracy of estimating the expected revenue and the cost of running trials, in addition to the everchanging prices of running a trial means that the conclusions of these two studies may become outdated.

Ding et al. (86) proposed a decision-making approach which incorporates a Bayesian hierarchical model that allows combining information across several treatments and includes a utility function which considers sampling costs and

possible future payoff. The measure of efficiency was the expected number of patients treated until the first promising treatment is identified in phase II. This measure of efficiency implies that the process of discovering new effective treatments stops when one is found, however, this is not the case in practice and successful phase II trials proceed to phase III.

Stallard (88) also considered optimising the sample size of phase II trials. He assumes that multiple phase II trials are conducted in sequence, each of which can lead to the decision to start a phase III trial or conduct another phase II trial with an alternative intervention. The approach described maximises a gain function including cost of drug development and the benefit from a successful therapy. The gain function is based on the expected gain per patient in the combined phase II and III program. Leung and Wang (89) built on Stallard's (88) work, and therefore used the same measure of efficiency, except instead of just maximising the expected gain per phase II trial, they maximise the rate of gain or total gain for a fixed length of time, since the rate of gain depends on the proportion of treatments that go on to the phase III trial. This is a similar approach as Pond and Abbasi (83), where they the total number of patients is fixed. Leung and Wang's (89) use of a utility function requires specification of cost and gains, which are ever-changing and therefore, this measure may not be generalisable to the future.

Since the sample size calculation of phase III is based on the treatment effect observed in phase II, Kirchner et al. (90) investigated the performance of the phase II/III program as a whole. They state that success in the whole program depends on the allocation of the resources to phases II and III by appropriate choice of the sample size and the rule applied to decide whether to stop the program after phase II or to proceed. Their optimisation was based on a utility function that takes into account the costs and incorporates cost and future revenue gain, corresponding to the expected net present value (NPV), after a successful phase III trial is found.

Hee and Stallard (91) also used a utility function to investigate the size of the population, specifically the impact of a small population on the design of multiple phase II trials. They proposed a Bayesian decision theoretic approach, and an optimal action is taken by considering the design of phase II and III trials, simultaneously. An action is chosen by optimising a gain function which includes economic gains and costs associated with drug development. A decision as to whether to proceed with the current phase II trial is made after observation of each patient.

Keiser, Kirchner et al. (92) also optimised the sample size of phase II trials, but from the perspective that the resulting phase III trials test multiple primary endpoints. They argue that the efficacy of a treatment is often not approved without the confirmation from multiple endpoints in phase III trials. They obtain optimal phase II sample sizes by evaluating a utility function, which again incorporates costs of recruiting patients into the trials and the gains of finding a successful treatment.

All of these articles' (Kirchner et al. (90), Hee and Stallard (91) and Keiser, Kirchner et al. (92)) use of a utility function, specifies cost and gains which are difficult to estimate correctly, and therefore give rise to potential inaccuracies. In addition, these costs may change in the future and are dependent on the setting they are based on, which consequently would mean their findings may not be relevant or generalisable.

In order to improve the rate of success of phase III trials Yao et al. (87) focused on optimising the sample size of phase II trials, which were assumed to be carried out consecutively, in the setting of vaccine development for cancer therapeutics. Specifically, they discussed the appropriate criteria for identifying a sufficiently encouraging therapy and to what extent the sample size depends on the chance that any individual new treatment will be successful in a phase III trial. The measure of efficiency was the total number of patients required before the first promising vaccine is identified. Wang and Leung (94) built on the work presented in Yao et al. (87) but they consider a sequential design of phase II trials. Following Yao et al. (23), they also propose a method where they optimise the number of patients required in each phase II trial.

Stallard (39) also proposed an analytical approach in order to optimise the sample size of phase II trials. He assumed that there is an unlimited number of treatments available for consecutive phase II testing. The author also assumes that the phase II trials are randomised. The measure of efficiency he aimed to minimise was the expected total number of patients required to lead to a successful phase III result, which over the duration of multiple phase II trials is equivalent to maximising the expected number of successful phase III trials. The difference between the frameworks used by Stallard (39) and Yao et al. (87) (and therefore Wang and Leung (94)) is very subtle; they all assume that multiple phase II trials are conducted, however, Stallard (39) assumes that a phase II trial is only successful if it leads to a successful phase III trial, whereas Yao et al. (87) do not incorporate phase III trial conclusions.

3.4 Discussion

In order to investigate the efficiency of different phase II trial design parameters, a measure to quantify the impact of the design choices is needed. The aim of this literature review was to evaluate the measures previously used by authors evaluating different aspects of phase II trials. An appropriate measure was considered to be one that included both the phase II and phase III conclusions. It also needed to be a flexible measure not confined to specific trials or data. It should also be generalisable to different scenarios and disease areas, and applicable to future advancements and changes. Finally, it needed to be appropriate for testing the efficiency of phase II trials when a series of trials are run. This is important in the current era where there is a plethora of treatments available, yet not enough resources to test them all. Table 3.3 summarises the measures of efficiency against the criteria I proposed to identify an appropriate measure to use in this research.

In order to align the methods of this research with the current era, the measure of efficiency used in articles that evaluated phase II trials under the individual trial context were deemed inappropriate. Assuming a framework with an individual phase II trial was perhaps suitable when there was a smaller number of therapies available for testing. However, Sargent and Taylor (95) highlight that currently there are too many drugs that need to be tested in a timely manner. Thus, a more realistic assumption is that several phase II and phase III trials can run either simultaneously or consecutively in a continuous process as developments in medicine constantly occur.

Under the multiple trials framework, several measures of phase II trial efficiency were identified. One of the measures was expected net present value. The main drawback of this measure was the fact that included it included costs and prices of running trials. These values are difficult to estimate with accuracy and even if it is estimated correctly, these costs are open to change due to the inflations in pricing. It is also not generalisable to other currencies. Another measure which had the same issues was the use of utility function with included costs and gains of finding an efficient treatment.

A further measure of efficiency used was the proportion of successful phase III trials out of the number of phase III trials run. This is the measure that is often used to highlight the inadequacy of phase II trials. Using this measure means that a limit is placed on the number of patients or trials to be used.

Table 3.3 checklist of the appropriateness of the measures reviewed

Measure of efficiency	Measure includes phase II and III trial outcomes	Series of trials incorporated	Measure has flexibility to apply to different parameters	Measure is generalisable & applicable to future
Proportion of concordant resampled phase II conclusions with parent phase III conclusions	✓	×	✓	×
Number of patients required in phase II trials	×	×	✓	✓
Proportion of incorrect decisions to proceed to a phase III trial, given that the experimental treatment is inefficacious, and the proportion of the incorrect decisions to terminate the development of the drug, given that it is efficacious	×	×	✓	✓
Probability of a phase II trial making an accurate conclusion when there is a treatment effect and when there is not	×	×	✓	✓
Number of patients required in phase II and III trials	✓	×	✓	✓
Utility function which takes into account the cost of the program and possible gains after successfully launching the product on the market	✓	×	✓	×
Probability of making the correct decision to proceed to phase III and the probability that phase III is successful	✓	×	✓	✓
Proportion of phase III trials conducted using promising agents.	✓	✓	✓	×
Expected net present value	✓	✓	✓	×
Expected number of patients treated until the first promising treatment is identified in phase II	×	✓	✓	×
Utility function, specifies cost and gains	✓	✓	✓	×
Total number of patients required to lead to a successful phase III result	✓	✓	✓	✓

The measure of efficiency which does not have such restrictions and consequently allows the evaluation of the long-term impact of a phase II design parameter is the number of patients required to lead to the first successful

phase III trial. This measure is flexible enough to explore the effects of different assumptions, such as different underlying treatment effects. It is also generalisable to changes in drug development, that may occur in the future, and to different cancer types. Additionally, it accommodates the need to conduct multiple trials, in order to increase the number of efficacious treatments available to patients. This measure has been used by several authors, such as Ding et al.(86), Yao et al. (87) and Wang and Leung (94), however, Stallard (39) stated that this measure is equivalent to the number of successful phase III trials over a long period of time. It is with this measure that the long-term efficiency of phase II trials, with different designs, can be evaluated. For these reasons it was chosen to quantify the efficiency of the design parameters of phase II trials, identified in the systematic review (Chapter 2).

Chapter 4 Evaluating phase II trial efficiency – Methodology

In the previous chapter, measures of efficiency of phase II trials, used by other researchers, were reviewed. With many treatments available for testing, but insufficient resources to do so, the most appropriate measure of efficiency is one that minimises the number of patients required in phase II trials and maximises the number of successful, subsequent phase III trials. Therefore, the number of patients required to lead to the first successful phase III trial, is the measure which best captures these quantities. This measure was used by Stallard (39) where he aimed to optimise phase II trial sample sizes. Therefore, a brief description of the methods employed by Stallard (39) is provided, before describing in detail the methods I use in this research. This chapter also explains which methods are adopted from Stallard (39) and where I build on his methods to meet the aims of this research.

4.1 Stallard's (39) methods and assumptions

The aim of Stallard's research (39) was to optimise the phase II trial sample sizes. This optimisation problem was viewed from the perspective of a large pharmaceutical company or funding body, who have the capacity to test a large number of experimental therapies in multiple trials, either consecutively or simultaneously. This perspective has also been used by other researchers such as Pond and Abbasi (83) and Ding et al. (86). Their approaches are further discussed, later in this chapter (Section 4.2.1). In order to make this process as efficient as possible Stallard (39) aimed to maximise the number of experimental treatments that showed efficacy for a variety of phase II sample sizes.

Due to the phase II trials' place in the clinical evaluation of a new therapy, the phase II trial cannot be investigated in isolation. Stallard (39), therefore, considers the optimisation problem with the whole drug development pathway in mind. Figure 4.1 depicts the drug development process used; this image was adapted from Stallard (39). It was assumed that the treatments available have already been deemed safe in phase I trials, and therefore are available for testing in phase II trials. Each treatment is tested consecutively; if the treatment shows sufficient efficacy it proceeds to a phase III trial. However, if the treatment is found to be futile, the process terminates and another therapy is tested in a new phase II trial. Once a treatment proceeds to phase III, the

therapy is tested and is deemed successful if it shows superiority over the standard treatment. If the therapy is unsuccessful at phase III, the investigations for this particular therapy terminates and a different treatment is then selected to be tested in another phase II trial. The process continues until the first successful phase III trial is found. Stallard (39) assumed that the objective of the optimisation problem is to minimise the expected total number of patients required to lead to a successful phase III trial. Stallard (39) states that over a long run of experimentation, this outcome is equivalent to maximising the expected number of successful phase III trials.

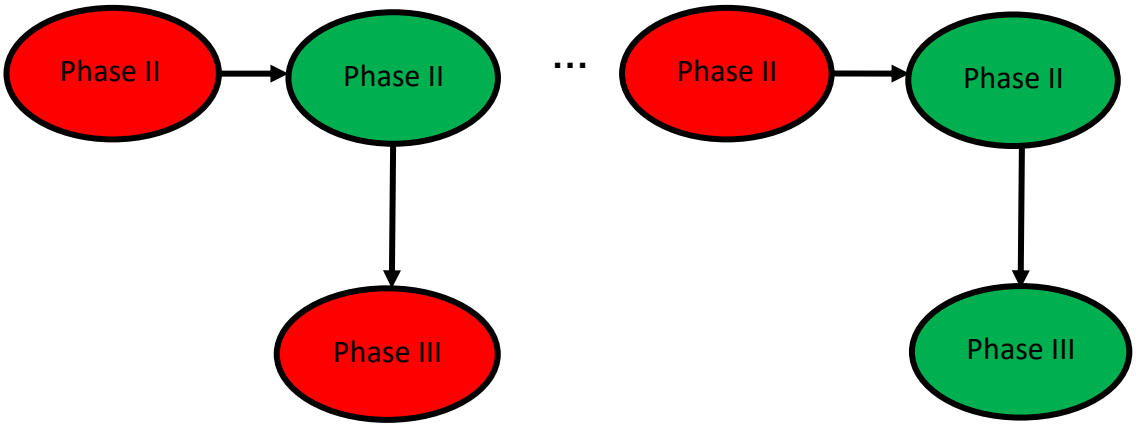


Figure 4.1 An illustration depicting the phase II-III component of the drug development process and assumptions of the research project; green indicating trial success and red indicating trial failure

4.1.1 Obtaining the measure of phase II trial efficiency

Stallard (39) assumed that the phase II and III trials are single-stage and randomised with patients assigned into two parallel groups with a 1:1 ratio. He also assumed that the endpoint in phase II and III trials is the same, continuous and follows a normal distribution with a known variance. In addition, he assumed that there is an underlying true treatment effect distribution for each therapy that is evaluated and each of these distributions are identical.

The clinically relevant difference between the means for the treatments in the control and experimental arms was assumed to be fixed and known, denoted by δ. The patient responses were assumed to have a known variance equal to 1, so that it represents a standardised effect size. Since Stallard (39) assumed the phase II and III trials were randomised two-arm studies, the sample size of the phase II and III trials, n_1 and n_2, were obtained using the following formula:

$$n_i = \frac{4\sigma^2\left(z_{1-\alpha_i} + z_{1-\beta_i}\right)}{\delta^2}, i = 1, 2$$

where $\sigma^2 = 1$, since the variance was known and equal to 1 (as mentioned above), $i = 1$ represents the parameters associated with the phase II trial, while $i = 2$ represents the parameters associated with the phase III trial, $z_{1-\alpha_i}$ and $z_{1-\beta_i}$ are the standard z-score evaluated at the one-sided type I error rate, α_i, and the power, $1 - \beta_i$, respectively.

The phase III trial operating characteristics were fixed and thus the sample size of the phase III trial was also fixed, while the phase II trial sample size depends on the values of α_1 and β_1. He utilised power functions for the trials which represented the probability of a significant result at α_i when the true standardised treatment effect is equal to θ:

$$P_i(\theta, \alpha_i, \beta_i) = \Phi\left(\frac{\left(z_{1-\alpha_i} + z_{1-\beta}\right)\theta}{\delta} - z_{1-\alpha_i}\right), i = 1, 2 \qquad (4.1)$$

$$P_i(\theta) = P_i(\theta, \alpha_i, \beta_i), i = 1, 2$$

Therefore, $P_1(\theta)$ and $P_2(\theta)$ denote the probability that a true effect size, θ, leads to a significant result in a phase II and a phase III trial, respectively. Given the fact that a phase III trial is only conducted after a significant result in the preceding phase II trial, the probability of conducting the phase III trial is given by $P_1(\theta)$, while the probability of conducting a phase III trial and finding it to be successful is equal to the product of $P_1(\theta)P_2(\theta)$.

Stallard (39) came to the conclusion that the total expected sample size (over phase II and III) per significant result in a phase III trial is equal to

$$\frac{n_1 + n_2\, E(P_1)}{E(P_1 P_2)} \qquad (4.2)$$

Where n_1 is the sample size required for the phase II trial and depends on the type I and II error rates selected for the phase II trial, denoted by α_1 and β_1, respectively; n_2 is the sample size required for the phase III trial and depends on the type I and II error rates set for the phase III trial, denoted by α_2 and β_2, respectively; and $E(P_1)$ is the expected probability of a successful phase II trial over the assumed prior distribution of the treatment effect; while $E(P_1 P_2)$ is the expected probability that the phase II trial was a success and the subsequent phase III trial is also successful.

As mentioned above, $E(P_1)$ and $E(P_1 P_2)$ depend on the form of the assumed prior distribution for the treatment effect θ. Stallard (39) presented two examples with different prior distributions, in order to demonstrate what the optimal sample sizes of phase II trials are under the assumptions given. The first prior distribution he assumed was the two-point prior, with the aim of illustrating how the calculations are implemented. The treatment effect, θ, is assumed to follow

a two-point prior with mass at values 0 (no effect) and δ (effect of size equal to clinically relevant difference), such that $p(\theta = \delta) = \pi$ and $p(\theta = 0) = 1 - \pi$ so that, π is the probability that the therapies, assessed in consecutive phase II trials, are effective.

In order to obtain the formula for the expected number of patients per successful phase III trial, for this two-point prior distribution, $E(P_1)$ and $E(P_1P_2)$ need to be calculated, as shown in Equation 4.2. In calculating $E(P_1)$ and $E(P_1P_2)$, I use the fact that if X is a discrete random variable with probability mass function $p(x)$, then the expectation of X is defined as $E(X) = \sum_x x\, p(x)$.

In the two-point prior example the probabilities of the random variable are $p(\theta = \delta) = \pi$ and $p(\theta = 0) = 1 - \pi$, therefore the expected value of P_1 is given by:

$$E(P_1) = p(\theta = 0) \times P_1(0, \alpha_1, \beta_1) + p(\theta = \delta) \times P_1(\delta, \alpha_1, \beta_1)$$

Since, $p(\theta = 0) = 1 - \pi$ and $p(\theta = \delta) = \pi$, these can be substituted into the equation to give

$$E(P_1) = (1 - \pi) \times P_1(0, \alpha_1, \beta_1) + \pi \times P_1(\delta, \alpha_1, \beta_1) \quad (4.3)$$

$P_1(0, \alpha_1, \beta_1) = \alpha_1$ and $P_1(\delta, \alpha_1, \beta_1) = 1 - \beta_1$; hence substituting these values into Equation 4.3 yields

$$E(P_1) = (1 - \pi)\alpha_1 + \pi(1 - \beta_1) \quad (4.4)$$

Similarly to calculate the expectation of P_1P_2, the same methods are utilised:

$$E(P_1P_2) = p(\theta = 0)P_1(0, \alpha_1, \beta_1)P_2(0, \alpha_2, \beta_2) + p(\theta = \delta)P_1(\delta, \alpha_1, \beta_1)P_2(\delta, \alpha_2, \beta_2)$$

Recall $p(\theta = 0) = 1 - \pi$, $p(\theta = \delta) = \pi$, $P_1(0, \alpha_1, \beta_1) = \alpha_1$, $P_1(\delta, \alpha_1, \beta_1) = 1 - \beta_1$ and since $P_2(0, \alpha_2, \beta_2) = \alpha_2$ and $P_2(\delta, \alpha_2, \beta_2) = 1 - \beta_2$, thus these can be substituted into the equation to give:

$$E(P_1P_2) = (1 - \pi)\alpha_1\alpha_2 + \pi(1 - \beta_1)(1 - \beta_2) \quad (4.5)$$

Hence, the expected number of patients required per successful phase III trial, under the assumption of a two-point prior distribution, can be obtained by substituting Equations 4.4 and 4.5 into Equation 4.2. Varying the value of α_1 and β_1, while fixing the values of α_2 and β_2, can be used to investigate which pair of error rates give an optimal choice for phase II trials.

Stallard (39) recognised that the two-point prior was not very realistic and was only presented for demonstrative purposes. Therefore, Stallard's (39) second example was based on a more realistic situation in which the prior distribution for θ was assumed to take the normal form, specifically $\theta \sim N(\mu_0, \sigma_0^2)$. Similar

methods to the two-point prior example were employed in order to obtain the formula for the expected number of patients per successful phase III trial, given a normal prior distribution. In this case I use the result that if X is a continuous random variable with a probability density $f(x)$ then $E(X) = \int_{-\infty}^{\infty} xf(x)\,dx$.

Therefore, the expected probability of success in a phase II trial, $E(P_1)$, can be found by integrating, under the assumed prior normal distribution and the power function of the phase II trial (Equation 4.1) with respect to θ. Similarly, the probability of success in both phase II and III trials, $E(P_1P_2)$, can be found by integrating, under the normal prior distribution, the power function of the phase II trial and the power function of the phase III trial (Equation 4.1) with respect to θ.

Mathematically, this translates to Equations 4.6 and 4.8, respectively:

$$E(P_1) = \int_{-\infty}^{\infty} \frac{1}{\sigma_0}\phi\left(\frac{\theta-\mu_0}{\sigma_0}\right)\ \phi\left(\frac{(z_{\alpha_1}+z_{\beta_1})\theta}{\delta} - z_{\alpha_1}\right) d\theta \qquad (4.6)$$

Let $f(\theta,\mu,\sigma^2) \equiv \frac{1}{\sigma}\phi(\frac{\theta-\mu}{\sigma})$ and $P_1(\theta) = \int_{-\infty}^{\theta} f(x,\mu_1,\sigma_1^2)\ dx = \int_{-\infty}^{0} f(x+\theta,\mu_1,\sigma_0^2)\ dx$

Therefore, $E(P_1) = \int_{-\infty}^{\infty} f(\theta,\mu_0,\sigma_0^2)\int_{-\infty}^{0} f(x+\theta,\mu_1,\sigma_1^2)\ dxd\theta$

$$E(P_1) = \int_{-\infty}^{0}\int_{-\infty}^{\infty} f(\theta,\mu_0,\sigma_0^2)f(x+\theta,\mu_1,\sigma_1^2)\ d\theta dx$$

$$E(P_1) = \int_{-\infty}^{0} f(x,\mu_1-\mu_0,\sigma_0^2+\sigma_1^2)\,d\theta dx$$

Hence, the expected probability that a phase II trial is successful can be found when $P(X<0)$ where $X \sim N(\mu_1-\mu_0,\sigma_0^2+\sigma_1^2)$ (4.7)

Similarly, Stallard (39) found $E(P_1P_2)$ to be equal to

$$E(P_1P_2) = \int_{-\infty}^{\infty} \frac{1}{\sigma_0}\phi\left(\frac{\theta-\mu_0}{\sigma_0}\right)\ \phi\left(\frac{(z_{\alpha_1}+z_{\beta_1})\theta}{\delta} - z_{\alpha_1}\right)\ \phi\left(\frac{(z_{\alpha_2}+z_{\beta_2})\theta}{\delta} - z_{\alpha_2}\right) d\theta \qquad (4.8)$$

and $P(X_1<0, X_2<0)$ where

$$\begin{pmatrix}X_1\\X_2\end{pmatrix} \sim BVN\left(\begin{pmatrix}\mu_1-\mu_0\\\mu_2-\mu_0\end{pmatrix},\begin{pmatrix}\sigma_0^2+\sigma_1^2 & \sigma_0^2\\ \sigma_0^2 & \sigma_0^2+\sigma_2^2\end{pmatrix}\right) \qquad (4.9)$$

where $\mu_i = \frac{z_{\alpha_i}\delta}{z_{\alpha_i}+z_{\beta_i}}, i = 1,2$ and $\sigma_i = \frac{\delta}{z_{\alpha_i}+z_{\beta_i}}, i = 1,2$

Given values for the type I and II error terms for phase II and III trials, the probabilities were obtained and thus substituted into Equation 4.2 to calculate the expected number of patients per successful phase III trial. The values of the type I and II errors which yield the smallest expected number of patients yielded the optimal sample size for phase II trials.

4.2 Methods and assumptions employed in this research

In this section, I will detail the methods I have adopted from Stallard (39) and how I adapt his methods to meet the aims of this research. Recall that, as mentioned previously, the aim of this research is to investigate the effect of different phase II design parameters on the efficiency of phase II trials. As reported in the systematic review (Chapter 2), the design choices of phase II trials are varied, with multiple options available, specifically, the use of randomisation and single-arm designs, running multiple stage designs, the sample size, determined by the choice of the type I and II error and the endpoint choice in phase II trials. Thus, the objective of this research is to explore the effect of these design choices on phase II trial efficiency. Specifically, the questions I will address are:

- Endpoints of phase II trials:
    1. Do the phase II and III endpoints need to be correlated in order to improve efficiency?
    2. What other factors influence the efficiency of phase II trials; is correlation the most important factor to ensure phase II efficiency or is there other factors that also have an impact?
- Design of phase II trials:
    1. Is a randomised design more efficient than a single-arm design using either two-stage or single-stage designs?
    2. Is it more efficient to use a single-stage design over a two-stage design?
    3. Which design is the most efficient?
- Sample size of phase II trials:
    1. Does increasing the sample size increase efficiency or is there a point where growing the sample size would lead to inefficiencies?
    2. What is the effect of different type I and II errors on the efficiency of phase II trials?

Answering these questions will equip researchers, designing phase II trials, with the information they need to understand the impact of their choices of the design parameters. In addition, recommendations about the most efficient

design can be made. The context and methods used to answer these research questions are described below.

4.2.1 Context

The literature review (Chapter 3) highlighted different contexts that authors used to investigate the efficiency of phase II trials. It was concluded that the most appropriate context for assessing the efficiency of phase II trials is from the perspective of running multiple phase II and III trials in a continuous manner. This is due to the fact that in the current era there are a number of treatments available for testing, but the resources available to test these treatments are extremely limited (96). It is, therefore, this context that is used in this research to explore the effects of different design parameters on the efficiency of phase II trials.

This context has been used by many authors, including (but not limited to) Pond and Abbasi (83), Ding et al. (86) and, as explained above, Stallard (39). Pond and Abbasi (83) compare the efficiency of randomised two-stage and single-arm two-stage phase II trials from the point of view of a “clinical research organisation, industrial sponsor, or cooperative group”. They specify that, via simulation, they will run multiple phase II trials in a specific group of patients over a known period of time. Consequently, they limit the number of patients to 1000. The fact that Pond & Abbasi (83) limit the population size is unrealistic and not widely generalisable, as the population size is often not known, and usually much larger. This is evidenced by the fact that cancer patients are constantly increasing, with about 1000 new cases each day (97).

A more realistic approach was the one proposed by Ding et al. (86) and Stallard (39). Ding et al. (86) consider the phase II trials to run sequentially with an indefinite number of new treatments available. Stallard (39) also considers optimising phase II trials from the point of view of a “large funder of clinical trials, such as a public sector research body or large pharmaceutical company”. He states that such bodies have the capacity to run multiple trials, however, each trial needs be executed as efficiently as possible. Therefore, the context of the evaluations of phase II design parameters is a simple model of the drug development process. I assume an infinite number of treatments are available for testing in phase II, i.e., they have already been deemed safe and a recommended phase II dose is available, and that testing occurs in a sequential manner. Each treatment is tested consecutively first in a phase II trial and if successful proceeds to phase III (Figure 4.1).

Both Stallard (39) and Ding et al.'s (86) context lends itself to minimising the number of patients required to lead on to the first successful phase III trial. It is, therefore, this measure that is used in this research, however, where previous authors have aimed to optimise this measure, I will report the effect of different design parameters of phase II trials on this measure. I will therefore build a profile of how each parameter choice, in different combinations, affects the efficiency of phase II trials. Ultimately, this will allow trialists to design efficient phase II trials, under different scenarios.

4.2.2 Statistical evaluations

The narrative for the evaluations of the endpoint, design and sample size of the phase II trials is that there exists an underlying true treatment effect, Δ, for the treatments under evaluation. This is assumed to follow a normal distribution. This was the distribution chosen by Stallard (39), stating that it is a realistic scenario. This is due to the fact that it incorporates an array of effective and ineffective treatments, and therefore incorporates the uncertainty about the true treatment effects. By specifying a distribution for the treatment effects, the efficiency of phase II trials can be assessed for both effective and ineffective treatments. A treatment effect, θ, is randomly selected from this distribution to be evaluated in phase II, and if the treatment shows sufficient efficacy, the same treatment is tested with the same true treatment effect, θ, in phase III.

Stallard (39) states that he used Bayesian decision theory to evaluate phase II sample sizes. Even though the methods employed here are strictly frequentist, some of his methods are used as a basis of this research. The true treatment effect distribution was inspired by Stallard (39), where he assumed a prior distribution about the treatment effect of the available treatments. The treatment under evaluation is assumed to have an effect drawn from that distribution. In Bayesian methodology the prior distribution would be updated to obtain the distribution conditional on the observed data. This is the posterior distribution and would be used to draw inferences in the analysis of the phase II and III trials. However, the role of the true treatment effect distribution in this research, and in Stallard's (39) work, is used in a manner analogous of the frequentist methods. When evaluating the effect of an experimental treatment, compared with a concurrent control or historical control, the effect of the experimental treatment, is drawn from the assumed distribution of available treatments. The observed data is then drawn from a distribution that has a mean given by this treatment effect. The observed effect between the experimental and concurrent control or historical control is compared to a prespecified minimum clinically

important difference (MCID), which is estimated by trialists and experts based on previous research. Under the frequentist approach, the true treatment effect is unknown and is fixed. A test is conducted to decide whether or not the observed treatment effect is clinically and statistically promising as denoted by MCID. In order to make these conclusions regarding the observed treatment effect, the probability of all possible data in the sample space given the pre-specified fixed value of the unknown parameter (the treatment effect), is obtained, and then used to see how concordant the observed data are with the pre-specified assumption of the MCID.

The conclusions made at the end of the trial are based on observed data. This means that there is scope for errors to occur. The errors that may arise are known as the type I and II errors. The type I error is the probability that the treatment has no or limited effect, yet is indicated to be of benefit to patients in the trial. The other error that can also occur is the type II error, which is defined as the probability of not accepting an efficacious treatment. The power of the trial is the probability of making the correct decision, and is the complement of the type II error. To ensure that the conclusions of the trial can be made with high certainty, these errors are controlled by the trialists at the design and analysis stages: the type I error is usually set in the design and the power set by the choice of sample size. It is important to note that at the analysis stage, decisions are based on hypothesis tests, while at the design stage, decisions are based on the pre-specified operating characteristics of these hypothesis tests, in other words, the design of the trial is dependent on the choice of the type I and II errors. Details of the designs of phase II and III trials is provided later in Section 4.2.3.

While the true treatment effect distribution is assumed to follow a normal distribution with a mean of zero throughout the investigations, its variance changes depending on the parameters investigated, in this research. In Chapter 5, where the aim is to investigate the effect of the correlation between phase II and III trial endpoints, the variance of the true treatment effect distribution varies between 0.1-10 in increments of 0.3. This was chosen as it was found that the correlation was dependent on the variance of the true treatment effect (more details on this is provided in Chapter 5). However, when investigating the effect of design and sample size of phase II trials (Chapters 6 and 7, respectively), the true treatment effect follows a standard normal distribution, with a mean of zero and variance of one. In Chapters 5, 6 and 7, a sensitivity analysis is conducted to assess the robustness of the conclusions to the assumption that the mean treatment effect is zero on average. Table 4.1 summarises the values for the

true treatment effect distribution and the parameter values for the designs of the phase II and III trials. Details and explanations for the choices of the designs are provided in the next section (Section 4.2.3).

Table 4.1 summary of parameter values and design of phase II trials

Parameters	**Value**		**Notes**
	Chapter 5	**Chapters 6 &7**	
Underlying treatment effect, Δ			
Mean, μ	0	0	Standard normal distribution – realistic example from Stallard (39)
Variance, σ^2	0.1-10 (increases of 0.3)	1	
Phase III design			
Type I error, α_3	0.05, two-sided	0.05, two-sided	conventional levels: 5% for type I error, 80% or 90% for power (98)
Type II error, β_3	0.2	0.2	
Clinically significant difference (csd), δ_3	0.3	0.3	Most common in (98)
Endpoints	Continuous	Continuous	
True treatment effect, θ	Randomly selected from Δ	Randomly selected from Δ	
Phase III control, μ_1	0 (fixed)	0 (fixed)	True mean response in control arm
Phase III experimental, μ_2	Equivalent to $\theta + \mu_1$	Equivalent to $\theta + \mu_1$	True mean response in experimental arm
Phase II design			

Parameters	Value		Notes
	Chapter 5	**Chapters 6 &7**	
True treatment effect, θ	Randomly selected from Δ **(continuous scale)**	Randomly selected from Δ **(continuous scale)**	In chapters 6 and 7 the phase II endpoint is binary therefore θ, which is the mean difference needs to be transformed to the log-odds scale using $\theta \frac{\pi}{\sqrt{3}}$
Phase II control, p_1	N/A endpoint continuous	0.25 (fixed)	Proportion of success in control arm (concurrent or historical)
Phase II experimental, p_2	N/A endpoint continuous	$\frac{\left(e^{\theta\frac{\pi}{\sqrt{3}}}p_1\right)}{(p_1\left(e^{\theta\frac{\pi}{\sqrt{3}}} - 1\right) + 1)}$	Proportion of response in phase II experimental arm
Phase II control, μ_1	0 (fixed)	N/A endpoint binary	True mean response in control arm
Phase II experimental, μ_2	Equivalent to $\theta + \mu_1$	N/A endpoint binary	True mean response in experimental arm
Endpoints	Continuous	Binary	
Clinically significant difference (csd), δ_2	0.3 (same as the phase III csd)	0.2 (Difference in proportions)	Systematic review
Type I error, α_2	One-sided 0.05	One-sided 0.05 (Chapter 6 only, varied in Chapter 7)	Systematic review

Parameters	Value		Notes
	Chapter 5	Chapters 6 &7	
Type II error, β_2	0.2	0.2 (Chapter 6 only, varied in Chapter 7)	Systematic review

4.2.3 Trial designs

4.2.3.1 Phase III trials

Throughout all the evaluations the phase III design is fixed. It was assumed to be randomised with a 1:1 ratio of patients allocated to the control and experimental arm, with a purpose of confirming whether the experimental drug is superior to the control arm therapy. The design of the phase III trial was fixed with a type I error rate, $\alpha_3 = 0.05$, power, $1 - \beta_3 = 0.8$ and a standardised targeted treatment effect, $\delta_3 = 0.3$. The clinically significant difference that is targeted in phase III is assumed to be 0.3 on the continuous scale. This was found to be the most common targeted effect in a systematic review exploring common operating characteristics in phase III trials (98).The type I and II errors were chosen as they represent what is typically used in phase III clinical trials (98). Using these values, the sample size of the phase III trial was determined using Equation 4.10,

$$n_3 = \frac{4\left(z_{1-\frac{\alpha_3}{2}} + z_{1-\beta_3}\right)^2}{{\delta_3}^2} \qquad (4.10)$$

where $z_{1-\frac{\alpha_3}{2}}$ and $z_{1-\beta_3}$ are the standard z-score evaluated at the 5% level and 20% level, respectively. For each of the phase III trials that were initiated, the significance level was two-sided. The formula yielded a phase III trial total sample size of 348 patients.

Patients for the phase III trial were randomly sampled from the true distributions for each treatment arm – patients accrued into the control arm were sampled from the true control arm distribution, and similarly patients enrolled into the experimental arm were sampled from the true experimental arm distribution. Since the endpoint in the phase III trials was assumed to be continuous (e.g., size of tumour in cm), the true patient distributions for phase III trials were assumed to be normally distributed. The value for the mean of the true control distribution was assumed to be $\mu_1 = 0$, while the mean for the true experimental distribution was assumed to be μ_2, which was obtained, as explained above.

The standard deviation for both true patient distributions were assumed to be the same and took values $\sigma_1 = \sigma_2 = 1$.

A t-test was used to determine whether the phase III trial indicated that the experimental treatment is better than the standard therapy. The phase III trial was deemed successful if the p-value is smaller than or equal to the type I error which was set to $\alpha_3 = 0.05$, otherwise the phase III trial was deemed to be unsuccessful.

4.2.3.2 Phase II trials

The design of the phase II trial differs when I am evaluating different design parameters. In Chapter 5, when evaluating the relationship between the endpoints, the phase II and III trial endpoints are both continuous. As the same treatment is tested in phase III, if it is found to be efficacious in the preceding phase II trial, the clinically significant difference (CSD) $(= 0.3)$ and power $(= 0.8)$ in phase II is the same as that in the phase III trial design. Therefore, the only difference that arises in phase II compared to the phase III is the choice of the significance level. In the phase II trial, it is a one-sided significance level of 0.05, while in phase III it is a two-sided significance level of 0.05. It is not surprising that the sample sizes of the phase II and III trials are very similar. This set up does not reflect what occurs in the drug development process: phase II and III trials sample sizes are different and the endpoints are typically not the same. However, this is done to lay the theoretical foundations so that the relationship between the phase II and III trial endpoints can be explored. The results from these evaluations will then be used to feed into more realistic scenarios where the sample sizes in phase II are smaller and the endpoints in phase II and III are different.

In Chapters 6 and 7, the phase II design is assumed to have a binary outcome. This was the most common endpoint type found in the systematic review. For those articles that used a binary outcome the median difference in proportions commonly targeted was 0.2. For this reason, this was the clinically significant difference used to design the phase II trials in these chapters. In Table 4.1, the phase II control rate, p_1,refers to the largest unacceptable response rate and represents the fixed true response rate in the control population (since this is a single-arm trial). The true phase II experimental rate, p_2 is equal to p_1 plus the true treatment effect, θ , i.e., $p_2 = p_1 + \theta$, and refers to the true underlying response rate of the experimental treatment. In specific phase II designs, such as Simon's two-stage design p_2 is defined as the smallest acceptable response rate required to be observed in the experimental arm to warrant further

evaluation in phase III, and reflects the targeted treatment effect. This value is required to calculate the sample size and decision criteria for the trial. In this thesis, this value is referred to as $p_1 + \delta_2$, and is fixed at 0.45, since $p_1 = 0.25$ and the targeted treatment effect, δ_2, is fixed at 0.2.

Success in phase II trials is reached if the treatment effect tested is significant at a one-sided significance level of 0.05 (in Chapter 6), power of 0.8 (in Chapter 6) and a difference in proportions larger than 0.2 between the control (historical or concurrent) and experimental arm. The treatment effect, selected from the normal distribution, when tested in the phase II trials with a binary endpoint needs to be transformed to the log odds scale using $\theta \frac{\pi}{\sqrt{3}}$. Therefore, if the p-value < 0.05 then phase II is deemed successful. In Chapter 7, when assessing the effect of the phase II trial sample sizes, n_2, the choice for the type I and II errors (α_2 and β_2 respectively) will vary to a range of values, however, when assessing the effect of phase II trial endpoints and designs, α_2 is fixed at the one-sided level of 0.05, as previously mentioned, and β_2 is set to 0.2.

The findings of each evaluation will determine the design choice of the phase II trial. In Chapter 5 the objective is to explore the relationship between the endpoints and the findings in this chapter will be used in subsequent evaluations of the design in Chapter 6 and sample size in Chapter 7. Similarly, the most efficient design in Chapter 6 will be the design used to explore the effect of the sample size in Chapter 7. Further details of the sample size calculation and designs of the phase II trials will be explained in the specific chapters for each of the parameter evaluations. Refer to Table 4.1 for the summary of the design parameters of the phase II and III trials.

4.2.4 Evaluating phase II trial efficiency

The chosen measure used in this research to quantify the efficiency of phase II trials, is the one used by Stallard (39), namely, the number of patients required to lead on to the first successful phase III trial. As explained in section 4.1.1, $E(P_1)$ denotes the expected probability of a successful phase II trial averaged over the prior distribution (39). Similarly, the expected probability of successful phase II and III trials is denoted by $E(P_1P_2)$. Therefore, to calculate the expected conditional probability of success in a phase III trial given the preceding phase II was successful is defined as $\frac{E(P_1P_2)}{E(P_1)}$. Stallard (39) uses this result to define the number of patients required to lead to the first successful phase III trial (repeated here for ease of reference):

$$\frac{n_2 + n_3\, E(P_1)}{E(P_1P_2)} \qquad (4.2)$$

Using this measure, I will evaluate the effect of the correlation between endpoints in phase II and III trials, the design of phase II trials, when they are single-arm, randomised, single-stage or two-stage, and the sample size of phase II trials.

Initially, the evaluations are conducted analytically: the methods used by Stallard (39), described in section 4.1.1, are adapted to allow the exploration of the effect of the relationship between the phase II and III trial endpoints. Simulations, conducted on the statistical software R, are then used to explore the effect of the design and sample size of phase II trials. Since these parameters have multiple varieties to explore, simulations provide flexibility and efficiency; the parameter choices explored in Chapters 6 and 7 would have been computationally intensive.

The most efficient set up of the simulations was to define a population of patients, N, that can be entered in multiple phase II and III trials. This was set to be a large number so that it can reflect what occurs over a very long period of time, and so it does not affect the results and conclusions. Stallard (39) notes that the measure of phase II efficiency, namely the number of patients required to lead to the first successful phase III trial, is equivalent in its conclusions to the number of successful phase III trials, N_{trial}. Hence, this was the measure used to compare the designs evaluated in the simulations (Chapters 6 and 7). However, in order to obtain the number of patients required to lead to the first successful phase III trial, a simple calculation $\left(\frac{N}{N_{trial}}\right)$ is done to obtain the number of patients required, so as to make the conclusions for each evaluation comparable.

Chapter 5 Investigating the effect of the relationship between phase II and III endpoints

In Chapter 4, I outlined the methods for this research, stating that the aim is to investigate the effect of the endpoint, design and sample size of phase II trials on their efficiency. In this chapter, the first parameter is investigated. Here, I aim to discover what the impact of the relationship between phase II and III trial endpoints is and how influential it is, in terms of the number of patients required to lead to the first successful phase III trial. The problem is answered analytically and provides a basis for future investigations in the coming chapters.

5.1 Introduction

In clinical trials, an endpoint is used to measure how a patient responds to an intervention. It can quantify the wellness, certain body functions or survival of patients (99). In oncology, phase III trials usually measure patients' overall survival, or in some disease areas disease-free or progression-free survival. This is an example of a true clinical endpoint (100) and usually requires a long follow-up time. While this may be acceptable in confirmatory phase III trials, phase II trials typically use a short-term endpoint, since their purpose is to quickly and efficiently move treatments on to phase III trials, for further investigation. Using the most appropriate endpoint can influence the conclusions of the trial and therefore contribute to the advancement or hindrance of medical discoveries.

Oftentimes, oncology phase II trials are conducted with the proportion of objective responders as the primary measure of efficacy. This is reflected in the findings of Chapter 2 where the majority of phase II trials, included in the systematic review, used binary outcomes (80%; response rate, 51%; dichotomised time-to-event, 25%; safety rate, 4%). Langrand-Escure et al.'s (9) systematic review also reiterates this: they report that 80.7% of phase II trials carried out between the years of 2010 to 2015 used response rate as the primary endpoint. Nonetheless, there is still much dispute in the literature in regards to the most appropriate endpoint for oncology phase II trials. An et al. (10) discuss a number of issues with response rate as an outcome measure in phase II trials, including the lack of concordance between response rate in phase II trials and endpoints used in definitive phase III trials. They also state

that for solid tumours, dichotomising tumour measurements, which is a continuous measure, can lead to a loss of potentially valuable information. In addition, the cut off points for each category in the binary endpoint is arbitrarily defined. Also, with the emergence of cytostatic treatments, which may not directly cause tumour shrinkage, stable disease can no longer be categorised as a negative outcome in phase II trials as patients with stable disease may have long-term survival benefits. Despite these issues, they found that response rate was a more appropriate endpoint than the continuous endpoints they proposed (total sum of tumour measurements, average sum of tumour measurements, relative change from baseline and average change from baseline), as it better predicted the phase III endpoint, overall survival, and therefore resulted in phase III success.

Sharma and Karrison et al. (71) found that randomised phase II designs with a continuous outcome, specifically, the log ratio of tumour sizes between the two arms, yield better results for predicting phase III trial success compared with randomised designs with progression-free survival and single-arm designs. However, Fridlyand et al. (73) and Kaiser (101) recommend the use of progression-free survival (PFS) as the primary phase II outcome, as they found that it led to an increased probability of correctly terminating the development of a futile treatment. With the increase in cytostatic therapies, Stone et al. (102) also advocate the use of PFS in phase II trials.

An important role of the phase II trial is the fact that it is used to inform decisions in the phase III trial, one of which is the choice of the phase III endpoint. Therefore, the phase II trial endpoints cannot be considered in isolation, rather they have to be investigated with the phase III endpoint in mind, as the choice of endpoint in phase II should be closely related to the outcome used in phase III trials. In order to incorporate the phase III trial endpoint, the investigations carried out in this chapter focus on the correlation between treatment effects on the endpoints used in phase II and III trials, rather than the type of endpoints used in phase II trials. When choosing an endpoint for the phase II trial, there are both clinical and statistical considerations to be taken into account. Clinical considerations include whether the endpoint addresses the trial objectives and can capture the benefit of the treatment, which can be measured in different ways, including how a patient feels, functions or survives (103). The statistical considerations, which is the focus of this chapter, include the distribution of the endpoint and how the phase II endpoint relates to the phase III.

The evaluation of the relationship between endpoints in phase II and III and the validation of surrogate endpoints are closely related but have a different focus (80). Correlation between treatment effects on differing endpoints of trials is evaluated in the validation of surrogate endpoints. A surrogate endpoint may be used to replace a true endpoint in a trial. This can happen for a number of reasons, including when the true endpoint is difficult to measure or requires a long follow-up time (100). However, surrogate endpoints cannot be used to replace a true endpoint unless they have been properly validated. One approach to surrogate endpoint validation is to evaluate the trial-level and individual-level surrogacy. Trial-level surrogacy captures “the precision one can achieve in the prediction of a trial-specific treatment effect in the true endpoint from the effect on the surrogate. It is based on a linear regression model built using trial specific treatment effects on the true and surrogate endpoints observed in previous” clinical trials (104). Individual-level surrogacy is defined as correlation between the surrogate and the true endpoint (80), on the basis of individual patient data. Values close to one for both these parameters indicate that the surrogate endpoint is valid to replace the true endpoint.

Surrogacy validation depends on these two measures however, the methods used to evaluate the strength of relation between phase II and III endpoints can be based on the trial-level surrogacy, in addition to the treatment effect of a phase II endpoint relative to the treatment effect of the phase III endpoint (80), and does not require establishing perfect surrogacy. In this chapter, I will explore the effect of the relationship between the endpoints used in phase II and III trials on the efficiency of phase II trials. This chapter will also highlight what the extent of the impact of the relationship between the endpoints in phase II and III trials and whether there are any other parameters, such as (but not limited to) the variance in the effect of the phase II and III trial endpoints, that can impact the choice of the phase II trial endpoint.

Since only the trial-level surrogacy is required to evaluate phase II endpoints, elements of the surrogacy methodologies are employed. Burzykowski et al. (100) used these methods and present a hierarchical model to capture the relationship between different endpoint types. In this chapter, the hierarchical model is employed, not to validate a surrogate endpoint, but rather to explore the impact of the strength of relationship between the endpoints used in phase II and III trials. This is done within the context of the methods outlined in Chapter 4: using a frequentist approach to design the phase II and III trials, while assuming a distribution for the true treatment effect.

5.2 Methods

Burzykowski et al. (100) indicated that the easiest situation, when evaluating the surrogacy of endpoints, is where the treatment effect of the phase II endpoint, and that of the phase III are realisations of random variables which are normally distributed. In addition, Stallard (39) optimised the sample size of phase II trials under the assumption that they both use continuous normally distributed endpoints. Consequently, here the phase II and III endpoints are assumed to be continuous. Similar to Stallard (39), I assumed an underlying true treatment effect, Δ, for the treatments under evaluation, that follows a normal distribution. This distribution, Δ, represents the effects of all the possible treatments available for investigation and was assumed to have mean μ and standard deviation σ.

Let θ_1 denote the true treatment effect for the surrogate endpoint in the phase II trial and θ_2 denote the true treatment effect for the true endpoint in the phase III trial. Both θ_1 and θ_2 are normally distributed with the same mean Δ and variance τ^2, i.e., $\theta_1 \sim N(\Delta, \tau^2)$ and $\theta_2 \sim N(\Delta, \tau^2)$. This means that θ_1 and θ_2 are independent given Δ, the mean for both treatment effects, which is itself random and normally distributed $\Delta \sim N(\mu, \sigma^2)$, as previously mentioned. In other words, θ_1 and θ_2 are independent but have the same mean, which is random itself. Assuming the same mean for the phase II and phase III effects is vital for the exploration of the correlation between the treatment effects and how the correlation is derived is discussed below. Figure 5.1 shows a visual representation of the assumptions made. Assuming that the mean treatment effects for both phase II and III trials are the same, the joint prior distribution for the treatment effects $f(\theta_1, \theta_2)$, is calculated. Despite the fact that the treatment effect distributions of an endpoint are the same, in phase II and III trials, the actual endpoints used in both phases may be different.

In general, the joint distribution for θ_1 and θ_2 combines information about the distribution of treatment effects and about surrogacy of endpoints; this is because the two design parameters (treatment effect distribution and surrogacy of endpoints) are closely related. In this situation surrogacy has two levels: within treatment which represents the correlation between the endpoints given θ_1 and θ_2, i.e. a patient who has a high value of the surrogate endpoint will have a high value of the true endpoint, as these are correlated. The second level is the between treatment effect which is captured by the correlation between the true treatment effects for the phase II and III endpoints, θ_1 and θ_2, i.e., a treatment which has a large effect on the surrogate endpoint will have a large effect on the true endpoint if the correlation is high. The within treatment level of

surrogacy is also referred to as the individual patient level surrogacy, denoted by R^2_{ind} and the between treatment level is referred to as the trial-level surrogacy, denoted by R^2_{trial}. As previously mentioned, when validating whether an endpoint is a true surrogate of another, both these levels of surrogacy are used. However, here, the objective is to investigate the effect of the relationship between the treatment effect on the phase II and III trial endpoints on phase II trial efficiency, rather than validating a surrogate endpoint in order to replace a true one. In this setting, the relationship between the endpoints is summarised by the correlation between the treatment effect distributions at phase II and III, only, i.e., the trial-level surrogacy, and not at the individual patient level. Therefore, in this research the joint distribution for θ_1 and θ_2 combines information about the distribution of the treatment effects and trial-level surrogacy.

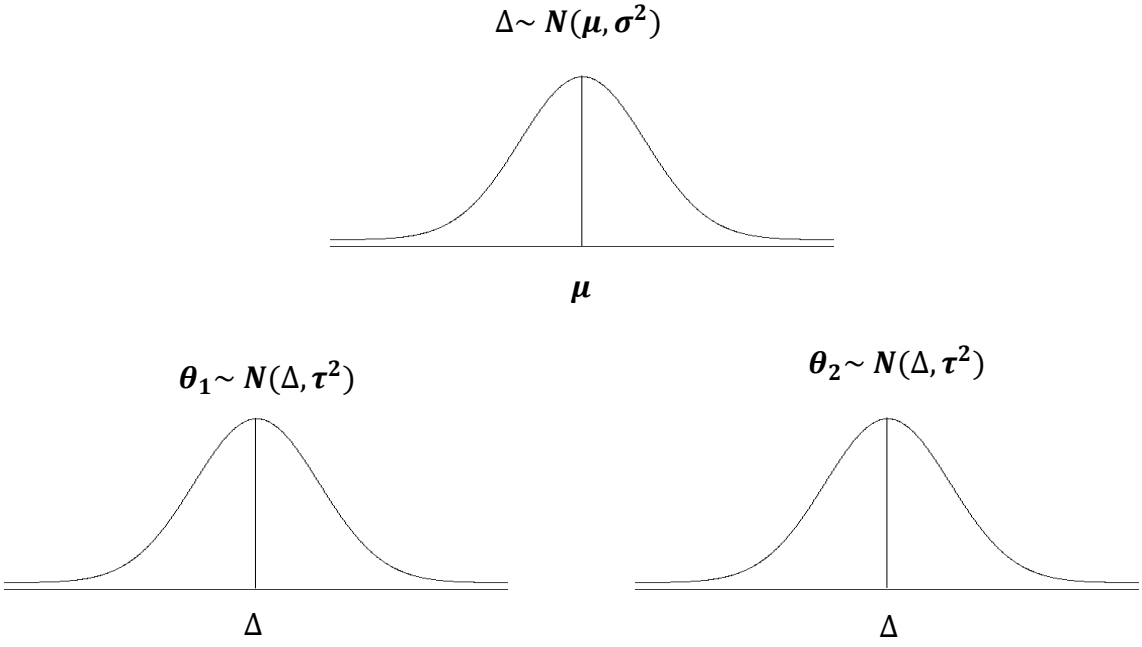


Figure 5.1 The true treatment effect distributions for phase II, θ_1, and phase III, θ_2, both follow a normal distribution with mean Δ, where Δ is the distribution of effects of the available treatments $\Delta \sim \boldsymbol{N(\mu, \sigma^2)}$

The chosen measure used to quantify the phase II trials' efficiency is the expected total number of patients required to lead to a successful phase III trial, as discussed in Chapter 4. Recall, the total expected number of patients required to lead to the first successful phase III trial is given by:

$$\frac{n_1 + n_2 E(P_1)}{E(P_1 P_2)} \tag{5.1}$$

Where n_1 and n_2 denote the total sample size required in the phase II and III trials, respectively. The probability of success in a phase II trial is denoted by $E(P_1)$, and $E(P_1P_2)$ is the probability of success in phase III given the success of the preceding phase II trial.

In this chapter, it is assumed that both the phase II and III trials are two-arm randomised controlled trials with an equal number of patients in both arms, comparing an experimental treatment with a control. Success in a phase II or a phase III trial is defined as the experimental therapy showing statistically significant evidence that it is more promising than the control treatment, i.e., superiority. If the phase II trial is found to be successful, then investigation of the experimental treatment would continue in a phase III trial. A phase III trial is only initiated if the preceding phase II trial is a success; if it failed then the treatment under investigation is dropped and no further investigations are carried out for that treatment, and another phase II is initiated with a different treatment.

The type I and type II error rates for the phase II and III trials were fixed: they were set to $\alpha_1 = 0.05$ (one-sided) in phase II, $\alpha_2 = 0.05$ (two-sided) in phase III, while the type II errors were equal in both phases $\beta_1 = \beta_2 = 0.2$. These were used to calculate the total sample sizes for the phase II and III trials, n_1 and n_2 respectively. The sample size formula used was the standard formula for a two-arm trial comparing a continuous endpoint with an equal number of patients enrolled to both arms.

$$n_i = \frac{4\sigma_i^2\left(z_{1-\alpha_i} + z_{1-\beta_i}\right)^2}{\delta_i^2}; i = 1,2 \qquad (5.2)$$

The targeted treatment difference between the experimental and the control arm denoted by, δ_i was fixed, for both the phase II and III trials. In order to reflect what occurs in a typical phase III clinical trial the targeted treatment effect was set to $\delta_2 = 0.3$; this is based on a systematic review exploring the quality of reporting of phase III trial operating characteristics (98). The most common effect size used for a continuous outcome was 0.3. The targeted treatment effect for phase II trials was also set to $\delta_1 = 0.3$. This was chosen as the same endpoint scale is used in phase II and III trials (in this chapter only). The variance of the underlying true treatment effect for the population of patients in the phase II and III trials was estimated to be $\sigma_1 = \sigma_2 = 1$, so that they represent a standardised effect. Using the sample size Equation 5.2, the total sample size for the phase III trials was calculated to be $n_2 = 348$ patients (note $\frac{\alpha_2}{2}$ is used as the significance level is two-sided); 174 subjects required in

each arm. The total sample size in the phase II trial was calculated to be $n_1 = 274$ patients; this means that a total of 137 patients were in each of the control and experimental arm.

The between treatment correlation, i.e. trial level surrogacy measure is captured by the joint distribution of the true treatment effect distributions for the phase II and III trial endpoints, $f(\theta_1, \theta_2)$. Therefore, the joint distribution was calculated and was used to calculate the probability of success in a phase II trial, $E(P_1)$ and the probability of success in phase II and phase III trials, given that the phase II succeeded, $E(P_1 P_2)$.

5.2.1 Calculating the joint prior treatment effect distribution, $f(\theta_1, \theta_2)$

In order to calculate the probability density function, $f(\theta_1, \theta_2)$, the probability densities of Δ, θ_1 and θ_2 will be multiplied and integrated as follows:

$$f(\theta_1, \theta_2) = \int f(\theta_1, \theta_2 | \theta)\, f(\theta)\, \mathrm{d}\theta$$

$$f(\theta_1, \theta_2) = \int_{-\infty}^{\infty} \frac{1}{2\pi\tau^2}\, e^{-\frac{1}{2}\left[\frac{(\theta_1-\theta)^2-(\theta_2-\theta)^2}{\tau^2}\right]}\, \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left[\frac{(\theta-\mu)^2}{\sigma^2}\right]}\, \mathrm{d}\theta$$

$$f(\theta_1, \theta_2) = \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{\frac{3}{2}}\tau^2\sigma}\, e^{-\frac{1}{2}\left[\frac{(\theta_1-\theta)^2-(\theta_2-\theta)^2}{\tau^2}+\frac{(\theta-\mu)^2}{\sigma^2}\right]} \mathrm{d}\theta$$

From here it can be derived that $f(\theta_1, \theta_2)$ follows a bivariate normal distribution:

$$f(\theta_1, \theta_2) \sim BVN\left(\binom{\mu}{\mu}, \begin{pmatrix} \sigma^2+\tau^2 & \sigma^2 \\ \sigma^2 & \sigma^2+\tau^2 \end{pmatrix}\right) \quad (5.3)$$

Since the trial level surrogacy is captured through the distribution of $f(\theta_1, \theta_2)$, the correlation between the treatment effects was derived from Equation 5.3. Therefore, the correlation between θ_1 and θ_2 is $\frac{\sigma^2}{\sigma^2+\tau^2} (= R^2_{trial})$.

5.2.2 Calculating the probability of success in phase II trials, $E(P_1)$

Equation 5.3 is used to calculate the probability of success in a phase II trial, $E(P_1)$. In order to do this, $\bar{x}_1$ is assumed to be the observed treatment effect in a phase II trial. It is assumed to be normally distributed, $\bar{x}_1 | \theta_1, \theta_2 \sim N\left(\theta_1, \frac{4\sigma_1^2}{n_1}\right)$. Since the phase II trials are analysed using the frequentist approach, success in the phase II trial only occurs if and only if $\bar{x}_1$ is larger than the critical value associated with the phase II trial, k_1. It is concluded that the phase II trial was successful if the observed treatment effect, $\bar{x}_1$ is larger than the critical value; k_1 is dependent on the type I error rate

and the estimated standard deviation associated with the phase II trial population, σ_1, and is given by the following formula:

$$k_1 = z_{1-\alpha_1}\frac{2\sigma_1}{\sqrt{n_1}} \tag{5.4}$$

Therefore, since the distribution of $\bar{x}_1$ only depends on θ_1:

$$P(\bar{X}_1 > k_1) = \int_{k_1}^{\infty}\int_{\infty}^{-\infty} f(\bar{x}_1|\theta_1)\, f(\theta_1)\, d\theta_1 d\bar{x}_1$$

Using the fact that $f(\bar{x}_1|\theta_1) \sim N\left(\theta_1, \frac{4\sigma_1^2}{n_1}\right)$ and $f(\theta_1) \sim N(\mu, \sigma^2+\tau^2)$, the probability of success in a phase II trial becomes:

$$P(\bar{X}_1 > k_1) = \int_{k_1}^{\infty}\int_{\infty}^{-\infty} \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2+\tau^2}} \frac{1}{\sqrt{8\pi\frac{\sigma_1^2}{n_1}}} e^{\left\{-\left[\frac{(\bar{x}_1-\theta_1)^2}{8\frac{\sigma_1^2}{n_1}} + \frac{(\theta_1-\mu)^2}{2(\sigma^2+\tau^2)}\right]\right\}} d\theta_1 d\bar{x}_1$$

$$P(\bar{X}_1 > k_1) = \int_{k_1}^{\infty}\int_{\infty}^{-\infty} \frac{\sqrt{n_1}}{2\sqrt{2}\,\pi\sigma_1\sqrt{\sigma^2+\tau^2}} e^{\left\{-\left[\frac{(\bar{x}_1-\theta_1)^2}{8\frac{\sigma_1^2}{n_1}} + \frac{(\theta_1-\mu)^2}{2(\sigma^2+\tau^2)}\right]\right\}} d\theta_1 d\bar{x}_1$$

Therefore, $\overline{x}_1 \sim N\left(\mu, \frac{4\sigma_1^2}{n_1} + \sigma^2 + \tau^2\right)$

From here, the probability of success in a phase II trial is given as:

$$E(P_1) = P(\bar{X}_1 > k_1) = 1 - \Phi\left(\frac{k_1 - \mu}{\sqrt{\frac{4\sigma_1^2}{n_1} + \sigma^2 + \tau^2}}\right) \tag{5.5}$$

5.2.3 Calculating the probability of success in phase II and III trials, $E(P_1P_2)$

Equation 5.3 was used, again, to calculate the probability of success in phase II and III trials, $E(P_1P_2)$. In order to do this, let $\bar{x}_2$ be the observed treatment effect in a phase III trial. Similar to $\bar{x}_1$, it was assumed to be normally distributed $\bar{x}_2|\,\theta_1, \theta_2 \sim N\left(\theta_2, \frac{4\sigma_2^2}{n_2}\right)$. It should be noted that $\bar{x}_1$ and $\bar{x}_2$ are independent (given θ_1 and θ_2) as the data come from different patients. Since the phase II and III trials are analysed using the frequentist approach, success in the phase III trial can only occur if and only if $\bar{x}_2$ was found to be larger than the critical value associated with the phase III trial, k_2, where

$$k_2 = z_{1-\alpha_2}\frac{2\sigma_2}{\sqrt{n_2}} \tag{5.6}$$

The probability of success in both phase II and III trials is calculated by obtaining the joint distribution of $\bar{x}_1, \bar{x}_2, \theta_1, \theta_2$, shown below:

$f(\bar{x}_1, \bar{x}_2, \theta_1, \theta_2) = f(\theta_1, \theta_2) f(\bar{x}_1|\theta_1) f(\bar{x}_2|\theta_2)$

Using the fact that $f(\theta_1, \theta_2) \sim BVN\left(\begin{pmatrix}\mu\\ \mu\end{pmatrix}, \begin{pmatrix}\sigma^2+\tau^2 & \sigma^2\\ \sigma^2 & \sigma^2+\tau^2\end{pmatrix}\right)$,

$f(\bar{x}_1|\theta_1) \sim N\left(\theta_1, \frac{4\sigma_1^2}{n_1}\right)$ and $f(\bar{x}_2|\theta_2) \sim N\left(\theta_2, \frac{4\sigma_2^2}{n_2}\right)$ gives:

$$f(\theta_1, \theta_2) f(\bar{x}_1|\theta_1) f(\bar{x}_2|\theta_2) = \frac{1}{2\pi(\sigma^2+\tau^2)\sqrt{1-\left(\frac{\sigma^2}{\sigma^2+\tau^2}\right)^2}} \exp\left\{-\frac{1}{2\left(1-\left(\frac{\sigma^2}{\sigma^2+\tau^2}\right)^2\right)}\left[\frac{(\theta_1-\mu)^2}{(\sigma^2+\tau^2)} - \frac{2\frac{\sigma^2}{\sigma^2+\tau^2}(\theta_1-\mu)(\theta_2-\mu)}{\sigma^2+\tau^2} + \frac{(\theta_2-\mu)^2}{\sigma^2+\tau^2}\right]\right\} \times \frac{1}{\sqrt{8\pi}\frac{\sigma_1}{\sqrt{n_1}}} \frac{1}{\sqrt{8\pi}\frac{\sigma_2}{\sqrt{n_2}}} \exp\left\{-\frac{1}{2}\left(\frac{(\bar{x}_1-\theta_1)^2}{\frac{4\sigma_1^2}{n_1}} + \frac{(\bar{x}_2-\theta_2)^2}{\frac{4\sigma_2^2}{n_2}}\right)\right\}$$

The correlation ρ denotes the R^2_{trial} so therefore, let $\rho = \frac{\sigma^2}{\sigma^2+\tau^2}$ and $\gamma = \sigma^2 + \tau^2$

Therefore,

$$f(\theta_1, \theta_2) f(\bar{x}_1|\theta_1) f(\bar{x}_2|\theta_2) = \frac{1}{2\pi\gamma\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(\theta_1-\mu)^2}{\gamma} - \frac{2\rho(\theta_1-\mu)(\theta_2-\mu)}{\gamma} + \frac{(\theta_2-\mu)^2}{\gamma}\right]\right\} \times \frac{1}{2\sqrt{2}\,\pi\frac{\sigma_1}{\sqrt{n_1}}\frac{\sigma_2}{\sqrt{n_2}}} \exp\left\{-\frac{1}{2}\left(\frac{(\bar{x}_1-\theta_1)^2}{\frac{4\sigma_1^2}{n_1}} + \frac{(\bar{x}_2-\theta_2)^2}{\frac{4\sigma_2^2}{n_2}}\right)\right\}$$

$$f(\theta_1, \theta_2) f(\bar{x}_1|\theta_1) f(\bar{x}_2|\theta_2) = \frac{1}{4\sqrt{2}\,\pi\gamma\sqrt{1-\rho^2}\frac{\sigma_1}{\sqrt{n_1}}\frac{\sigma_2}{\sqrt{n_2}}} \exp\left\{-\frac{1}{2}\left[\frac{(\theta_1-\mu)^2 - 2\rho(\theta_1-\mu)(\theta_2-\mu) + (\theta_2-\mu)^2}{\gamma(1-\rho^2)} + \frac{(\bar{x}_1-\theta_1)^2}{\frac{4\sigma_1^2}{n_1}} + \frac{(\bar{x}_2-\theta_2)^2}{\frac{4\sigma_2^2}{n_2}}\right]\right\}$$

It follows that $f(\theta_1, \theta_2) f(\bar{x}_1|\theta_1) f(\bar{x}_2|\theta_2) = f(\bar{x}_1, \bar{x}_2, \theta_1, \theta_2)$ which follows a multivariate normal distribution which in matrix form can be expressed in the following way:

$$\begin{pmatrix}\bar{x}_1\\ \bar{x}_2\\ \theta_1\\ \theta_2\end{pmatrix} \sim MVN\,(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Where $\boldsymbol{\mu}$ is the vector of means for each of the parameters, $\bar{x}_1, \bar{x}_2, \theta_1, \theta_2$, and $\boldsymbol{\Sigma}$ is the variance-covariance matrix.

By comparing coefficients in the exponent term, to the terms in $\begin{pmatrix}\bar{x}_1-M_1\\ \bar{x}_2-M_2\\ \theta_1-M_3\\ \theta_2-M_4\end{pmatrix}^T \begin{pmatrix}Q_{11} & Q_{12} & Q_{13} & Q_{14}\\ Q_{21} & Q_{22} & Q_{23} & Q_{24}\\ Q_{31} & Q_{32} & Q_{33} & Q_{34}\\ Q_{41} & Q_{42} & Q_{43} & Q_{44}\end{pmatrix} \begin{pmatrix}\bar{x}_1-M_1\\ \bar{x}_2-M_2\\ \theta_1-M_3\\ \theta_2-M_4\end{pmatrix}$ (where M_i denote the means

and Q_{ij} denotes the variance-covariance matrix), the means, $\boldsymbol{\mu}$, and variance-covariance matrix, $\boldsymbol{\Sigma}$, was found to be:

$$\begin{pmatrix} M_1 \\ M_2 \\ M_3 \\ M_4 \end{pmatrix} = \begin{pmatrix} \mu \\ \mu \\ \mu \\ \mu \end{pmatrix}$$

And

$$Q = \frac{1}{4(1-\rho^2)\gamma\frac{\sigma_1^2\sigma_2^2}{n_1 n_2}} \begin{pmatrix} (1-\rho^2)\gamma\frac{\sigma_2^2}{n_2} & 0 & -(1-\rho^2)\gamma\frac{\sigma_2^2}{n_2} & 0 \\ 0 & (1-\rho^2)\gamma\frac{\sigma_1^2}{n_1} & 0 & -(1-\rho^2)\gamma\frac{\sigma_1^2}{n_1} \\ -(1-\rho^2)\gamma\frac{\sigma_2^2}{n_2} & 0 & 4\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2} + (1-\rho^2)\gamma\frac{\sigma_2^2}{n_2} & -4\rho\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2} \\ 0 & -(1-\rho^2)\gamma\frac{\sigma_1^2}{n_1} & -4\rho\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2} & 4\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2} + (1-\rho^2)\gamma\frac{\sigma_1^2}{n_1} \end{pmatrix}$$

Calculating the inverse of Q, denoted by S, the variance-covariance yields the joint distribution between $\bar{x}_1, \bar{x}_2, \theta_1, \theta_2$, referred to as Equation 5.7:

$$f(\bar{x}_1, \bar{x}_2, \theta_1, \theta_2) \sim MVN\left(\begin{pmatrix} \mu \\ \mu \\ \mu \\ \mu \end{pmatrix}, \begin{pmatrix} \frac{4\sigma_1^2}{n_1} + \sigma^2 + \tau^2 & \sigma^2 & \sigma^2 + \tau^2 & \sigma^2 \\ \sigma^2 & \frac{4\sigma_2^2}{n_2} + \sigma^2 + \tau^2 & \sigma^2 & \sigma^2 + \tau^2 \\ \sigma^2 + \tau^2 & \sigma^2 & \sigma^2 + \tau^2 & \sigma^2 \\ \sigma^2 & \sigma^2 + \tau^2 & \sigma^2 & \sigma^2 + \tau^2 \end{pmatrix}\right)$$

Note that Equation 5.5 could have also been derived from Equation 5.7. The proof to verify that Q and S are the inverse of one another (i.e., $QS = I$) is shown in Appendix A. The marginal distribution for $f(\bar{x}_1, \bar{x}_2)$ used to calculate the conditional probability of success in phase II and III trials, $E(P_1 P_2)$ is thus given by Equation 5.8:

$$f(\bar{x}_1, \bar{x}_2) \sim BVN\left(\begin{pmatrix} \mu \\ \mu \end{pmatrix}, \begin{pmatrix} \frac{4\sigma_1^2}{n_1} + \sigma^2 + \tau^2 & \sigma^2 \\ \sigma^2 & \frac{4\sigma_2^2}{n_2} + \sigma^2 + \tau^2 \end{pmatrix}\right) \quad (5.8)$$

5.2.4 Analytic Evaluations

Using the algebraic solutions described, the probabilities of success in phase II trials, $E(P_1)$, and conditional success in phase II and III trials, $E(P_1 P_2)$, were obtained. These values were then substituted into Equation 5.1 in order to obtain the expected number of patients required to lead to a successful phase III trial. Since the design parameter under investigation here is the relationship between the treatment effects on endpoints in phase II and III trials, which is

summarised by $\rho(= R^2_{trial})$, a range of values were investigated for ρ. However, two variables affect the value of ρ: the variance about the true treatment effect distribution, σ, and the variance about the true effect of the treatments on the endpoints used in phase II and III trials, τ. Therefore, a range of values for each of these variables were explored; σ ranged from 0.1 to 10, in increments of 0.3, while τ ranged from 0 to 2 in increments of 0.5. These were selected as they are extreme values of these parameters and therefore allow the evaluations to be exhaustive. The correlation depends on σ, so when $\sigma = 0$ the correlation is $\rho = 0$ and increases as σ increases, approaching 1 as σ approaches infinity. This is due to the fact the $\rho = \frac{\sigma^2}{\sigma^2+\tau^2}$ therefore, when $\tau = 0$ and $\sigma > 0$ the correlation is equal to 1, and the treatment effects are perfectly correlated (e.g., the same endpoint was used in both trial phases and the observed treatment effects were the same). However, when $\tau > 0$, the treatment effects of the endpoints are not perfectly correlated; in fact an increase in τ means the treatment effects of the endpoints are less correlated as τ tends to infinity. Setting $\tau = 0$ is equivalent to having a single endpoint throughout the phase II and III evaluations, in other words, the same endpoint is used in phase II and III trials.

As mentioned in Chapter 4, the mean of the true treatment effect distribution was set to $\mu = 0$. This implies that the treatments available for testing have no effect, on average. The impact of different means for the true treatment effect was explored through a sensitivity analysis. Two situations were considered: the average effect of the treatments available was assumed to be positive, i.e., $\mu = 0.5$, and negative $\mu = -0.5$. Recall also that the targeted treatment effect is the same in phase II and III and set to 0.3. As mentioned above, success in the phase II trial can only occur if and only if the observed mean is larger than some critical value, k_1. The critical value was found from the normal distribution and is dependent on the type I error rate, the standard deviation associated with the observed patients in the phase II trial and the sample size of the phase II trial, as shown in Equation 5.4. Given these values the probability of success in phase II trials, $E(P_1)$ was obtained using the standard function “pnorm” in R (105) (Appendix B shows the R code for the investigations in this chapter).

The probability of success in phase II and III trials, $E(P_1P_2)$, was obtained in a similar way. However, in addition to phase II trial values, phase III trial variables were specified. The type I and II error rates for the observed phase III trial were assumed to be the same as the error rates in phase II, except that a two-sided significance level is used in phase III and one-sided in phase II. Success in the phase III trial only occurred if and only if the observed mean was larger than some critical value, k_2, (Equation 5.6). Given these values for

$k_1, k_2, \mu, \sigma_1, \sigma_2, n_1, n_2, \sigma$ and τ the probability of success in phase II and III trials, $E(P_1P_2)$, was obtained using the standard function "pmvnorm" in R (106) (Appendix B).

Hence with values for $E(P_1)$ and $E(P_1P_2)$, the expected number of patients required to lead to the first successful phase III trial were obtained for a range of ρ values in order to explore the effect of phase II and III endpoints on the efficiency of phase II trials. Table 5.1 provides a summary of the parameter values and definitions.

Table 5.1 parameter definition and values

Parameter	**Nomenclature**	**Value/range**
$\boldsymbol{\mu}$	Mean effect of the treatments available	0
$\boldsymbol{\sigma}$	Standard deviation of effect of treatments available	$0.1 - 10$
$\boldsymbol{\tau}$	Standard deviation of the treatment effects in both phase II and III trials	$0 - 2$
$\boldsymbol{\rho} = \frac{\sigma^2}{\sigma^2 + \tau^2}$	Trial-level surragacy/ R^2_{trial}	$0.02 - 1$
Phase II trials		
$\boldsymbol{k_1 = z_{1-\alpha_1} \frac{2\sigma_1}{\sqrt{n_1}}}$	Phase II critical value	0.2
$\boldsymbol{\sigma_1}$	Standard deviation of phase II population, known	1
$\boldsymbol{n_1}$	Phase II sample size	274
$\boldsymbol{\alpha_1}$	Phase II significance level	One-sided 0.05
$\boldsymbol{\beta_1}$	Phase II type II error	0.2
Phase III trials		

Parameter	Nomenclature	Value/range
$\boldsymbol{k_2 = z_{1-\alpha_2} \frac{2\sigma_2}{\sqrt{n_2}}}$	Phase III critical value	0.21
$\boldsymbol{\sigma_2}$	Standard deviation of phase III population, known	1
$\boldsymbol{n_2}$	Phase III sample size	348
$\boldsymbol{\alpha_2}$	Phase III significance level	Two-sided 0.05
$\boldsymbol{\beta_2}$	Phase III type II error	0.2

5.3 Results

5.3.1 Same phase II and III endpoint, perfect correlation

In the single endpoint scenario, which is the simplest scenario, it is assumed that the endpoint used in the phase II trials is the same as the one used in the phase III trials, i.e., that the treatment effects on the endpoints in phase II and III are the same. In order to investigate this scenario, the correlation, ρ, associated with the true treatment effects of the phase II and III trials ($\theta_1$and θ_2), is equal to 1 ($\rho = 1$). In order to attain a perfect correlation between the endpoints in phase II and III trials, the variances (τ) of the individual phase II and phase III treatment effects, θ_1 and θ_2, about the true treatment effect, Δ, are set to zero. This means that the true treatment effect for the endpoints in phase II and III trials, are perfectly correlated. Hence the question, in this situation, is reduced to the effect of the variance about the true effects of the treatments available on the efficiency of phase II trials, σ.

Recall that the true treatment effect follows a normal distribution with mean $\mu = 0$, implying that on average the treatments available have no effect. In addition, the targeted treatment effects in both phase II and III trials were both set to 0.3. With this in mind, the probabilities of success in phase II trials, and phase II and III were obtained, following the methods described above, and are shown in Figure 5.2. It is clear that as the standard deviation of the true treatment effect distribution, σ, increases the probability of success in phase II trials and the probability of success in phase II and III trials also increases, under the assumption of a true treatment effect following a normal distribution with mean

0. It is also clear that the probability of success tends towards 0.5 as σ increases. This is because the treatment effects are far from zero when the variance is large, but as the mean is zero have equal probability of being a large positive or a large negative, with only one of these leading to a phase II success. Figure 5.2 also shows that the probability of success in both phases is smaller than or equal to the probability of success in phase II trials alone. It is clear that both probabilities of successes increase when σ increases and then begin to plateau when σ is greater than 2.

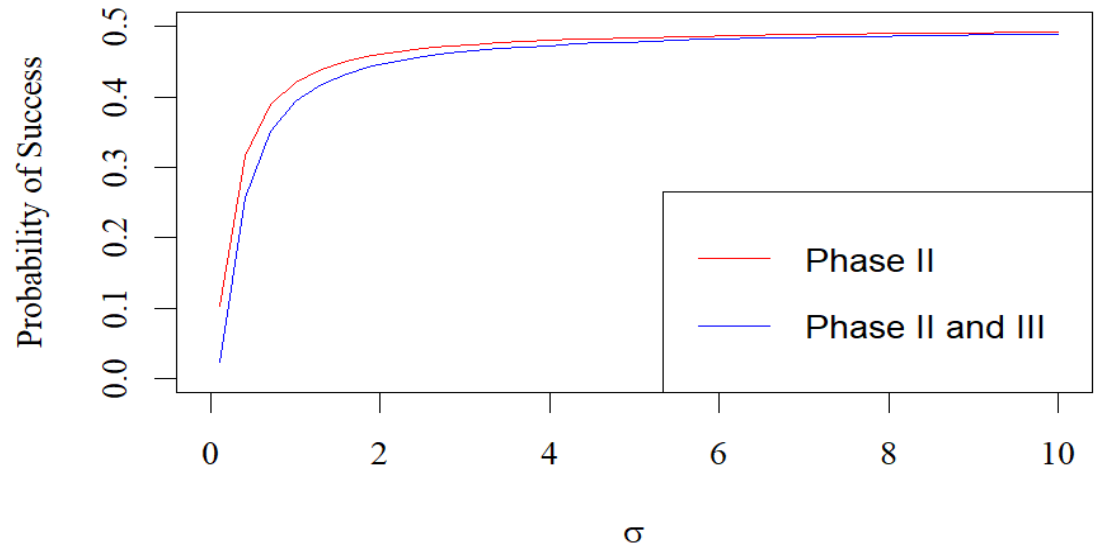


Figure 5.2 Shows the probabilities of success in phase II trials,$E(P_1)$, and in phase II and phase III trials, $E(P_1P_2)$.

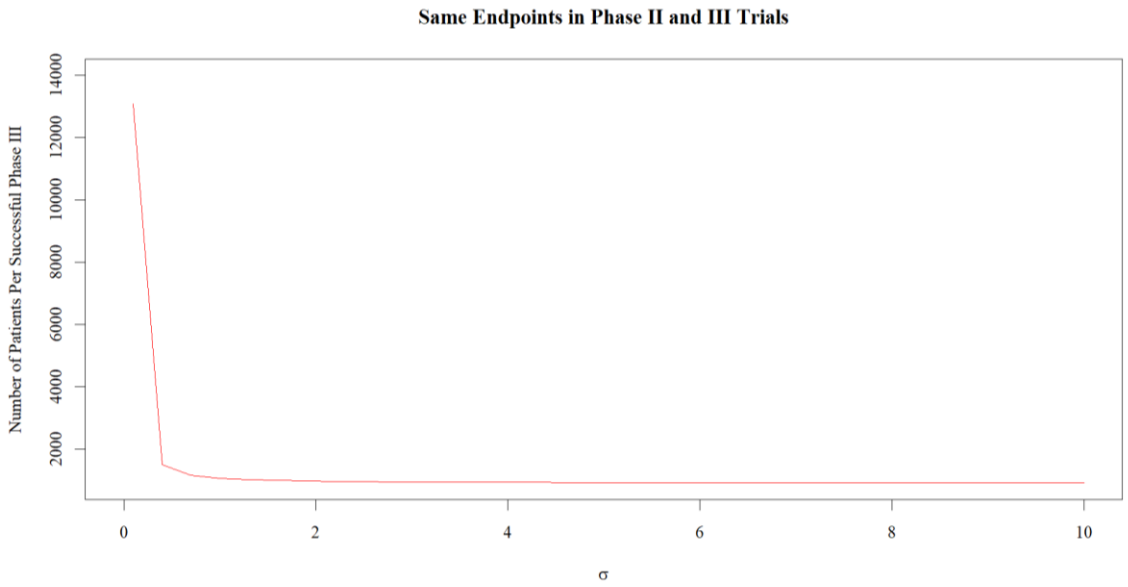


Figure 5.3 The relationship between the variance in the true treatment effect (σ) and the total number of patients required before a successful phase III trial is observed

The probabilities were substituted into Equation 5.1 to obtain the total expected number of patients required until the first successful phase III trial is observed. It is clear from Figure 5.3 that the total expected number of patients required before a successful phase III trial is observed decreases as σ (variance in the treatment effect) increases. When σ goes from 0.1 to 0.4, a drastic decrease in the number of patients required occurs: the number of patients required is almost ten times more at $\sigma = 0.1$ than when $\sigma = 0.4$. It also shows that after approximately $\sigma = 2$ the number of patients starts to plateau. This implies that a large variance (>2) in the true treatment effect distribution does not significantly affect the efficiency of phase II trials, as indicated by the constant level in the number of patients required. With a mean of zero, a larger σ means that some treatments will have a very large treatment effect, leading to the low number of patients required per successful phase III trial. This coincides with the results presented in Figure 5.2 where the probabilities of success also start to plateau at that value of σ. This means that beyond a certain level for the variance about the true treatment effects, it does not play a big role in the efficiency of phase II trials, particularly when the treatment effects of the phase II and III trial endpoints are perfectly correlated or the same.

5.3.2 Different phase II and phase III endpoints

In the different endpoint scenario, which is more realistic, it is assumed that the treatment effects on the endpoints in both the phase II and III trials are not the same but are related. The level of this relationship is captured by $\rho = R^2_{trial}$, and is dependent on the variances of the true treatment effects of the phase II and III trials, τ. The correlation between the treatment effects for the two endpoints, used in phase II and III trials, ranges from approximately $\rho = 0$, implying the two endpoints are not correlated, and $\rho = 1$ suggesting that they have a perfect correlation. Increasing τ from 0.5 to 2, in increments of 0.5, decreases the value of the correlation, ρ.

Using the same values for $k_1, k_2, \sigma_1, \sigma_2, n_1$ and n_2, as in the single endpoint scenario, the effect of the correlation between the endpoints, ρ, and the two variables that affect it, σ and τ, was explored. Figure 5.4 shows the relationship between the variance in the underlying treatment effect (σ) and the total number of patients required before a successful phase III trial is observed for a range of values for τ. Initially the effect of increasing τ, and therefore decreasing the correlation, ρ, appears to increase the number of patients required to lead to the first successful phase III trial, and therefore decreases efficiency. However, it should be noted that when this occurs the variance of the true treatment effect

distribution is small. This implies that the effect of the variance of the true treatment effect is larger than the correlation effect. It is clear that an increase in τ increases the expected number of patients required. This can be explained by the fact that as τ increases, the correlation parameter, ρ decreases, given that σ is fixed. The effect of the τ decreases as σ increases, as the difference between the lines in Figure 5.4 reduce. Similar to the single endpoint scenario, when $\tau = 0.5, 1, 1.5$ and 2, the effect of σ wanes off when $\sigma > 2$, as indicated by the fact that the number of patients required plateaus. This implies that as we increase the variance to a certain level, the total number of patients required in the phase II and III trials decrease, however it no longer has an effect after a certain point. Note Appendix C provides the graphs for the probabilities of success in phase II and phase II and III trials, used to obtain the total number of patients required per successful phase III trial, for when $\tau = 0.5, 1, 1.5$ and 2.

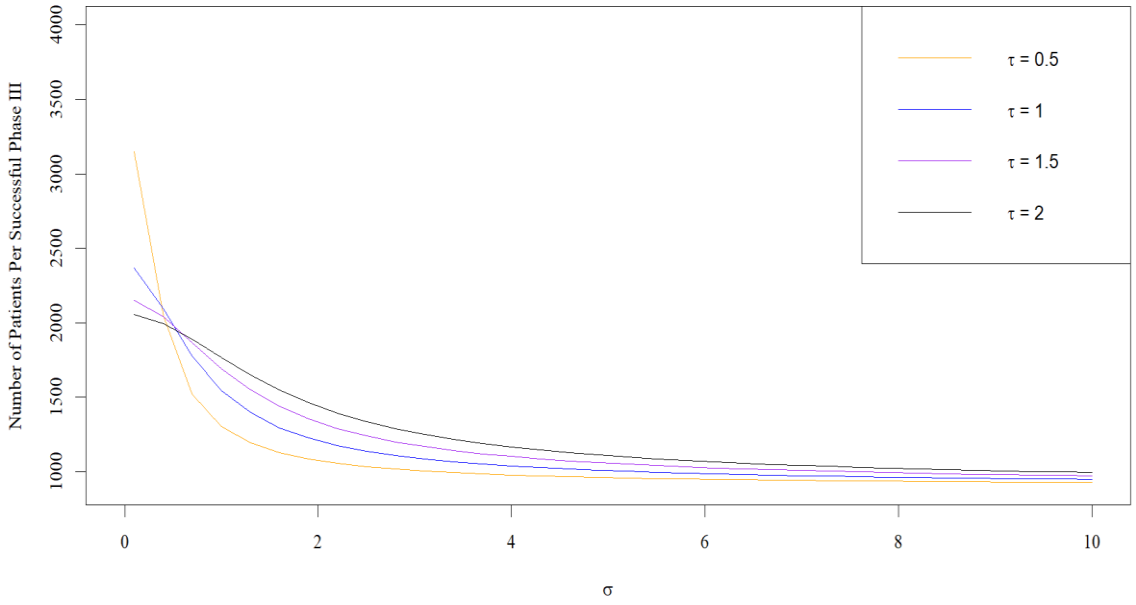


Figure 5.4 Shows the relationship between the variance in the true treatment effect (σ) and the total number of patients required before a successful phase III trial is observed for different values of τ

The extent of the impact of ρ on the efficiency of phase II trials is explored and is shown in Figure 5.5. It shows the relationship between correlation, ρ and the total number of patients required before a phase III success is observed, was drawn. When $\tau = 0.5, 1, 1.5$ and 2, it is clear that the expected number of patients per successful phase III trial decreases as the correlation increases; the degree of decrease in the expected number of patients decreases as τ increases. The larger the value of τ the smaller the difference in the number of patients required to lead to the first successful phase III trial, and as the correlation increases the difference between the efficiency of phase II trials

becomes very minimal. This implies that the correlation has a larger influence than the variance of the treatment effect on the phase II and III trial endpoint, except when the variance is very small, so that all available treatment effects are close to 0 and the number of patients per successful phase III trial is very large as shown earlier.

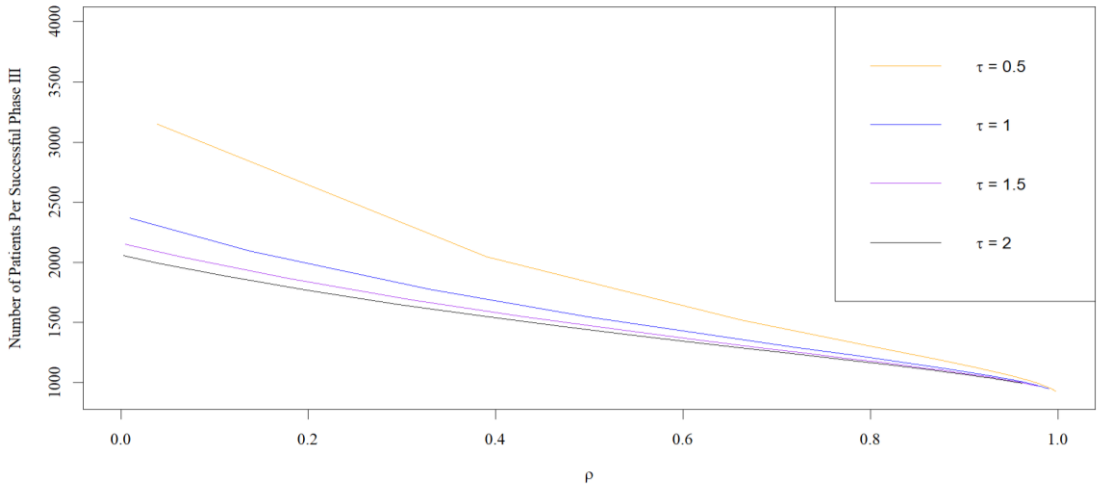


Figure 5.5 Shows the relationship between the correlation between the true treatment effects θ_1 and θ_2 and the total number of patients required before a successful phase III trial is observed

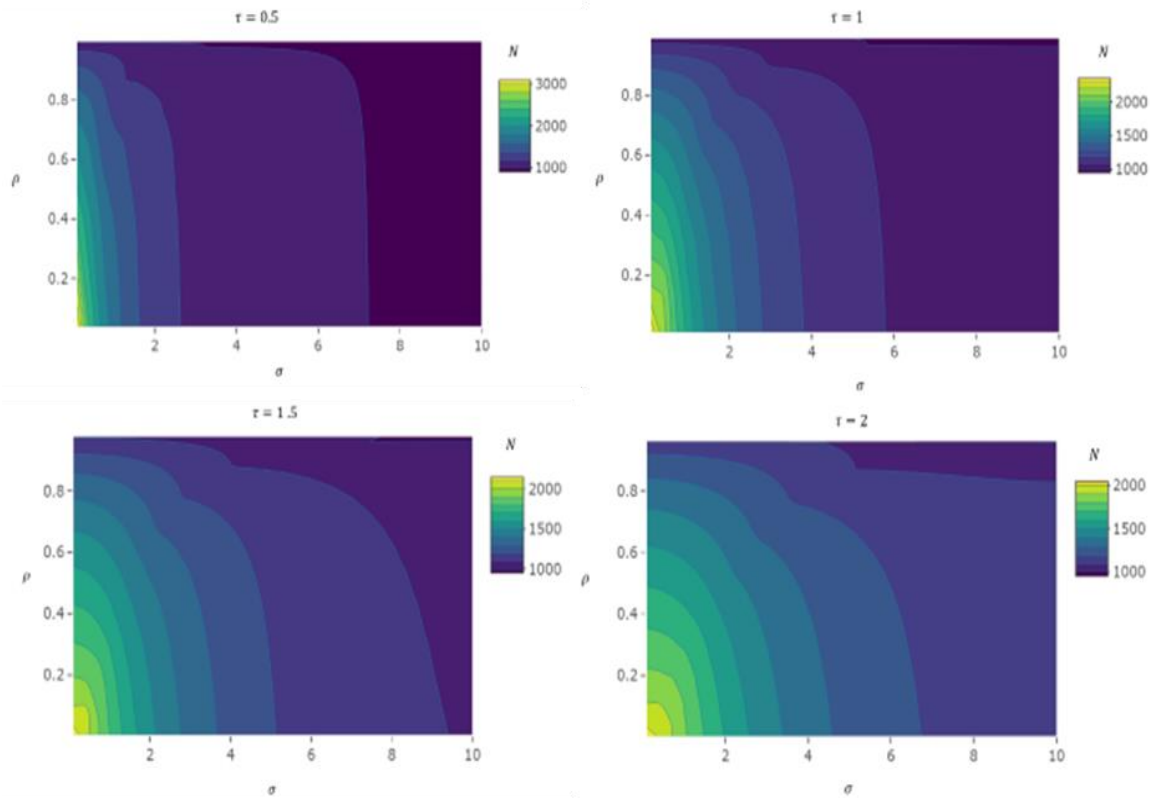


Figure 5.6 Contour plots showing the relationship between ρ and σ and their effect on the expected number of patients required to lead to a successful phase III trial (N)

Combining the information in Figure 5.4 and Figure 5.5, reveals that a high variance in the treatment effect in phase II and III trials, τ, yields more efficient phase II trials when the variance of the true treatment effect, σ, is small. Conversely, when σ is large (between 0.4 and 2) and τ is small the efficiency of phase II trials increases. This is further emphasised in the contour plots shown in Figure 5.6 where the panels show the plots for $\tau = 0.5,1,1.5$ and 2; as τ increase the efficiency of phase II trials decreases. Additionally, it's clear that increasing ρ increases the efficiency, and increasing σ also increases efficiency. However, the contour plots confirm that the effect of σ is larger than the effect of ρ.

5.3.3 Sensitivity analysis

The robustness of the results made was assessed via a sensitivity analysis by changing the assumption that the mean of the treatment effects available for testing was equal to zero. Two further scenarios were conducted: when the treatment effect mean was positive and equal to $\mu = 0.5$ and when it was negative and set to $\mu = -0.5$. Figure 5.7 shows the results for $\mu = 0.5$.

It is clear that when the mean of the treatment effects available is positive, the number of patients required to lead to the first successful phase III trial is reduced. In the single endpoint scenario (top graph in Figure 5.7), it is clear that as σ increases the number of patients increases and then plateaus at about $\sigma = 2$. This is because when the variability increases about the mean of the treatment effects there is more chance that the treatment being tested is not beneficial. The effect of σ is reduces when the variability becomes very large. When $\tau > 0$ the endpoints in phase II and III are different. As τ increases the number of patients required increases i.e., the phase II trials become less efficient. This is due to the fact that increasing τ decreases ρ. In the different endpoint scenario, an increase in σ increases ρ, therefore the middle and bottom graphs in Figure 5.7 show that increasing σ reduces the number of patients required (which is the opposite to what happens in the single endpoint scenario), due to the fact that ρ increases with σ. Consequently, when the mean of the treatment effects available is positive ($\mu = 0.5$) the correlation is a key contributor to the efficiency of phase II trials.

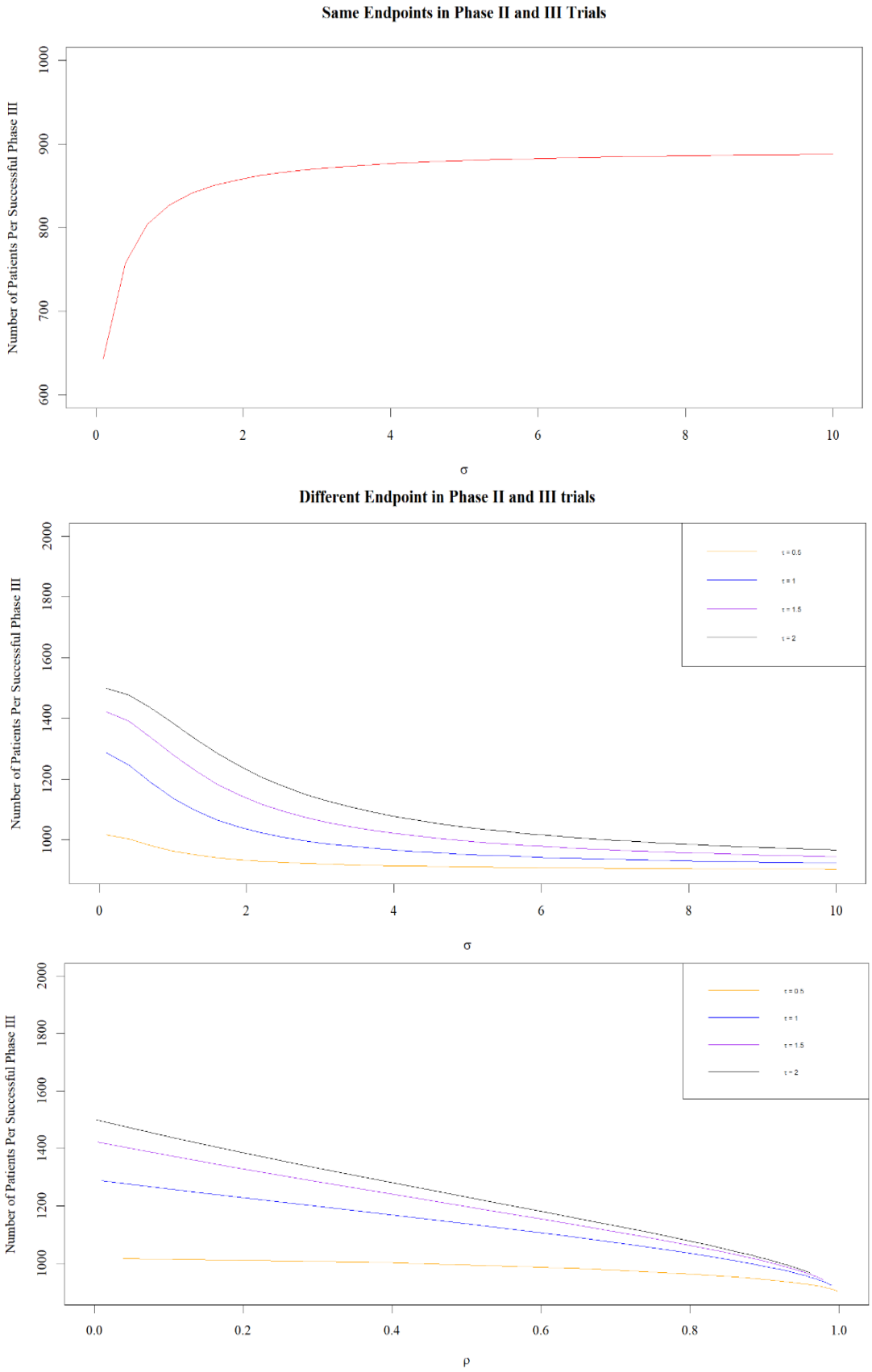


Figure 5.7 The effect of the variability about the treatment effect and the correlation when $\mu = 0.5$ in both the single endpoint scenario (top) and different endpoint scenario (middle & bottom)

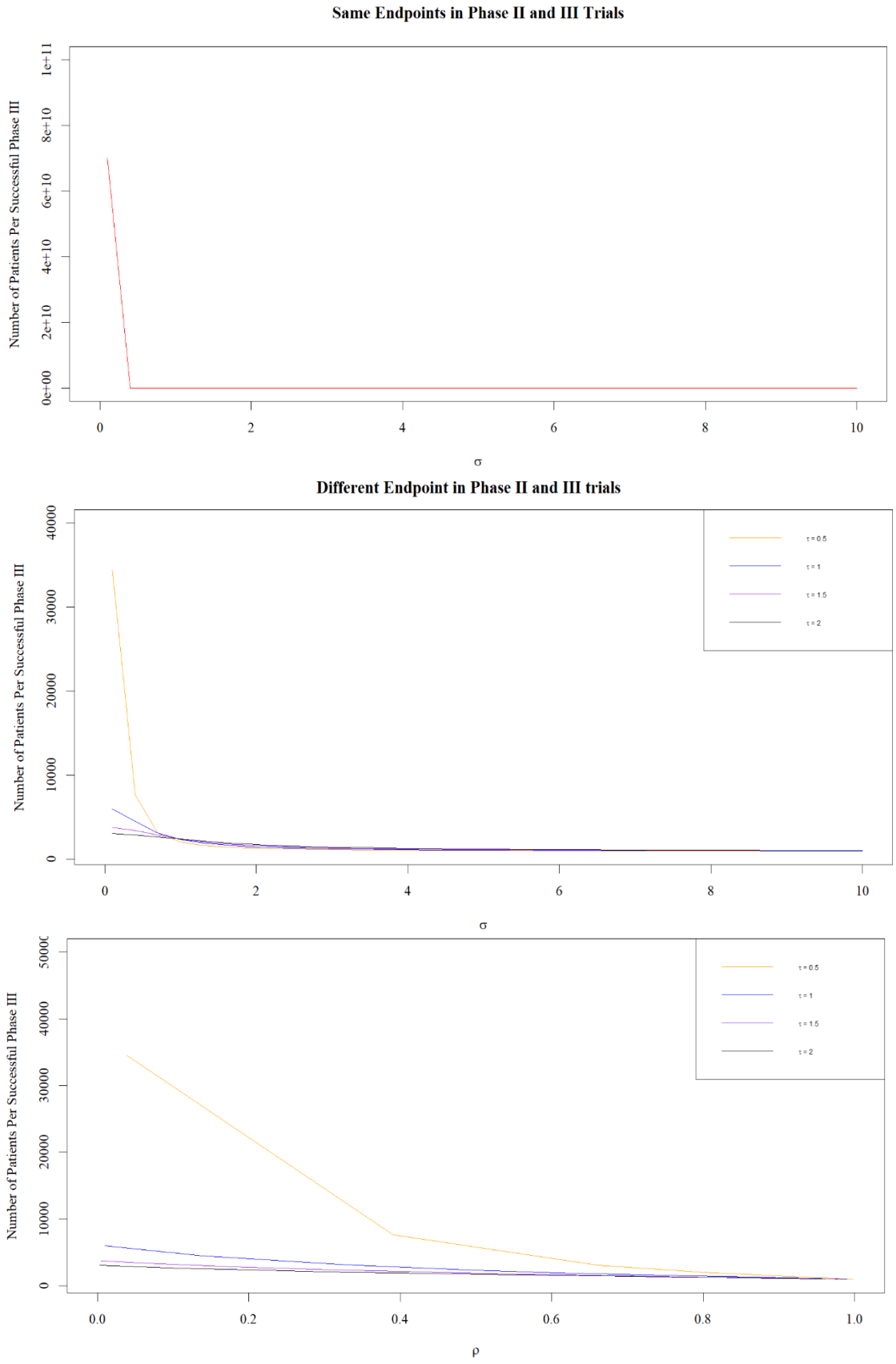


Figure 5.8 The effect of the variability about the treatment effect and the correlation when $\mu = -0.5$ in both the single endpoint scenario (top) and different endpoint scenario (middle & bottom)

Figure 5.8 shows the results for the scenario where the treatment effect of the available treatments is negative, set at $\mu = -0.5$. In the negative mean scenario, the number of patients required increased dramatically compared to the scenarios where $\mu = 0.5$ and $\mu = 0$. This is particularly prominent when the endpoints in the phase II and III trial are the same (top graph in Figure 5.8) and when the variability about the mean of the treatment effects is small. This is to be expected since the treatment effects are centred about a non-beneficial treatment, so that the majority of trials run would conclude that the treatment is futile. In the different endpoint scenario, the effect of increasing τ increased the number of patients required to lead to the first successful phase III trial (refer to middle graph of Figure 5.8), despite the fact that ρ decreases when τ increases. This occurs when the variability about the mean of the treatment effects is small. This implies that the effect of the variance of the true treatment effect is larger than the correlation effect. This was the same result that arose when $\mu = 0$. Similarly, increasing σ reduces the number of patients required, however, the effect of σ reduces when $\sigma > 2$, as indicated by the fact that the number of patients required plateaus. The bottom graph of Figure 5.8 clearly shows that the expected number of patients per successful phase III trial decreases as the correlation increases, particularly when τ is small. This implies that the effect of the correlation exists so long as the variance in the treatment effects is sufficiently large.

5.4 Discussion

The use of short-term endpoints in phase II trials that are closely related to the phase III endpoint is a well-established phenomenon. Researchers have proposed many methods to try to find out whether the endpoint used in phase II is correlated with the endpoint used in phase III. Evaluating surrogacy is one such method. When validating whether an endpoint is a surrogate of another, both the correlation at the trial level, R^2_{trial} and the correlation at the individual patient level, R^2_{indiv} are used. If both these values are close to one then it can be concluded that an endpoint is a valid surrogate of a true endpoint. In this research, the effect of the relationship between phase II and III trial endpoints on phase II trial efficiency was explored. The intention of this research was not to validate whether an endpoint is a good surrogate of another, but to measure the effect of different strengths of correlation between the treatment effects on phase II trial success. Therefore, only the correlation between the treatment effects on the endpoints used in phase II and III trials was of interest and so the measure used was R^2_{trial} or ρ. This is due to the fact that it is also important to

note that the focus of this research is to look at phase II and III endpoints where a valid surrogate endpoint does not replace the true endpoint in the phase III trial. It is, therefore, expected that the two endpoints may be strongly correlated but not perfectly correlated.

From the single endpoint scenario, it was assumed that the phase II and III trials were both designed to measure the exact same endpoint. This, of course, does not reflect what usually occurs in oncology trials, as the endpoint of the phase II trial is different to that used in phase III, due to their different objectives. However, in doing so, the variance about the treatment effects was found to have an impact on the efficiency of phase II trials. The effect of the variance is because some treatments have big effects and some small ones, with this more pronounced as the variance increases. As such, it was concluded that the efficiency of phase II trials increases when the variance of the true treatment effects increases. Furthermore, beyond a certain level of the variance, particularly, after it becomes large ($\sigma > 2$), the effect on the efficiency of phase II trials plateaus. This may be expected, given the assumption that the underlying true treatment effect is drawn from a normal distribution with mean 0, therefore the larger the variance, the more likely a true treatment effect greater than the targeted effect of 0.3 is observed.

In the different endpoint setting, the correlation between the treatment effects in phase II and phase III were assumed to range from no correlation to a near-perfect correlation. It was concluded that the efficiency of phase II trials increases when there is a strong correlation between the treatment effects in phase II and III trials. It was also clear that the variance in the underlying treatment effects in phase II and III trials has a large effect on the efficiency of phase II trials. If the variance of the underlying treatment effect is adequately high ($\sigma > 2$) then the phase II trial is likely to be efficient regardless of the correlation between the treatment effects on the phase II and III trials endpoints. This was clear in the number of patients required to lead to the first successful phase III trial: the difference between the number of patients required when $\rho = 0$ and $\rho = 1$ is small so long as σ is large. This shows that the more influential parameter is the value of the variance of the true treatment effect. This may be expected, similarly as observed in the single endpoint scenario, given the underlying assumption of no treatment effect on average.

The conventional choice for the phase II endpoint is response rate (categorised using RECIST) (107). With the emergence of different types of treatments (cytostatic, as opposed to cytotoxic), the call for different endpoints soon grew (35) and much debate exits regarding the most appropriate endpoint to use in

phase II. As previously mentioned, Sharma and Karrison et al. (71) proposed the use of a continuous outcome. While other authors such as, Fridlyand et al. (73) and Kaiser (101) recommend the use of PFS. While these authors considered the choice of endpoint using resampling methods, there is little consideration in the literature to the statistical aspects for this choice. The key aspect that may lead to better efficiency in phase II trials is to ensure that the treatment effects on the phase II and III trial endpoints are correlated. Chen et al. (80) also found that this relationship is important in choosing the phase II outcome. The findings presented here also highlight that the variance in the true effects of the available treatments also plays a role in the efficiency of phase II trials. This, however, can be overcome by ensuring that the treatment effects of the phase II and III trials are correlated.

As noted, the scenario investigated here is where the average treatment effect is zero. This has a significant bearing on the results, therefore, a sensitivity analysis was carried out to explore the impact of the mean of the true treatment effect. Values of the mean were adjusted to $\mu = -0.5$ and $\mu = 0.5$. These represent the average treatments available have a negative effect, so that the standard treatment is better than the experimental, and a positive average effect signifying that the experimental treatment is better than the control. In the negative scenario the results and conclusions are the same as the zero average effect scenario: the variance of the true treatment effect has an overriding impact on the efficiency of phase II trials over the correlation, particularly when it is small. However, in the positive average effect scenario the correlation is more important than the variance, in terms of phase II trial efficiency. In all scenarios, it can be concluded that the stronger the correlation between the phase II and III trial treatment effects results in more efficient phase II trials, suggesting that the conclusions are robust to the mean value of the treatment effect distribution.

Formal investigations into the effect of the average treatment effect μ were outside the scope of this research, as the focus is to guide the design choices a trialist can make, hence the parameters investigated here are limited to this. The mean of the true treatment effect can only be estimated by historical data on similar treatments, for example collating estimates in a meta-analysis. However, with such high attrition rates in oncology phase III trials (46) (as discussed in Chapter 1) an average of no effect can be deemed appropriate as it reflects the current treatment availability in cancer. As developments are made though, this trend may change. In such cases the treatment effect distribution can be modelled using a more flexible distribution, such as a beta

distribution, which can give rise to different distribution shapes, instead of the normal distribution, assumed here. The fact that the true treatment effect was assumed to follow a distribution at all was based on the work done by Stallard (39). This was deemed a realistic scenario for the treatments available, particularly in the current era of cancer drug development, where a plethora of new treatments are available for testing (96). Further research may investigate the effect of the true treatment effect distribution as a parameter that affects phase II trial efficiency, and more broadly, can focus on parameters that are outside a trialists control, to investigate their impact.

The endpoints of phase II and III trials that were assumed in this chapter were limited to continuous endpoints which follow a normal distribution. As highlighted earlier, this is the simplest scenario and perhaps does not reflect reality in most cases. This scenario was chosen to allow for analytical evaluation to be conducted. Incorporating binary outcomes in phase II and continuous outcomes in phase III would need intense computations of mathematical models. Therefore, the investigations of this chapter are used as the foundations of future chapters, where the phase II trial endpoint is different from that of the phase III endpoint. An example of a type of continuous endpoint that may be used in oncology phase II and III trials, is the tumour size after a certain follow-up period. However, in many phase II and III trials in cancer response rate and overall survival are usually used, respectively. Both the test statistics on these endpoints are asymptotically normal (108). This means that the use of the normal distribution for the treatment effect in phase II and III trial endpoints is valid.

In this research specific endpoints relevant to cancer were not considered, as only the relationship between treatment effect of the endpoints in phase II and III trials was modelled. Specific endpoint types were not investigated in this research, however, previous research has already been dedicated to investigating the appropriateness of specific endpoints (34, 71-74, 101, 109) and these were discussed in detail in the Literature Review (Chapter 3). However, more research is required in order to understand the effect of an appropriate endpoint for phase II trials, and as indicated in this research, ensuring it is correlated with the endpoint used in phase III trials; this should, therefore, be part of the endpoint verification process.

In conclusion, what has been highlighted in this chapter is that when designing a phase II trial and considering a suitable endpoint, there is a trade-off between the correlation between the treatment effects and the variance of the true treatment effect. It is therefore useful to know the effects each of these

parameters has on the efficiency of phase II trials, particularly when faced with a situation where a choice between different phase II endpoints is required, where one is more closely correlated with the phase III endpoint but has a higher variance than another. Historical data can be used to gather information about the correlation of the treatment effects on the endpoints being considered for the phase II trial and the subsequent phase III trial; it is clear that the impact of these parameters play a vital role in the efficiency of phase II trials and should be added into their design process.

Chapter 6 Investigating the impact of different phase II trial designs

From the systematic review (Chapter 2), phase II trials are designed using a variety of methods, including specific designs such as Simon's single-arm two-stage design, randomised designs and single-arm single-stage designs. The purpose of this chapter is to evaluate the effects of different designs on the efficiency of phase II trials, where the efficiency of phase II trials is defined as the number of patients required to lead to the first successful phase III trial, as found in the literature review (Chapter 3). The designs under investigation in this chapter include those that were found in the systematic review, namely randomised single-stage, A'hern's single-arm single-stage (65) and Simon's single-arm two-stage design (63). In addition, the efficiency of Jung's randomised two-stage design (93) is also evaluated (despite not appearing in the systematic review), as it is a randomised version of Simon's single-arm two-stage design (63). Evaluating the efficiency of these designs allows trialists to make informed decisions when setting up a new phase II trial.

Throughout this chapter, I use the terms: single-stage single-arm trials to refer to A'hern's design (65), single-arm two-stage when discussing Simon's design (63) and randomised two-stage design to refer to Jung's design (93).

6.1 Introduction

When designing phase II trials there are many elements that need to be considered. Brown et al. (22) conducted a systematic review which aimed to provide a summary of the elements applicable to cancer phase II trials. These elements are important to consider in order to extract the maximum benefit from the phase II trial. One of the considerations they highlighted was the statistical design of the phase II trial, including whether the phase II trial should incorporate randomisation and whether the trial should be a single-stage, two-stage or multi-stage design. These parameters, among others mentioned in Brown et al. (22), can impact the outcome of phase II trials, and consequently the success rates of phase III trials. Hence, the aim of this chapter is to quantify the effect of the design of phase II trials in terms of their ability to successfully screen new treatments.

In the past, phase II trials were typically conducted as a single-arm design, where patients are enrolled to receive the novel therapy (12). Outcomes are compared to fixed historical control data from recent studies with standard of

care treatment, which is incorporated in the hypotheses. A test is conducted at a pre-specified level of significance to conclude whether the new therapy is statistically better than the historical control data. The level of significance is also known as the type I error rate (α) and it represents the probability of recommending an ineffective treatment for further evaluation. The type II error (β) is defined as the probability of incorrectly rejecting an effective treatment. The power ($1 - \beta$) of the trial is the complement of the type II error. Both the type I and II errors are required in designing phase II trials using a frequentist approach.

Despite the single-arm design's simplicity, it has obvious disadvantages, including the fact that it compares current data to potentially outdated historical data, rendering the comparison moot. For this reason, randomised designs with concurrent arms are often recommended as they distinguish true effects of the novel treatment. While this advantage of randomisation makes it desirable, it is not always practical due to the fact that it requires two to four times more patients than the single-arm design (31). Sharma, Stadler and Ratain (110) highlight other drawbacks of randomised phase II trials: they state that they require more initial effort in the design and conduct than single-arm designs and therefore require a prolonged amount of time to complete. They also state that randomised designs may be unethical as they subject patients to a potentially harmful treatment: if the experimental treatment is futile then patients are subjected to potential harm and if the standard therapy is futile then a beneficial treatment is withheld from them. However, despite these disadvantages, they do not believe that these are enough reasons for randomised phase II trials to remain the exception, as their advantages outweigh their disadvantages. Sharma, Stadler and Ratain's (110) recommendations come from a practical perspective of incorporating randomisation, however, its effect on efficiency of phase II trials has not been explored in the literature, under the context of running multiple trials over a long period of time.

Given the rise of cancer therapies with a cytostatic mechanism of action, and the lack of historical data available for these types of treatments, the need for randomisation in phase II trials is amplified. Despite this, the systematic review in Chapter 2 revealed that both single-arm and randomised designs are equally popular in recent years: of the 128 trials a total of 58 (45.3%) used single-arm designs and 67 (52.3%) trials opted to use randomised designs (the remaining three trials (2.3%) used multiple arms but were not randomised). While these statistics reveal that the use of randomised designs is on the rise, they are still not the gold standard in phase II trials, unlike their use in phase III trials. This implies that no consensus exists for the choice of design in phase II trials.

Hence, it is of profound importance to evaluate the effect of these designs on their ability to successfully screen new treatments. This, in combination with an understanding of the benefits and practicalities of each design, will allow researchers to design phase II trials that truly fulfil their purpose.

As mentioned in Chapter 2, the most popular single-arm two-stage phase II trial design used was that of Simon (63). Given a binary outcome of interest, such as patients' response to treatment, the design is based on testing a null hypothesis that the true response probability is less than or equal to a historical control value, while the alternative hypothesis states that the true response is larger than or equal to the historical control value plus a desirable target level. During the first stage of the trial, $n_{2,1}$ patients are accrued. If the number of responders in the first stage is lower than or equal to some critical value r_1, the trial is stopped for futility, otherwise, the second stage of the trial can commence. During the second stage, $n_{2,2}$ patients are recruited and the total number of patients who responded in both stages is compared to the critical value r. If the total number of responders is larger than r, the treatment is recommended for further study in a phase III trial, otherwise futility is concluded at the end of the trial. Simon's design (63) only allows early termination for futility at the first stage, rather than efficacy.

While there are many two-stage randomised designs proposed, the systematic review (Chapter 2) showed that of the trials included, these were not used in recent phase II trials. However, for completeness and to provide a randomised comparator to Simon's single-arm design (63), I consider, also, the design of Jung (93), who proposed a two-stage randomised design which builds on the structure of Simon's two-stage single-arm design (63). The design is based on testing whether the experimental arm has a higher response rate than the control arm. In the first stage of the trial $n_{2,1}$ patients are accrued in each of the experimental and control arms. The number of responders in each arm is observed; let x_1 denote the number of patients who responded in the experimental arm and y_1 are the responders in the control arm. Success in the first stage of the trial is achieved if the difference between the numbers of responders is larger than or equal to some critical value, $x_1 - y_1 \geq a_1$, otherwise, the trial is terminated, concluding that the experimental treatment is futile.

Upon the success of the first stage of the trial, $n_{2,2}$ patients are accrued into each of the experimental and control arms in the second stage, and the number of responders in this stage are observed. Let x_2 denote the number of patients who responded in the experimental arm and y_2 are the responders in the control arm. The total number of responders in both stages is then calculated

for both arms so that X denotes the total number of responders in the experimental arm and Y denotes the total number of responders in the control arm. Success in the trial is determined by evaluating the difference between the total numbers of responders in both arms and comparing it to some critical value, a. If $X - Y \geq a$ then the experimental treatment is recommended for further study in a phase III trial, otherwise the experimental therapy is deemed futile.

Despite the frequency of use of two-stage designs in phase II trials, their effects on the phase II trial efficiency, in comparison with single-stage designs, are not yet known. Therefore, just as the effects on the phase II trial efficiency of randomised and single-arm phase II trials will be investigated in this chapter, the effect on the phase II trial efficiency of the two-stage designs will also be compared in order to improve the efficiency of phase II trials, in terms of successfully screening new treatments.

6.2 Methods

Each of the phase II designs investigated, namely A'hern's exact design (single-stage single-arm), single-stage randomised design, Simon's optimal design (two-stage single-arm) and Jung's design (two-stage randomised), are assumed to measure a binary outcome, e.g., response, as opposed to the continuous outcome used in Chapter 5. This is chosen to represent realistic design choices in phase II trials in oncology. Despite investigating different designs, each requiring different sample sizes, the operating characteristics are held the same in each design. For all four designs the control rate was assumed to be fixed at $p_1 = 0.25$. This means that 25% of patients in the control arm (for randomised designs) or historical control (for single-arm designs) respond to the standard treatment. For simplicity, the effect of the standard treatment ($p_1 = 0.25$) is fixed for all phase II trials and does not change throughout the evaluations, however, in randomised designs this value is the average response rate in that arm of patients, while in single-arm designs it is assumed to be a fixed value of what has been previously observed, so that it represents how historical control rate is obtained. The four phase II designs were also designed using a one-sided significance level (of those that reported it, this was the most used in phase II trials evaluated in the systematic review (Chapter 2)) of $\alpha_2 = 0.05$ and power $1 - \beta_2 = 0.8$. The clinically significant difference that the phase II trials were designed to detect was set at $\delta_2 = 0.2$, i.e., a 20% difference in the response rate between the two-arms, whether concurrent in the randomised design or historical control in the single-arm design, indicating a response rate of 45% to be clinically significant for the experimental arm. As mentioned in Chapter 4, the

choice of the targeted treatment effect was derived from the systematic review (Chapter 2), where the average effect size was 0.2, for those phase II trials that reported an effect size and had a binary endpoint.

In order to calculate the sample size for each design the appropriate methods were used corresponding to the reference literature. For the randomised single-stage phase II trials with a binary endpoint the following formula was used:

$$n_{2,R} = \frac{2 \times \left(p_1(1 - p_1) + (p_1 + \delta_2)\left(1 - (p_1 + \delta_2)\right)\right)\left(\left(z_{1-\alpha_2} + z_{1-\beta_2}\right)^2\right)}{\delta_2^2} \qquad (6.1)$$

Where $z_{1-\alpha_2}$ and $z_{1-\beta_2}$ are the standard Z scores evaluated at the $(1-\alpha_2)\%$ significance level and 80% power, respectively. As mentioned above, in the randomised single-stage design p_1 $(= 0.25)$ is the expected patients' response rate to the control treatment in phase II trials. The total number of patients required in the randomised single-stage phase II trial with this design was $n_{2,R} = 134$.

The sample size for the single-arm single-stage phase II trial was obtained from the tables presented in A'hern (65). The same operating characteristics were used as above and the sample size for this design was found to be $n_{2,S} = 36$ for a one-sided 5% significance level and with 80% power. In the single-arm single-stage design $p_1 = 0.25$ is patients response rate derived from historical data. Success in phase II trials using this design is concluded when the number of responders is more than or equal to 14 patients.

With the same operating characteristics as the single-stage randomised and single-arm designs, the total and first-stage sample size for the single-arm two-stage phase II trials were obtained following Simon (63). Several designs satisfy the constraints of the type I and II errors (α and β); the minimax design is the design with the smallest total sample size that corresponds to both error constraints and the optimal design is the design that has the minimum expected value of the total sample size in both stages (63). The minimax design was chosen as it is the one with the smallest total sample size and closer in size to the single-arm single-stage design. With these assumptions the total sample size required in the single-arm two-stage phase II trials was $n_{2,ST} = 36$, with $n_{2,ST1} = 17$ patients recruited in the first stage and the remaining $n_{2,ST2} = 19$ recruited in the second stage, if the treatment showed promise, i.e. passed the pre-specified boundary at stage one. Proceeding to the second stage would only occur if $r_1 = 4$ or more patients responded to the treatment in the first stage. The single-arm two-stage phase II trial would be deemed successful if the total number of patients who responded was more than or equal to $r = 13$.

Design	Total sample size and cut-off boundaries		
	Total	First stage	Second stage
Randomised single-stage	134	$n_{2,R} = 134$	-
Single-arm single-stage	36	$n_{2,S} = 36, \quad r = 14$	-
Single-arm two-stage	36	$n_{2,ST1} = 17, \quad r_1 = 4$	$n_{2,ST2} = 19, \quad r = 13$
Randomised two-stage	112	$n_{2,RT1} = 26, \quad a_1 = 2$	$n_{2,RT2} = 86, \quad a = 8$

Table 6.1 Summary of the total sample sizes of the four phase II designs using the operating characteristics $\boldsymbol{\alpha_2 = 0.05}$, $\boldsymbol{1 - \beta_2 = 0.8}$, $\boldsymbol{\delta_2 = 0.2}$, $\boldsymbol{p_1 = 0.25}$

The sample size for the randomised two-stage design was also calculated in a similar way. With the same operating characteristics the sample size required in each arm of the randomised two-stage phase II trials is $n_{2,RT} = 56$ (total of 112), with $n_{2,RT1} = 13$ patients recruited to each arm in the first stage (total of 26) and the remaining $n_{2,RT2} = 43$ recruited to each arm in the second stage (total of 86), if the treatment showed promise. Proceeding to the second stage would only occur if the difference between the number of patients who responded in both arms in the first stage was larger than or equal to $a_1 = 2$. The two-stage randomised phase II trial would be deemed successful if the difference between the total number of patients who responded in both stages was more than or equal to $a = 8$. It should be noted that the cut off boundaries and total sample size of Jung's design (93) are obtained using the exact binomial distribution, unlike the use of chi-squared distribution for the randomised single-stage design, which is why Jung's design results in a smaller sample size. Table 6.1 summarises the sample sizes and cut-off boundaries of each phase II trial design investigated in this chapter.

For the randomised phase II designs, patients were randomly sampled from the true underlying distributions corresponding to the control and experimental arms in the trial. Both of these were binomially distributed with probability of success of p_1 and p_2, respectively. For the randomised single-stage phase II trial the two groups were compared using the chi-squared test. Success in the randomised single-stage phase II trial is obtained by observing a statistically significant difference at the one-sided 5% level between the two groups, in favour of the

novel therapy. While in the randomised two-stage design the superiority of the novel therapy needs to be observed at both stages: the difference between the number of responders between the two groups is compared with the critical values. If the difference between them is larger than the critical value after both stages of the trial, then the phase II trial is deemed successful.

A similar construct was established for the single-arm designs, however, patients were only sampled from the true underlying distribution corresponding to the experimental arm. The response rate of the sample was then compared to the fixed historical control rate. The single-stage phase II trials were deemed successful by the binomial test if the response rate of the novel treatment was found to be statistically significantly better than the historical control rate. However, the success of the two-stage design was established by comparing the number of responders with the critical values; if the number of responders were higher than the calculated critical value at both stages of the trial, the phase II trial was deemed successful.

The value of p_2 used in the binomial distribution for all four designs is not fixed, rather it is related to the true treatment effect distribution for the available treatments, i.e., p_2 represents the underlying true treatment effect in the sample population, rather than the targeted treatment effect in the experimental arm (which is denoted by $p_1 + \delta_2$ in this thesis). Recall from Chapter 4, that the effects of the treatments available were assumed to follow a standard normal distribution, denoted by $\Delta \sim N\ (\mu = 0, \sigma^2 = 1)$. Consequently, the treatment effect that is randomly selected from this distribution, θ, is a mean difference and therefore on the continuous scale. This is used in the phase III trial since it is based on a continuous endpoint (as mentioned in Chapter 4; see Chapter 4 section 4.2.3.1 for more details regarding the phase III design). However, in the phase II trial, where the endpoint is binary, the treatment effect, θ, is transformed to the log-odds scale in order to obtain p_2, the underlying treatment effect in the experimental population, and is derived using the following Equation (6.2) (111).

$$p_2 = \frac{\left(e^{\theta\frac{\pi}{\sqrt{3}}}p_1\right)}{\left(p_1\left(e^{\theta\frac{\pi}{\sqrt{3}}} - 1\right) + 1\right)} \qquad (6.2)$$

Using this transformation, the true treatment effects for the novel therapy in phase II and III trials are equivalent but measured on a different scale. The relationship between the true treatment effects in the experimental arm of the phase II and III trials is depicted in Figure 6.1. The treatment effects on the two endpoints in phase II and III are functionally related, albeit in a non-linear way

(as shown in Figure 6.1). This correlation was included as the findings in Chapter 5 showed that a strong correlation between the treatment effects of the phase II and III endpoints increases the efficiency of phase II trials. Including it in the design evaluations also ensures that the only parameter investigated here is the effect of different designs of phase II trials. In Chapter 5, the treatment effects of the two endpoints were drawn from the same distribution but they cannot be directly derived from one another, thus correlation was not embedded in the model, unlike in this chapter (and in Chapter 7; see Section 7.2).

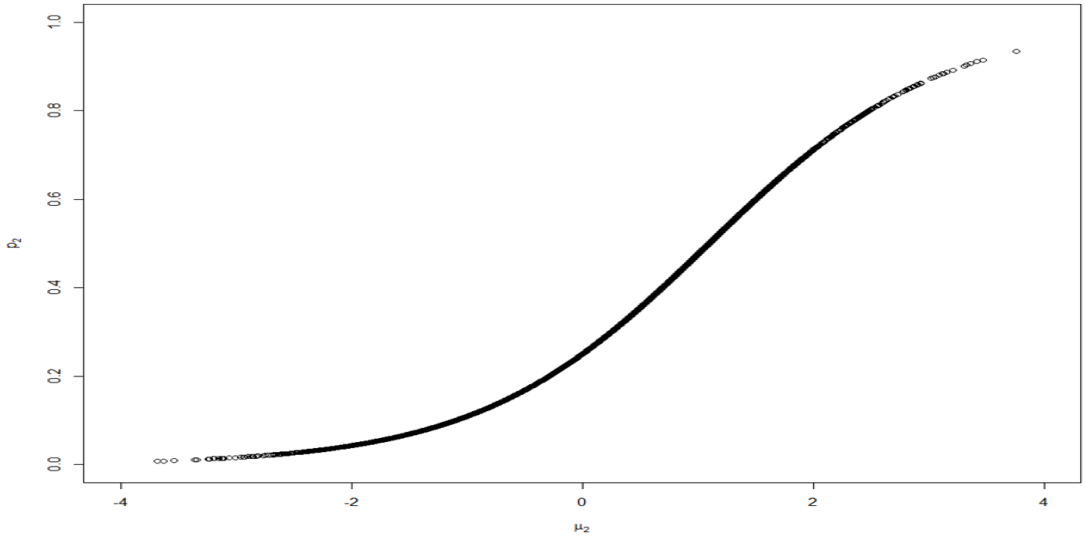


Figure 6.1 The relationship between the true treatment effects of the novel therapy in the phase II and III trials

The investigation of the phase II design with a binary outcome, resulted in any attempt at analytical evaluations to be computationally intensive. Simulations were therefore used to explore the effect of the designs and sample size (Chapter 7) of phase II trials. A large population of size $N = 500{,}000$ patients was set. To replicate the drug development process, patients in the population were sampled without replacement in the phase II and III trial pathway, sequentially, until not enough patients were available for both a phase II and a III trial to be performed. This meant that if a successful phase II trial were found, the pathway continues if enough patients were available to run the proceeding phase III trial, and if an unsuccessful phase II trial was found, the pathway would continue only if enough patients were available for another phase II and III trial.

The simulations were set up such that a treatment with corresponding treatment effect randomly selected from the true treatment effect distribution ($\Delta \sim N(\mu = 0, \sigma^2 = 1)$), is tested in phase II and III trials. A sensitivity analysis was conducted where $\mu = 0.5$, representing a positive treatment effect and $\mu = -0.5$, representing a negative effect. In doing so, the simulations would reveal if the

efficiency of the designs is affected by this assumption. The simulations were also conducted for each of the four designs separately i.e., simulations are conducted consecutively for one design at a time. Having obtained p_2 from the normal distribution, binomial phase II data are then simulated with this p_2. If a phase II trial reveals that the treatment is efficacious the same treatment is investigated in the proceeding phase III trial, i.e., successful phase II trial is followed by a phase III trial with the same novel treatment. However, development of the novel therapy was terminated if the phase II trial was unsuccessful and a new phase II trial is initiated with a different treatment i.e., with a different, randomly selected, treatment effect. Following the methods applied in Chapter 5, it was assumed that the process terminates after the first successful phase III trial is found. However, in order to compare the performance of the phase II trial designs over a long period of time, the measure of efficiency used in the simulations is the number of successful phase III trials, N_{trial}, as opposed to the number of patients required to lead to the first successful phase III trial (used in Chapter 5). This endpoint was suggested by Stallard (39), where he also considers conducting multiple phase II and III trials consecutively and uses the number of patients required to lead to the first successful phase III trial, but also states that over a long period of time these two measures are equivalent. As mentioned in Chapter 4, both these measures are presented in this chapter in order to make the findings throughout this thesis comparable. A simple calculation is conducted in order to obtain the average number of patients required per successful phase III trial: $\left(\frac{N}{N_{trial}}\right)$.

With this set up, the total number of phase II and III trials run was obtained. In addition, the total number of unsuccessful phase II trials were recorded. Of the phase III trials that were run, the total number of failed trials and successful trials were also noted and were used to further understand the impact the designs have on the efficiency of phase II trials. Using these values, the percentage success and failure rates were also obtained for each design investigated in this chapter by dividing the number of successful or failed trials by the total number of trials conducted. These supplement the results by providing a different perspective: while the number of successful phase III trials portrays the efficiency of a phase II design in the long-term (i.e., over a long sequence of trials, provided there is sufficient resources), the success/failure rates show the short-term efficiency of the phase II trial using these designs. For example, in the event that a researcher has resources and funding for one trial only, it would be desirable to design this trial with the highest probability of success. However, it should be noted that the main outcome of interest for this research is the number of successful phase III trials, as I am considering

efficiency from the perspective of the wider drug development pathway (more details in the Literature Review (Chapter 3)). With these assumptions and set up, the effect of the phase II trial design, on their efficiency, was investigated, using simulations in the statistical software R. Appendix D shows the R code employed to obtain the results.

6.3 Results

Table 6.2 shows the median number of phase II and III trials run, along with the number of failed and successful phase II and III trials, for one-stage randomised and single-arm trials and two-stage randomised and single arm trials. Table 6.3 shows percentage success and failure rates for the phase II and III trials run using the four designs.

Interestingly, Table 6.2 show that the single-stage and two-stage single-arm designs, perform in a similar way, however, the two-stage design lead to slightly more successful phase III trials. The design with the highest number of successful phase III trials (1144 (IQR 1140-1148)) is the single-arm two-stage design. The single-stage single-arm design yielded similar results with 1080 (IQR: 1076-1086) successful phase III trials found using this design in phase II. The similarity between the single-stage and two-stage single-arm designs is also reflected in the number of patients required to lead to the first successful phase III trial. On average, the number of patients required to lead to the first successful phase III trial using single-arm two-stage phase II trials is 438, while the single-arm single-stage design yielded 463. Despite the fact that the single-arm single-stage and two-stage designs have similar total sample size, the two-stage requires fewer patients (due to the fact that it can stop early for futility) and yields more successful phase III trials as the two-stage design provides an opportunity to test the efficacy of the treatment at the first stage and allows the trial to terminate if an insufficient number of patients respond to the treatment. It is for this reason that more phase II trials are run using the single-arm two-stage design compared to the single-stage design. Consequently, using the two-stage design means more phase II trials are initiated and as a result more successful phase III trials are yielded. Even though the number of single-stage and two-stage single-arm trials have a difference in the number of phase II and III trials initiated, the difference in the percentage success and failure rates using those designs is not discernible, as seen in Table 6.3.

The phase II design with the lowest number of successful phase III trials (n=685, IQR: (677-691)) is the randomised single-stage design. Comparing the randomised single-stage and two-stage designs, it is clear that the two-stage randomised design is more efficient as it yields 904 (IQR: 902-906) successful

phase III trials. This is also confirmed in the number of patients required per successful phase III trial: more patients are needed when using the single-stage randomised design than Jung's design (730 compared to 554, respectively). Similar to the single-arm trials, the randomised two-stage design allows more trials to be run, compared to the randomised single-stage design, however unlike the single-arm designs the percentage success and failure rates between the randomised single-stage and two-stage designs are quite different. Table 6.3 shows that the two-stage randomised design has a higher phase II trial failure rate than the single-stage randomised design (67.78% in comparison to 61.62%, respectively). In fact, the randomised single-stage design has the lowest phase II failure and the highest phase II success. Therefore, single-stage randomised phase II trials yield the best success rate in phase II trials. Therefore, in the short-term this design is more likely to allow a phase III trial to be initiated, than the other designs investigated.

Overall, it is clear from Table 6.3 that Jung's randomised two-stage design allows more trials to be started but that the success rate is much lower than the other phase II designs, suggesting that Jung's design has a lower power. Consequently, this would result in fewer treatments going through to phase III, so it would be expected that, of the ones that do go through to phase III, they would be successful, therefore resulting in higher success rates in phase III. In other words, it would be expected that the phase III success rate for Jung's randomised two-stage design to be higher than the other designs, as seen in Table 6.3.

The number of failed phase III trials is also of interest as it can provide an indication of efficiency. Similarities between the two-stage and single-stage designs can also be seen here. The designs with the lowest number of phase III trial failures are the single-arm and two-stage randomised design, with a total of just 32 (IQR: 29-35) trials. While the design with the highest number of failed phase III trials is Simon's single-arm two-stage design with a total of 53 (IQR: 49-55) failed trials. The single-stage single-arm design yields a similar number of failed phase III trials as Simon's two-stage design, on average (44 compared with 53, respectively).

Table 6.2 The median number of phase II and III trials run and the median number of successful and failed phase II and III trials for all four designs investigated

One-stage designs	**Median (IQR)**	$\left(\frac{\boldsymbol{N = 500,000}}{\boldsymbol{N_{trial}}}\right)$
Randomised phase II trials run	1868 (1852-1892)	

Randomised phase II trials failed	1151 (1129-1185)	
Randomised phase II trials successful	717 (708-723)	
Phase III trials run	717 (708-723)	
Phase III trials failed	32 (29-35)	
Phase III trials successful, N_{trial}	**685 (677-691)**	**730**
Single-arm phase II trials run	3004 (2968-3028)	
Single-arm phase II trials failed	1879 (1839-1905)	
Single-arm phase II trials successful	1125 (1123-1129)	
Phase III trials run	1125 (1123-1129)	
Phase III trials failed	44 (41-51)	
Phase III trials successful, N_{trial}	**1080 (1076-1086)**	**463**
Two-stage designs		
Randomised phase II trials run	2908 (2885-2978)	
Randomised phase II trials failed	1971 (1946-2050)	
Randomised phase II trials successful	937 (930-938)	
Phase III trials run	937 (930-938)	
Phase III trials failed	32 (29-35)	
Phase III trials successful, N_{trial}	**904 (902-906)**	**554**
Single-arm phase II trials run	3183 (3143-3232)	
Single-arm phase II trials failed	1987 (1944-2040)	
Single-arm phase II trials successful	1196 (1192-1198)	
Phase III trials run	1196 (1192-1198)	
Phase III trials failed	53 (49-55)	
Phase III trials successful, N_{trial}	**1144 (1140-1148)**	**438**

Table 6.3 comparison of the percentage of phase II and III trial successes and failures between the four designs investigated

Design	**% phase II fails**	**% phase II success**	**% phase III fails**	**% phase III successes**
One-stage randomised	61.62	38.38	4.46	95.53

One-stage single-arm	62.55	37.45	3.91	96.00
Two-stage randomised	67.78	32.22	3.42	96.48
Two-stage single-arm	62.43	37.57	4.43	95.65

6.3.1 Sensitivity analysis

A sensitivity analysis was conducted in order to assess the robustness of the assumption that the mean of the treatments available have an effect of zero. Two scenarios were considered: a positive treatment effect where $\mu = 0.5$ (results shown in Table 6.4) and a negative treatment effect where $\mu = -0.5$ (results shown in Table 6.5).

Table 6.4 The median number of phase II and III trials run and the median number of successful and failed phase II and III trials for all four designs investigated when there is a positive treatment effect ($\boldsymbol{\mu = 0.5}$)

One-stage designs	**Median (IQR)**	$\left(\frac{\boldsymbol{N = 500,000}}{\boldsymbol{N_{trial}}}\right)$
Randomised phase II trials run	1512 (1491-1518)	
Randomised phase II trials failed	658 (630-667)	
Randomised phase II trials successful	854 (852-862)	
Phase III trials run	854 (852-862)	
Phase III trials failed	22 (21-25)	
Phase III trials successful, N_{trial}	**832 (829-837)**	**601**
Single-arm phase II trials run	2153 (2128-2158)	
Single-arm phase II trials failed	940 (912-945)	
Single-arm phase II trials successful	1213 (1213-1216)	
Phase III trials run	1213 (1213-1216)	
Phase III trials failed	32 (28-36)	
Phase III trials successful, N_{trial}	**1183 (1178-1186)**	**423**
Two-stage designs		
Randomised phase II trials run	1979 (1958-1988)	
Randomised phase II trials failed	972 (950-982)	
Randomised phase II trials successful	1008 (1004-1010)	
Phase III trials run	1008 (1004-1010)	

Phase III trials failed	20 (18-25)	
Phase III trials successful, N_{trial}	**986 (980-991)**	**508**
Single-arm phase II trials run	2197 (2174-2238)	
Single-arm phase II trials failed	947 (925-992)	
Single-arm phase II trials successful	1248 (1246-1250)	
Phase III trials run	1248 (1246-1250)	
Phase III trials failed	32 (30-34)	
Phase III trials successful, N_{trial}	**1217 (1213-1220)**	**411**

Table 6.5 The median number of phase II and III trials run and the median number of successful and failed phase II and III trials for all four designs investigated when there is a positive treatment effect ($\mu = 0.5$)

One-stage designs	**Median (IQR)**	$\left(\frac{N = 500,000}{N_{trial}}\right)$
Randomised phase II trials run	2401 (2382-2432)	
Randomised phase II trials failed	1889 (1862-1932)	
Randomised phase II trials successful	512 (500-519)	
Phase III trials run	512 (500-519)	
Phase III trials failed	34 (28-40)	
Phase III trials successful, N_{trial}	**478 (468-485)**	**1046**
Single-arm phase II trials run	4491 (4472-4576)	
Single-arm phase II trials failed	3519 (3498-3576)	
Single-arm phase II trials successful	972 (963-974)	
Phase III trials run	972 (963-974)	
Phase III trials failed	61 (58-64)	
Phase III trials successful, N_{trial}	**910 (902-915)**	**550**
Two-stage designs		
Randomised phase II trials run	4686 (4654-4764)	
Randomised phase II trials failed	3876 (3840-3956)	
Randomised phase II trials successful	812 (808-815)	

Phase III trials run	812 (808-815)	
Phase III trials failed	43 (41-46)	
Phase III trials successful, N_{trial}	**769 (763-773)**	**651**
Single-arm phase II trials run	5251 (5220-5291)	
Single-arm phase II trials failed	4158 (4126-4200)	
Single-arm phase II trials successful	1092 (1090-1094)	
Phase III trials run	1092 (1090-1094)	
Phase III trials failed	68 (66-71)	
Phase III trials successful, N_{trial}	**1024 (763-773)**	**489**

It is clear that the mean of the treatment effects available for testing impacts the number of successful phase III trials found and therefore the number of patients required to lead to the first successful phase III trial. When the mean is positive, the number of successful phase III trials increases for all four designs. In contrast when the mean is negative, the number of successful phase III trials decreases for all four designs. However, the conclusions regarding the design with the greatest number of successful phase III trials remained the same in all scenarios investigated, hence the findings in this chapter are robust to the assumptions made about the mean of the treatment effect distribution. Simon's two-stage single-arm design required the least number of patients to lead to the first successful phase III trial and therefore yielded the highest number of successful phase III trials.

6.4 Discussion

An important element of setting up a phase II trial is its design. With so many designs available, it is vital that researchers understand the effect of each of these choices on the efficiency of the phase II trial, and ultimately, the drug development process as a whole. In this chapter, I have investigated the effects of randomised, single-arm, single-stage and two-stage designs. The efficiency of phase II trials was defined as the number of successful phase III trials. This measure is taken from the point of view of running multiple phase II and III trials consecutively, as may be done in large pharmaceutical companies. To supplement the main measure of efficiency, I also incorporated the number of failed phase III trials and the total number of phase II and III trials run, as this reveals the resources required to lead to the successful phase III trials found. To investigate the short-term benefit of each of the designs, I also presented the percentage success and failure rate of the phase II and III trials run.

The findings in this chapter suggest that single-arm designs are more efficient than randomised designs in terms of the number of successful phase III trials; the number of patients required to lead to the first successful phase III trial is also lower in single-arm designs than in randomised designs. This finding can be attributable to the fact that the sample size of the randomised designs investigated was almost three or four times more than that required in the single-arm designs (two-stage and single-stage designs, respectively). It can be concluded that, in the long-term, single-arm designs are more efficient than randomised designs, particularly, if there is a limitation on the number of patients available. The discrepancy in sample size between the randomised and single-arm designs was overcome by supplementing the findings with the percentage success and failure rates of the phase II and III trials. This short-term measure does not incorporate the designs' sample sizes, only the conclusions made using the design. Based on the findings, it was clear that the randomised single-stage design was the most efficient as it had the lowest phase II failure rate and highest phase II success rate.

One of the limitations of this research is the fact that the simulations do not capture the benefits of using a randomised design over a single-arm design, in terms of the variability of the historical control data. If the rate of the control arm success were known to be true then a one-arm trial would be more efficient, but if there is any doubt that it is not reliable then a two-arm trial would lead to more efficiency for the whole drug development process as it would be based on facts rather than speculation. An assumption in the simulations conducted is that the control reference value used in the one-arm trial comparisons is the same as the true control arm success probability used in the two-arm trial simulations (and equivalent to the implied control mean in the phase III simulations), which is the scenario where a two-arm design is not needed. Using a concurrent arm protects against the control group mean being different to that estimated. Since this advantage of randomised designs is not captured in the simulations, in addition to its short-term efficiency presented here, exploring the effect of its sample size is necessary. Another advantage of randomised trials that warrants its further investigation is the fact that randomised phase II trials balance prognostic factors between arms, which allows the efficacy of the treatment to be assessed more reliably (35).

The most efficient design is Simon's two-stage single-arm design as it yielded the highest number of successful phase III trials. However, A'hern's single-arm single-stage design yielded marginally less successful phase III trials, so the difference between the two designs was small, in terms of phase III successes. Despite the fact that Simon's two-stage single-arm design required more

resources than A'hern's single-arm design, namely patients, and therefore may cost more to run, it yielded the highest number of successful phase III trials consequently these costs are more worthwhile. Also, it should be noted, that the two-stage design can stop earlier than a single stage design (for futility), therefore patients and resources can be saved and accrued into more phase II trials that use Simon's two-stage design.

When comparing the randomised trials, namely Jung's two-stage design and the randomised single-stage design, Jung's design was the better option in the long term, as it yielded a considerably higher number of phase III trial successes. In fact, the number of phase III successes yielded using Jung's design in phase II is similar to the single-arm designs, relative to the sample size requirements of each design (almost a quarter of Jung's design). Even though this drastic difference in sample size exists, the efficiency of Jung's design in the long-term is very similar. Despite the fact that there is an increasing trend in the use and recommendations of randomised designs, as suggested in Chapter 2 and in several other literature (31, 96, 110), the results presented show that a two-stage design, rather than a one-stage design is more efficient and therefore the preferred option over the randomised one-stage design. This is more profound given the fact that in the systematic review (Chapter 2), Jung's design was not identified as one of the options that are used. It is therefore concluded that when considering a randomised design, Jung's two-stage design should be prioritised as an option for the phase II design, in the context of running multiple phase II trials.

In terms of the short-term effects of the designs investigated (i.e., an individual phase II trial and its subsequent phase III trial), it was clear that the worst performing design was Jung's two-stage design, as it had the highest rate of failure and the lowest rate of success in phase II trials. Since the treatment effects are sampled from a distribution with an average of no effect present, many treatments do not work. Jung's design stops early for these treatments, giving a high failure rate but meaning that other treatments can be assessed in other trials, which results in a high number of promising treatments going into phase III. The randomised single-stage designs performed the best overall, in terms of the short-term effects. It had the lowest phase II failure rate and highest phase II success rate, as previously mentioned. Given these results, it can be concluded that, in the short-term i.e., running an individual phase II trial, the least efficient design is Jung's two-stage randomised design, while the most efficient design was randomised single-stage design.

Several authors have previously compared the performance of randomised and single-arm designs in phase II. Taylor et al. (70) compared the efficiency of one-

arm phase II trials versus randomised phase II trials. The authors recommend two-arm phase II trials when the uncertainty in the historical control rate is large or when a large number of patients are available. However, when these two conditions are not met, the authors advocate the use of single-arm phase II trials. Pond and Abbasi (83) agree with Taylor et al. (70), as they also recommend the use of either trial design in certain situations. However, unlike Taylor et al. (70), they assumed that the phase II trials were two-stage: single-arm phase II trials were conducted using Simon's optimal design (63), and randomised phase II trials used Jung's design (93). Sambucini (79) used a Bayesian approach to compare single-arm and randomised phase II trials, in terms of their abilities of obtaining the correct decision regarding the new therapy. Similar to Pond and Abbasi (83) and Taylor et al. (70), they also conclude that randomised and single-arm phase II trials are both appropriate in certain situations: when the historical data is correctly estimated single-arm phase II trials are preferred. If this is not the case, a randomised phase II trial is preferred. It is clear from the literature that randomised and single-arm phase II trials have their uses in certain situations and they all point out that randomised designs are better when the historical control rate is unreliable. This research builds on these findings and adds another element to consider when choosing between these designs, namely, whether the phase II trials is conducted individually or in a series of experimentation. I have shown that depending on the context randomised and single-arm phase II trials can both be efficient. In addition, the fact that the efficiency of single-stage and two-stage designs were also investigated further elaborates on the literature, as only randomised and single-arm single-stage or randomised and single-arm two-stage designs were compared rather than assessing all four of these designs.

The designs investigated in this chapter were chosen due to their simplicity and popularity in phase II trials in oncology. However, there are other designs that can also be evaluated and compared in order to gain a deeper understanding of how they affect the efficiency of phase II trials. These designs could include different uses of randomisation, for example the randomised discontinuation trial, which subsets the enrolled patients, by important prognostic factors and randomising only these patients (28). Another design which can also be investigated is multi-stage designs. However, these designs were not included due to their complexity and lack of popularity in oncological trials.

These findings are subject to the design choices made for each of the phase II and III trials. In the phase II trials, the endpoint choice was assumed to be binary. This was chosen as the findings from the systematic review (Chapter 2), showed that binary outcomes are used most frequently in phase II trials. In

addition, the endpoint in phase III, was assumed to be continuous. Although phase III trials in oncology are not typically conducted with a continuous outcome, rather a time-to-event endpoint is used, such as overall survival, the test statistics based on both time-to-event outcomes and continuous ones are asymptotically normal (108). Hence, a continuous endpoint in phase III was deemed appropriate.

The choice of the operating characteristics of the phase II trials was based on the findings of the systematic review in Chapter 2. Most phase II trials are designed with a one-sided significance level, with the majority of trials selecting a value larger than 0.025 and smaller than or equal to 0.1. As such a one-sided significance value of 0.05 was chosen. The most common power chosen in phase II trials was between $90 - 80\%$. Thus, an 80% power was chosen for the phase II trials. It was important to fix the operating characteristics in this chapter as each design has different sample sizes. Chapter 7 explores the effect of different operating characteristics of phase II trials, in the best performing designs found in this chapter, namely Simon's single-arm two-stage, for its long-term efficiency, and randomised single-stage, for its short-term efficiency.

Other assumptions that may influence the results is the choice of the value of the response to the standard therapy. This was assumed to be the same for each phase II trial run regardless of design, i.e., it did not change throughout the simulations. This may not reflect what occurs in reality as oftentimes the information we obtain from one trial feeds into the next trial design. However, this was the same for all the trials investigated, i.e., was kept constant, so did not have any bearing on the conclusions derived from the simulations. The true effect changes between phase II trials as each trial was assumed to assess different treatments. If the same control treatment is assumed to be used in all trials it is reasonable to keep the response rate to the standard treatment fixed. However, future work incorporating changes in the value of the response to the standard therapy would be beneficial.

These findings are also subject to the assumptions made in the simulations. Firstly, the treatments available for testing was assumed to follow a standard normal distribution. This implied that there is an equal number of efficacious treatments as inefficacious treatments. In reality, this may not be the case: there may be an imbalance in efficacy of the available treatments. In order to assess the effect of assuming that the treatments available have no effect on average, a sensitivity analysis was conducted. The average effect of the available treatments was assumed to be either positive (0.5) or negative (-0.5). As expected, all the designs were more efficient (i.e., number of successful phase III trials increased) when there was a positive treatment effect on average.

Similarly, all the designs were less efficient when there was a negative treatment effect, on average. However, in terms of the most efficient design, Simon's two-stage design yielded the most successful phase III trials. Therefore, regardless of the mean of underlying treatment effect the most efficient design was Simon's two-stage design.

Another assumption made in the simulations was using a fixed number of patients available. The value chosen was exceptionally large and was only chosen to provide an end to the simulations. Increasing this value or decreasing it would change the number of trials run, as more or less patients would be required, respectively, however, the conclusions made would be unchanged. Hence, the findings are robust to this assumption.

It is acknowledged that one of the limitations of this research is that the conclusions made are limited to the two-stage and single-stage designs I have chosen to investigate. If a less stringent two-stage design was considered, this may have resulted in more phase III trials and a resulting increase in the number of phase III successes.

Another limitation is that the simulations were conducted one design at a time, therefore there is a potential that the same treatment that is randomly selected may not be the same one for each design, for example there may be more positive treatment effects tested in the randomised phase II trials than in the single-arm designs. However, this is unlikely to drastically affect the results as the average effect for the treatments available was assumed to be the same between the four designs investigated. Therefore, any discrepancies in the treatments tested would not affect the conclusions, particularly that a large number of patients, treatments and trials were assumed and conducted (respectively) for all four simulations.

In conclusion, the most efficient phase II designs are two-stage designs, with Simon's two-stage single-arm design being the most efficient, over a long run of experimentation, where multiple phase II and III trials are initiated consecutively. In the short-term scenarios, where a funding body has the means to run one phase II trial and its proceeding phase III trial, if successful, the randomised single-stage design was found to be efficient. Since Simon's two-stage single-arm and the randomised single-stage design are the most efficient, they are therefore incorporated in the sample size evaluations in Chapter 7.

Chapter 7 Investigating the impact of the sample size of a phase II trial

In Chapter 6, I found that, of the designs investigated, the most efficient design of phase II trials is Simon's two-stage single-arm design (63) as it yielded the most successful number of phase III trials. In addition, I found that randomised single-stage designs have the best success rate in phase II trials. As such, in this chapter, these designs are chosen to investigate the effect of different sample sizes of phase II trials, on their efficiency. The same methods are employed in this chapter as was described in Chapter 6, namely the simulations and the measure of efficiency used to compare the designs.

7.1 Introduction

The problem of the optimal sample size of the phase II trial has been frequently discussed in the literature. Despite this an optimum sample size for phase II trials has not yet been established. The sample size of the phase II trial typically requires a small number of patients, however guidelines about sample size are inconsistent. Taylor et al. (70) suggested that typical single-arm phase II trials have sample sizes between 30-80 patients, while Khan et al. (112) suggested that typical sample sizes range from 40-70 patients in single-arm phase II trials. Furthermore, Gotte et al. (82) recommended larger randomised phase II trials, with a maximum of 100 patients. While the FDA state that, in general, phase II trials can require up to several hundred patients (113). With different sample size recommendations and different designs, each requiring different sample sizes, available for researchers it is important to understand the impact each of these recommendations have on the success of phase II trials. Knowing this will improve the drug development process for both patients and the pharmaceutical industry. This is particularly useful for cancer patients where drastic improvements are required, due to the fact that phase III trials in oncology fail more frequently than other specialties (11). The aim of this chapter is not to determine an optimum sample size of the phase II trials, rather it is to investigate the effect of different sample sizes of phase II trials on their ability to screen new treatments. However, upon fulfilling this aim, recommendations about the size of phase II trials will be suggested.

Assuming the phase II trial will be analysed using a frequentist approach, the sample size is calculated so that the trial has a certain power to detect a clinically significant difference when conducting a hypothesis test at a pre-

specified (one-sided) level of significance. The level of significance is also known as the type I error rate and it represents the probability of recommending an ineffective treatment. The power of the trial is the complement of the type II error, which is defined as the probability of incorrectly discarding an effective treatment. The clinically significant difference is the targeted treatment effect required in order to conclude that the treatment is efficacious. The value of these three parameters influence the size of the trial, thus using different combinations of these values enables the investigation of the effect of different sample sizes of phase II trials on their ability to successfully screen new treatments.

Larger sample sizes of the phase II trial are more likely to detect a treatment effect. However, the objective here is to reveal whether there is a limit to the number of patients required in phase II trials. In other words, in this chapter the aim is to reveal whether increasing the sample size of phase II trials increases their efficiency or is there a point where increasing the sample size does not improve the efficiency of the phase II trial. Discovering this would save a lot of time and money that would otherwise be unnecessarily used. In addition, patients that would have opted into the clinical trial unnecessarily, may well be needed, and indeed, may benefit more, in other trials.

In my investigation of the effect of the sample size of phase II trials, the design of each phase II trial is fixed; they all are assumed to be either Simon's two-stage single-arm (63) phase II trials enrolling patients to take the experimental treatment, or randomised phase II trials with an allocation ratio of 1:1 into a control and an experimental arm. In either case, each phase II trial is assumed to have a binary outcome, namely, whether patients respond to the treatment or not e.g., as assessed by RECIST criteria v1.1 (114) . Differences in the design of the two-stage single-arm and randomised phase II trials only arise by altering the type I and II error rates. The effect of the sample sizes of the phase II trials was investigated assuming that the true treatment effect of all the available treatments followed the standard normal distribution, which included both efficacious and inefficacious treatments, so that the efficiency of the designs can be explored for a variety of treatment effects – moving truly efficacious treatments onto further study and rejecting inefficacious treatments early. Each phase II trial was assumed to test a different treatment whose true treatment effect was randomly selected from this distribution.

As previously described, the measure used to quantify the effect of the phase II trial sample sizes was explained by Stallard (39), where he states that over a long run of experimentation minimising the total number of patients required to

lead to the first successful phase III trial is equivalent to maximising the number of successful phase III trials. Assuming a fixed total number of patients were available, the total number of successful phase III trials was used to quantify the effect of the sample size of phase II trials on their ability to successfully screen new treatments.

7.2 Methods

The methods used in this chapter closely resemble those employed to carry out the design investigations in Chapter 6. As such the design of Simon's two-stage single-arm and randomised designs are the same in this chapter as in Chapter 6, namely, the designs are assumed to measure a binary outcome, with a fixed control rate, $p_1 = 0.25$. This means that 25% of patients in the control arm or historical control respond to the standard treatment. In addition, a clinically significant difference of $\delta_2 = 0.2$ was selected as the targeted treatment effect due to the findings in Chapter 2 (Systematic Review). The only difference in the methods between Chapter 6 and this chapter is that here a range of sample sizes for each design is used.

The sample size for the randomised design was determined using the following formula:

$$n_{2,R} = \frac{2 \times \left(p_1(1 - p_1) + (p_1 + \delta_2)\left(1 - (p_1 + \delta_2)\right)\right)\left(\left(z_{1-\alpha_2} + z_{1-\beta_2}\right)^2\right)}{\delta_2^2} \qquad (7.1)$$

Where $z_{1-\alpha_2}$ and $z_{1-\beta_2}$ are the standard score evaluated at the $(1-\alpha_2)\%$ significance level and $(1-\beta_2)\%$ power, respectively. To explore differing sample sizes the values of type I and II error rates were varied. The values for the type I error rate were set to $\alpha_2 = 0.01, 0.05, 0.1, 0.15, 0.2$, and were assumed to be one-sided, while the type II error was set to $\beta = 0.1 - 0.6$, in increments of 0.05, so that the power ranges from $(1-\beta_2) = 0.4 - 0.9$. The choice for the values of the type I and II error include what is typically chosen in trials (type I error rate 0.01-0.2, type II error rate 0.1- 0.2), as was found in the systematic review (Chapter 2). In addition, the extreme values included in the investigations were based on Stallard's (39) results, where he found that the optimal choice of α_2 and $1-\beta_2$ for the phase II trial were 0.2 and 0.4, respectively. Other extreme values were also included to allow the investigations to be exhaustive.

The sample size of Simon's design was determined using an R package called "clinfun"(115), using the same range of type I error and power values as described above, but with fixed $p_1 = 0.25$ and $p_1 + \delta_2 = 0.45$. This provided the designs, i.e., the total and first-stage sample sizes in addition to the cut-off

responses needed to conclude whether the stage or trial is successful, that satisfy these operating characteristics. Both the minimax and optimal designs are outputted but the minimax design was chosen as it was the design investigated in Chapter 6. The minimax designs that satisfy the operating characteristics and are investigated in this chapter are shown in Appendix E. From Chapter 6, recall that the total sample size for Simon's design is denoted by $n_{2,ST}$, with $n_{2,ST1}$ patients recruited in the first stage and the remaining $n_{2,ST2}$ recruited in the second stage, if the treatment showed promise, i.e., passed the pre-specified boundary. Proceeding to the second stage would only occur if more than r_1 patients responded to the treatment in the first stage. The single-arm two-stage phase II trial would be deemed successful if the total number of patients who responded was more than r.

For the randomised phase II designs, patients were randomly sampled from the true underlying distributions corresponding to the control and experimental arms in the trial. Both of these were binomially distributed with probability of success of $p_1(= 0.25)$ and p_2, respectively, i.e., patients accrued into the randomised design were sampled from a Binomial distribution, $Bin(\frac{n_{2,R}}{2}, p_1$ or $p_2)$. For the randomised single-stage phase II trial the two groups were compared using the chi-squared test. Success in the randomised single-stage phase II trial is obtained by observing a statistically significant difference at the $(1-\alpha_2)\%$ level between the two groups, in favour of the experimental arm.

Unlike p_1, the value of p_2, used in the binomial distribution for both designs, is not fixed, rather it is related to the true treatment effect distribution for the available treatments. Recall, from Chapter 4, that the treatment effects were assumed to follow a standard normal distribution, denoted $\Delta \sim N\ (\mu = 0, \sigma^2 = 1)$. Consequently, the treatment effect that is randomly selected from this distribution, Δ, is a mean difference and therefore on the continuous scale. This is used in the phase III trial since it is assumed that the phase III trial is designed with a continuous primary endpoint (as mentioned in Chapter 4; see Chapter 4 section 4.2.3.1 for more details regarding the phase III design). However, in the phase II trial, where the endpoint is binary, the treatment effect, θ (randomly selected from Δ), is transformed to the log-odds scale in order to obtain p_2 and is derived using the following Equation (7.2) (111).

$$p_2 = \frac{\left(e^{\theta\frac{\pi}{\sqrt{3}}}p_1\right)}{\left(p_1\left(e^{\theta\frac{\pi}{\sqrt{3}}} - 1\right) + 1\right)} \qquad (7.2)$$

This transformation was used to ensure that there is a correlation between the true treatment effects in the experimental arm of the phase II and III trials. This

model was previously used in Chapter 6 (for more details regarding this model see Section 6.2). Refer to Figure 6.2 in Chapter 6 for a depiction of the relationship between the true treatment effects in the experimental arm of the phase II and III trials.

A sensitivity analysis was conducted in order to assess the robustness of the assumption of the mean of the treatments effect of the available treatments. The mean of the distribution was set to 0, implying that the treatments available have no effect, on average. Two additional scenarios were assessed in order to reveal the phase II efficiency: when there is a positive treatment effect ($\mu = 0.5$) and when there is a negative treatment effect ($\mu = -0.5$).

Simulations were used to explore the effect of the sample size of Simon's design and randomised single-stage design. For each combination of α and β investigated, a large population of size $N = 500{,}000$ patients was assumed. To replicate the drug development process, patients in the population were sampled in the phase II and III trial pathway, sequentially, until not enough patients were available for both a phase II and a III trial. This meant that if a successful phase II trial was found, the pathway continues if enough patients were available to run the proceeding phase III trial, and if an unsuccessful phase II trial was found, the pathway would continue only if enough patients were available for another phase II and III trial.

In the randomised design setting the control sample was selected from the binomial distribution as mentioned above, while in the single-arm two-stage design (Simon's design (63)) a historical control rate with the same p_1 value as the randomised design, is used to compare to the patients accrued in the experimental arm. The simulations were set up so the experimental treatment under investigation has a treatment effect that is randomly selected from the true treatment effect distribution to be tested in phase II and III trials. If a phase II trial reveals that the treatment is efficacious the same treatment is investigated in the proceeding phase III trial, i.e., successful phase II trial is followed by a phase III trial with the same novel treatment. However, development of the novel therapy was terminated if the phase II trial was unsuccessful and a new phase II trial is initiated with a different treatment. It was assumed that the process terminates after the first successful phase III trial is found. This process is repeated for each combination of α_2 and $(1 - \beta_2)$.

In order to compare the performance of the phase II trial designs, over a long period of time, the measure of efficiency is the number of successful phase III trials, N_{trial}, as opposed to the number of patients required to lead to the first successful phase III trial (used in Chapter 5). This endpoint was suggested by

Stallard (39), where he also considers multiple testing of phase II and III trials and uses the number of trials required to lead to the first successful phase III trial, but also states that over a long period of time these two measures are equivalent. As mentioned in Chapter 4, both these measures are presented in this chapter in order to make the findings throughout this thesis comparable.

A simple calculation is conducted in order to obtain the number of patients required per successful phase III trial: $\left(\frac{N}{N_{trial}}\right)$. With these assumptions and set up, the efficiency of the phase II trial sample size for randomised and Simon's two-stage single arm design, was investigated, using simulations in the statistical software R. Appendix F shows the R code employed to obtain the results.

7.3 Results

7.3.1 Simon's two-stage single-arm phase II trials

The total sample sizes, n_2 yielded for Simon's two-stage single-arm (63) phase II trial ranged from 5 to 71 patients, for the varying combinations of type I error and power. A total of 54 combinations of α and β were investigated. Figure 7.1 shows the total number of successful phase III trials for the range of phase II sample sizes investigated, while Figure 7.2 is a panel plot of the number of phase III successes for different values of the power $(1-\beta)$. Overall, Figure 7.1 shows that there is a decreasing trend in the number of successful phase III trials yielded as the sample size of the phase II trial increases. Initially, however, the decreasing trend has a much smaller magnitude than when the sample sizes exceed 20 patients. This means that the two-stage phase II design is most efficient when there is a smaller number of patients. The efficiency of phase II trials is at its peak when the phase II trial reaches a total of approximately $5-15$ patients on average, corresponding to α values of $0.05-0.15$ and power values of $0.4-0.6$. Figure 7.1 reveals that increasing the sample size of single-arm phase II trials beyond a total of approximately 20 patients decreases the number of successful phase III trials, for the parameters investigated, specifically when the average effect of the treatments available was assumed to be 0 and with a $p_1 = 0.25$ and $\delta_2 = 0.2$. A decrease in the number of successful phase III trials means that the efficiency of the phase II trial decreases.

It is also clear from Figure 7.1 that the choice of the type I error has a very profound impact on the efficiency of the phase II trial. The designs with a type I error larger than or equal to 0.05 result in the largest number of successful phase III trials. While designs with a stringent type I error of $\alpha = 0.01$ yields a smaller number of successful phase III trials.

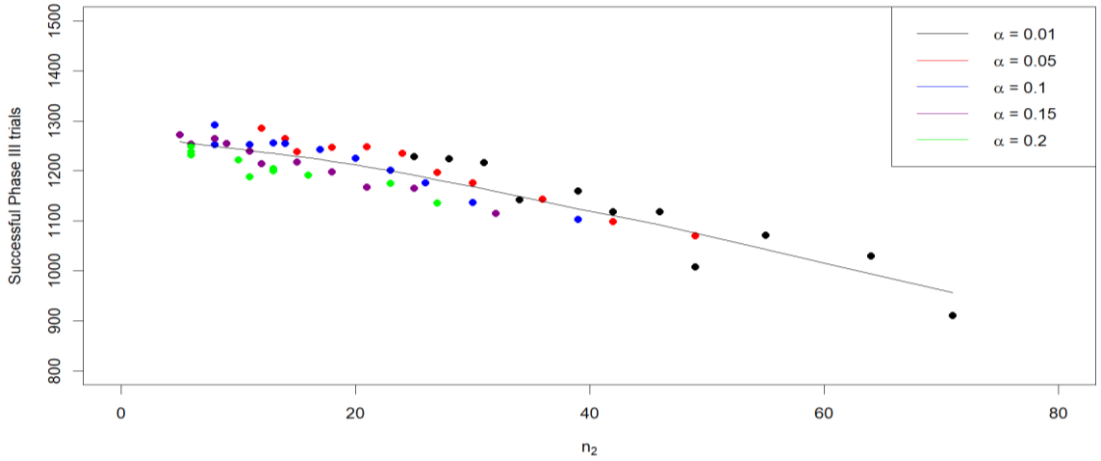


Figure 7.1 The total number of successful phase III trials depending on the number of patients in Simon's two-stage single-arm phase II trials differentiated by the value of α

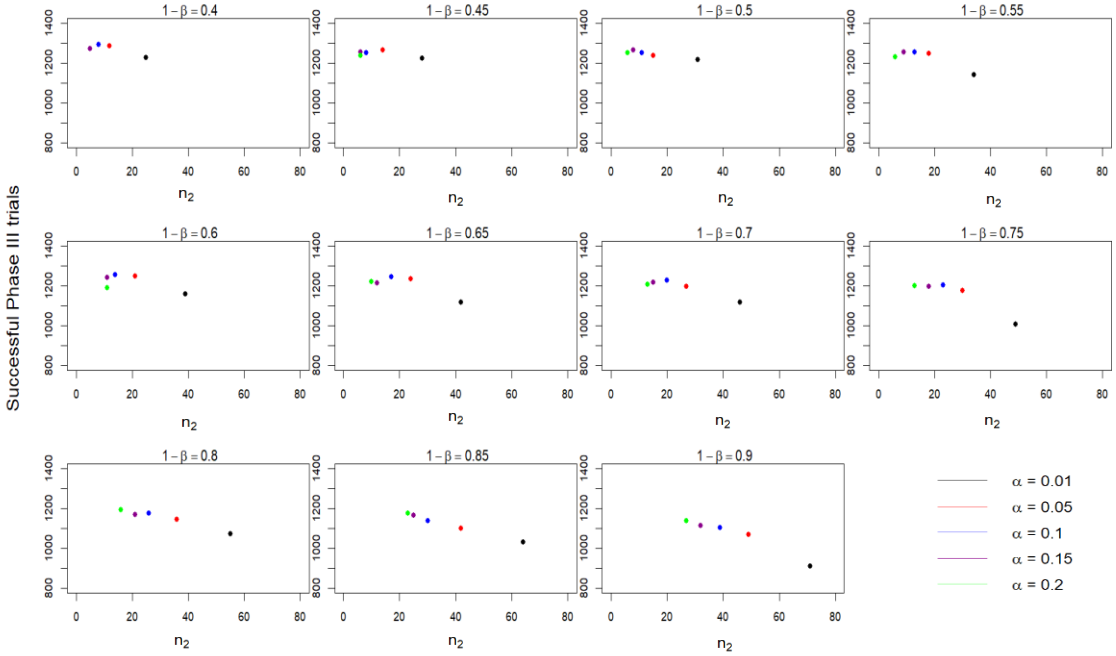


Figure 7.2 The total number of successful phase III trials, shown by the power of each Simon's two-stage single-arm phase II trial

As previously described, each phase II trial design had a range of type II error rates, ranging from $\beta = 0.1$ to 0.6 which correspond to power rates ranging from $(1 - \beta) = 0.4$ to 0.9, in increments of 0.05. Figure 7.2 shows the effect the power of each phase II design has on the number of successful phase III trials. It shows that as the power increases the points on the graphs shift to the right, that larger powers increase the sample size, as expected. Figure 7.2 shows that the small increasing increments in power does not impact the number of successful phase III trials, as the range of the points in each of the graphs, corresponding to the investigated powers, differ very little from each other. However, when comparing the number of phase III successes yielded when a

power of 0.4 and 0.9 is used, the advantage is clearly in favour of using a power of 0.4. Figure 7.1 and Figure 7.2 reveal that the choice of the type I error of the phase II trial is profound so long as the power is high. However, when the power is low the impact of the type I error is not prominent.

In order to further explore the impact of the sample size of the single-arm two-stage phase II trial has on their efficiency, the success rates of the phase II and phase III trials were obtained. The success rate of the phase III trials was calculated using the total number of successful phase III trials divided by the total number of phase III trials run. Similarly, the success rate of the single-arm two-stage phase II trials was obtained using the total number of successful phase II trials and the total number of phase II trials run. Figure 7.3 shows the success rates of the phase II and III trials, with the points corresponding to the designs with different type I error rates. The overall success rate of phase II trials is smaller than the success rate of phase III trials, with the phase II success rate ranging from approximately 0.29 to 0.44, while the phase III success rate ranges from approximately 0.88 to 0.99. It is clear that initially the phase II trial success rate increases as the sample size increases. However, at about 40 patients the success rate of the phase II trials plateaus. The phase III trial success rate increases as the sample size increases. The sample size increases with increasing power, which also increases the probability of success. Decreasing α (which decreases the probability of success) also increases the sample size. So, this is quite a complicated relationship, which also depends on the distribution of the treatment effects (e.g., if all treatments had no effect (at the null) then the probability of success would depend only on (in fact, equal) α, whereas if all were at the specified target value the probability of success would depend only on (equal) the power. The fact that the probability of success in phase III is higher than in phase II shows that phase II is leading to poor treatments being dropped, with this happening more when the sample size is larger.

Furthermore, it is clear from Figure 7.3 that the choice of the phase II type I error has a major effect on the success rates of the phase II and III trials. Stringent type I error rates ($\alpha \leq 0.05$) lead to higher success rates in phase III trials. Comparatively, phase II trials with less stringent type I error ($\alpha > 0.05$) have a lower success rate at phase III, but a high success rate at phase II. This implies that over a long run of experimentation stringent α's may not be the most efficient choice (as found when looking at the number of successful phase III trials), but in an individual phase II that leads to one phase III trial, a stringent α would be more appropriate, so that the success rate is optimised.

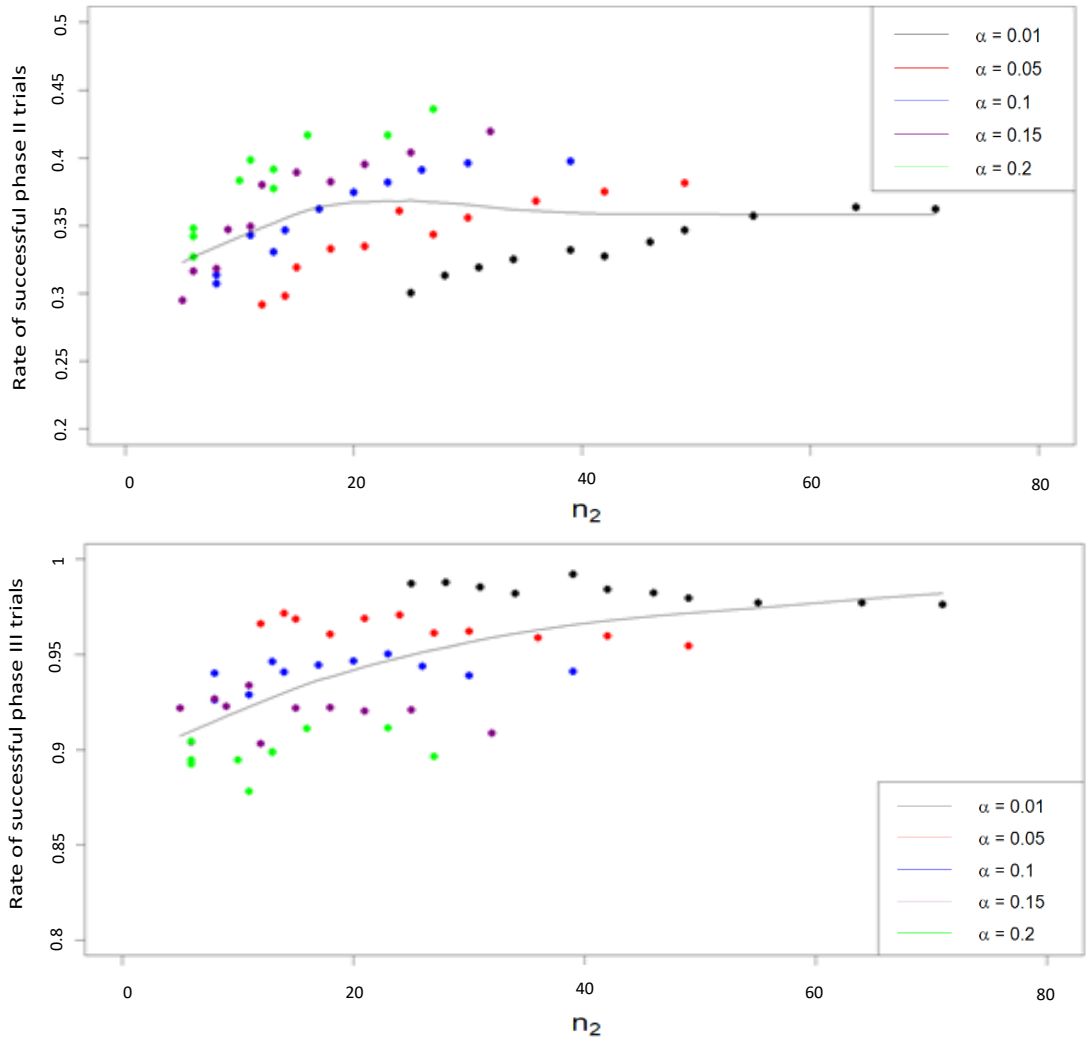


Figure 7.3 The success rate of the phase II trials (top); the success rate of the phase III trials (bottom)

7.3.1.1 Sensitivity analysis for Simon's two-stage design

As mentioned in Section 7.2, a sensitivity analysis was conducted to verify the robustness of the findings under the scenarios where there is a positive treatment effect and a negative effect, on average. This meant that the value of μ for the distribution of the treatments available was set to $\Delta \sim N(\mu = 0.5, 1)$ and $\Delta \sim N(\mu = -0.5, 1)$, respectively.

Changing the mean of the treatment effects of the available treatments affect the number of successful phase III trials found over a long period of experimentation. It is clear that when there is a positive treatment effect, on average, the number of successful phase III trials increases, while the opposite is true when the treatment effects are negative, on average.

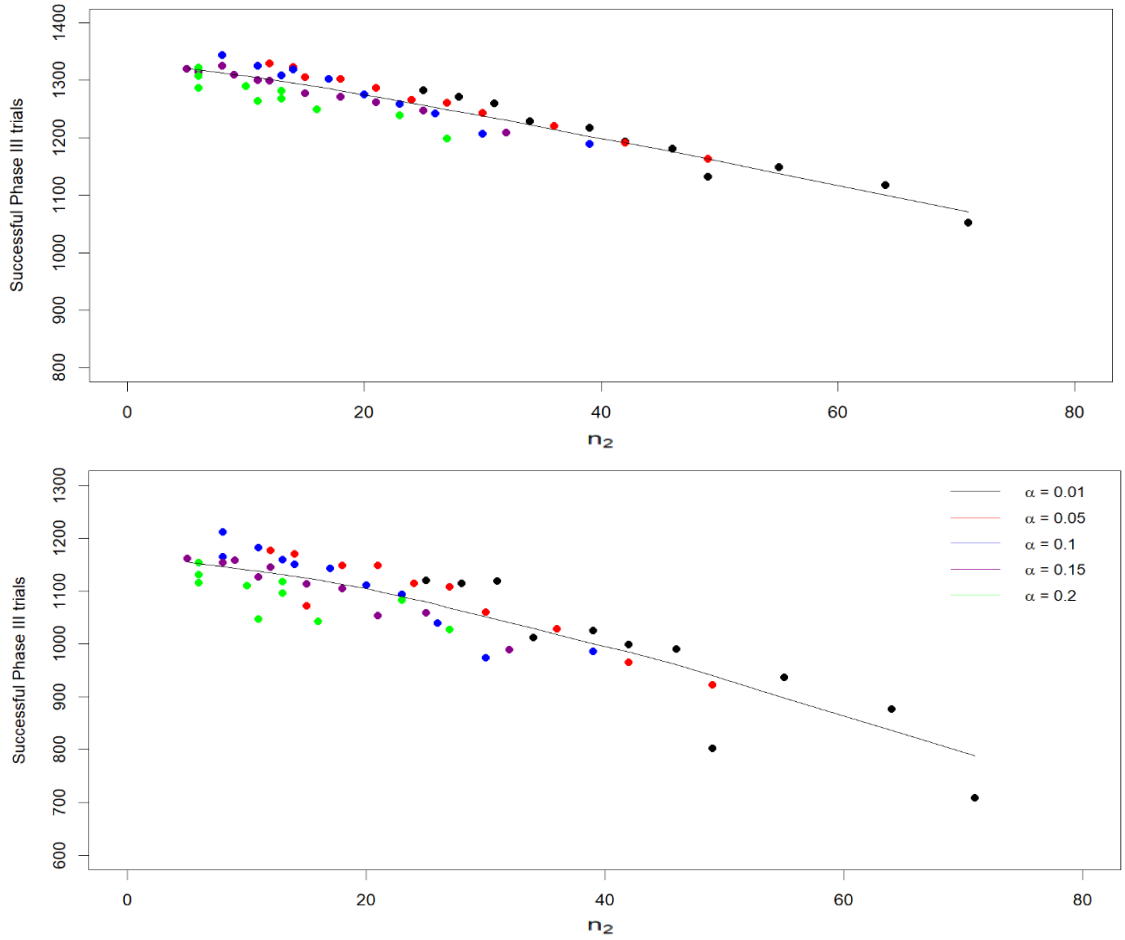


Figure 7.4 The total number of successful phase III trials depending on the number of patients in Simon's phase II trials differentiated by the value of α; positive treatment effect $\mu = 0.5$ (top); negative treatment effect $\mu = 0.5$ (bottom)

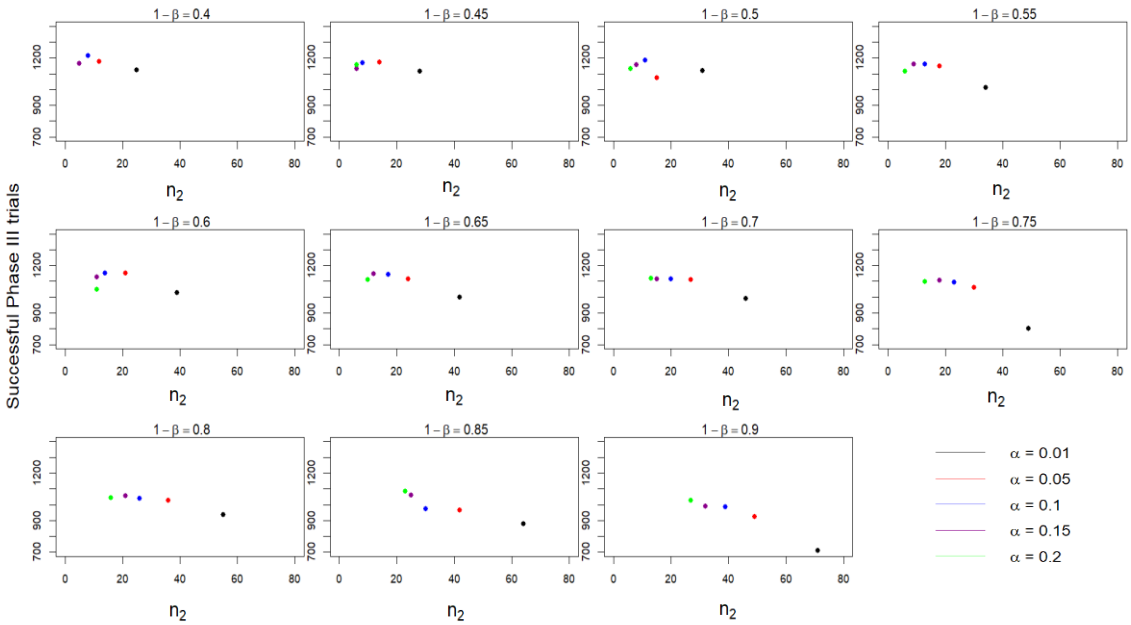


Figure 7.5 The total number of successful phase III trials, shown by the power of each Simon's two-stage single-arm phase II trial with the corresponding value of α for the negative treatment effect scenario

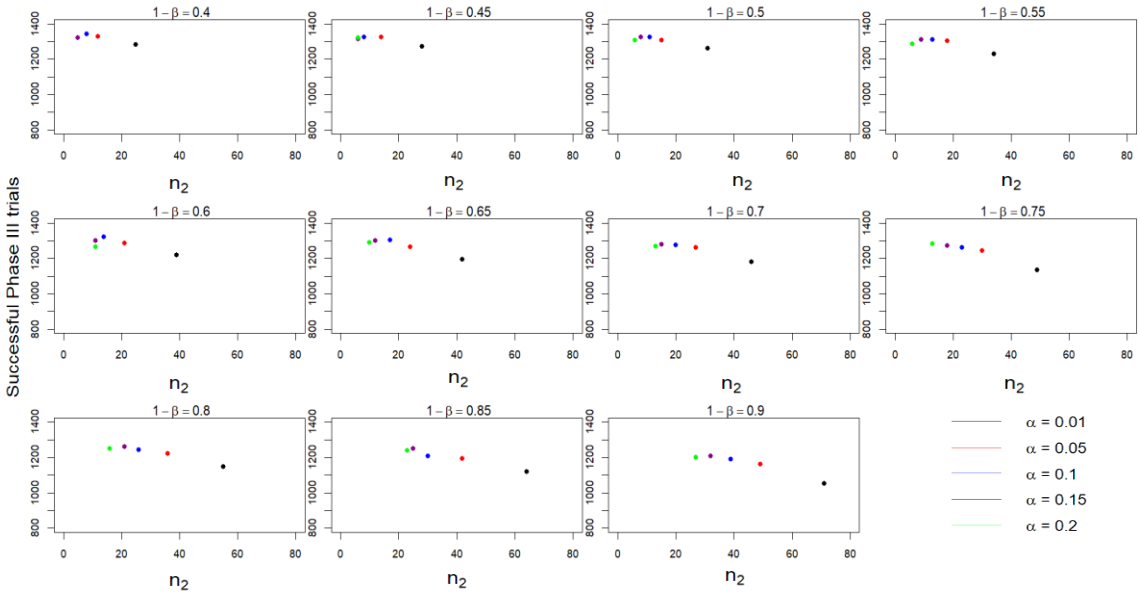


Figure 7.6 The total number of successful phase III trials, shown by the power of each Simon's two-stage single-arm phase II trial with the corresponding value of α for the positive treatment effect scenario

The optimal sample size is unchanged in the two scenarios, compared to the mean of 0 scenario, described above, however the variability for each value of α investigated is quite large in the negative treatment effect scenario. It is clear from Figure 7.5 that the variability occurs due to changes in power. The most efficient design is where the power is set between $0.4 - 0.6$. When the power is larger than 0.6 the effect of α becomes more prominent, as shown by the large difference between the less stringent α $(0.1, 0.15 \ \& \ 0.2)$ and the stringent values $(0.01 \ \& \ 0.05)$ in the number of phase III successes. Despite this difference in the number of successful phase III trials, these conclusions were unchanged in the positive treatment effect scenario, presented in Figure 7.6, indicating that the findings are robust to changes in the mean of the treatment effect distribution.

7.3.2 Randomised phase II trials

The total randomised phase II trial sample sizes yielded, ranged from 8 to 284 patients. The systematic review (Chapter 2) revealed that phase II trials are usually designed to recruit around 100 patients, on average. The range chosen for the sample size investigations is larger than that found in the systematic review however, it was selected for the purpose of being exhaustive. This would reveal whether there is a point in which increasing the sample size of randomised phase II trials leads to inefficiencies in the drug development process.

A total of 55 randomised phase II design combinations with different α and β's were investigated. Figure 7.7 shows the total number of successful phase III

trials for the range of sample sizes investigated. The number of successful phase III trials ranges from approximately 447 to 1074. Initially, the number of successful phase III trials increase as the sample size increases, however, after a certain point, it is clear that the number of phase III trial successes decrease as the sample size of the phase II trials increase. Given the assumptions made, namely an average effect of the treatments available of 0, a $p_1 = 0.25$ and $\delta_2 = 0.2$ this indicates that, randomised phase II trials shouldn't be designed to be too small ($n_2 < 10$). In addition, the most efficient sample size for randomised designs ranges from a total of $14 - 30$ patients and anything above this actually causes a decline in the efficiency of phase II trials.

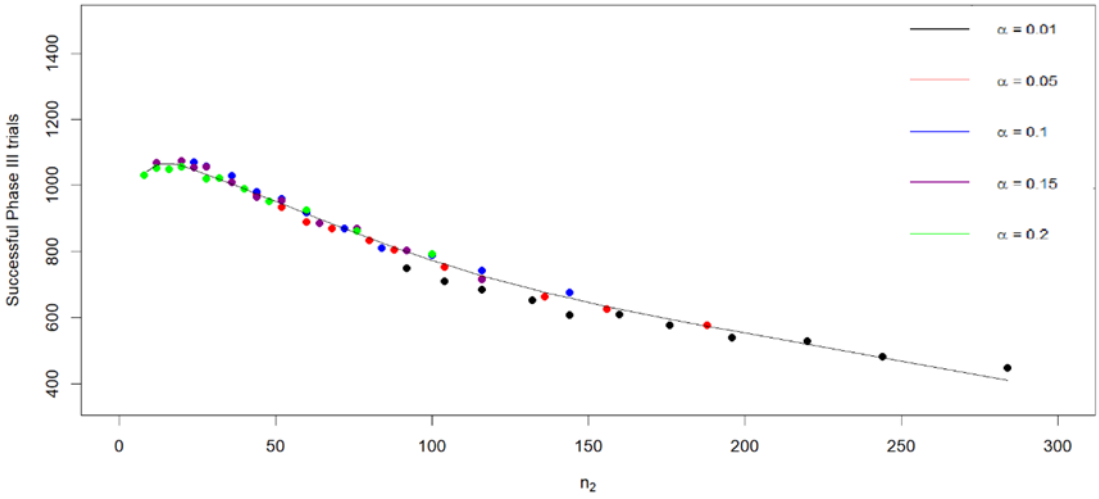


Figure 7.7 The total number of successful phase III trials depending on the number of patients in two-arm phase II trials

In the randomised design scenario, it can be seen that the type I error has a profound effect on the number of phase III trials, and therefore the efficiency of the phase II trials. Figure 7.7 shows that stringent α's ($\alpha \leq 0.05$) yield the lowest number of successful phase III trials, while less stringent type I errors $0.1 \leq \alpha \geq 0.2$ result in a higher number of successful phase III trials.

The impact of the power was assessed in Figure 7.8. As previously stated, the number of successful phase III trials decreases with an increase in sample size. Consequently, too high a power yields the lowest number of phase III trial successes. The number of successful phase III trials drastically decreases as the power increases. The range of power which results in the highest phase III trial successes is between $0.4 - 0.6$. However, it is also clear that the value of α has a negative impact on the efficiency of phase II trials when it is equal to 0.01 and 0.05. So long as the power is at $0.4 - 0.6$ and α is larger than 0.1 then the randomised design in phase II would be optimised, over a series of experimentation.

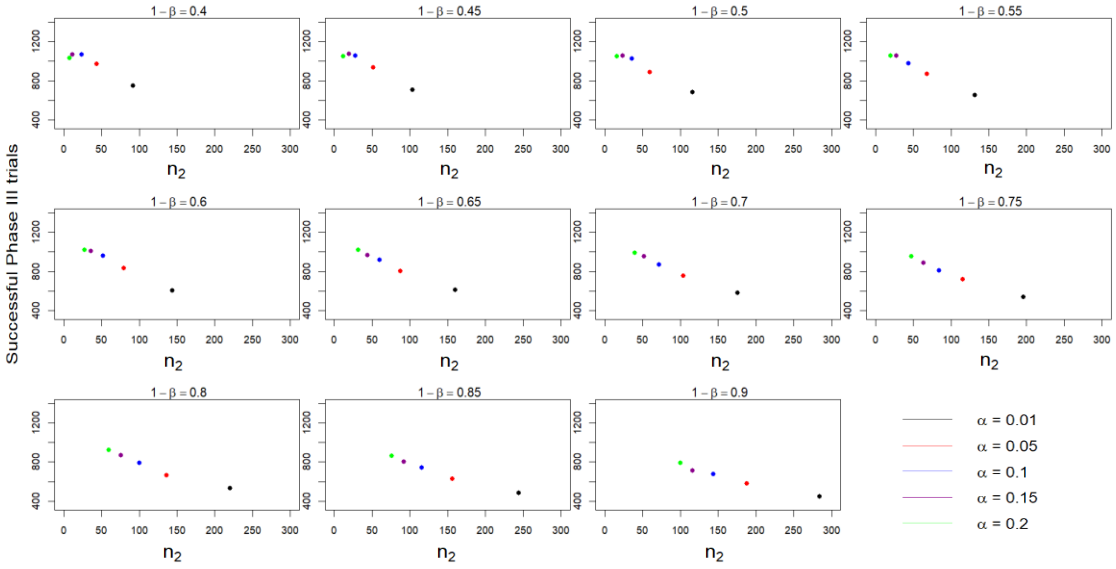


Figure 7.8 The total number of successful phase III trials, shown by the power of each randomised phase II trial design

The impact of the sample size was supplemented by exploring the success rates of the phase II and III trials. Figure 7.9 shows the success rates of the phase II and III trials, with the points corresponding to the designs with different type I error rates. The overall success rate of phase II trials is smaller than the success rate of phase III trials, with the phase II success rate ranging from approximately 0.30 to 0.44, while the phase III success rate ranges from approximately 0.77 to 0.99. It is clear that both the phase II and III trial success rates increase as the randomised phase II sample size increases. However, the magnitude of success rate in phase III is much larger than the phase II. The success rates in phase II and III are affected by the sample size of a randomised design in phase II, in much the same way as that of Simon's two-stage design. The phase III trial success rate increases as the sample size increases. The sample size increases with increasing power, which also increases the probability of success. Decreasing α (which decreases the probability of success) also increases the sample size. Similar to Simon's design, the fact that the probability of success in phase III is higher than in phase II shows that phase II is leading to poor treatments being dropped, with this happening more when the sample size is larger.

Of interest is how the type I error affects the success rates of the phase II and III trials. It is clear that very stringent type I errors ($\alpha = 0.01$ and $\alpha = 0.05$) result in the highest success rates in phase III trials but yield the lowest success rates in phase II, overall. This means that randomised phase II trials with stringent type I errors are more likely to lead to successful phase III trials, as the stringent α rarely allows futile treatments to proceed to phase III, therefore only those that are truly efficacious at phase II are allowed to proceed to phase III. This

contradicts the earlier conclusion regarding the use of different values of α in randomised phase II trials, that less stringent α rates in randomised phase II trials lead to a higher number of successful phase III trials. However, this can be reconciled, as the number of successful phase III trials is a long-term measure of efficiency, suitable for carrying out multiple phase II and III trials over a long period of time, whereas the success rate indicates the short-term benefit of the phase II trial. For example, in an individual phase II that leads to one phase III trial, a stringent α would be more efficient, in terms of the percentage success rate of phase III. It should be noted, though, that making the phase II type I error rate more stringent means that the phase II trials are less likely to lead to a phase III, but when they do, the phase III trials are more likely to be successful.

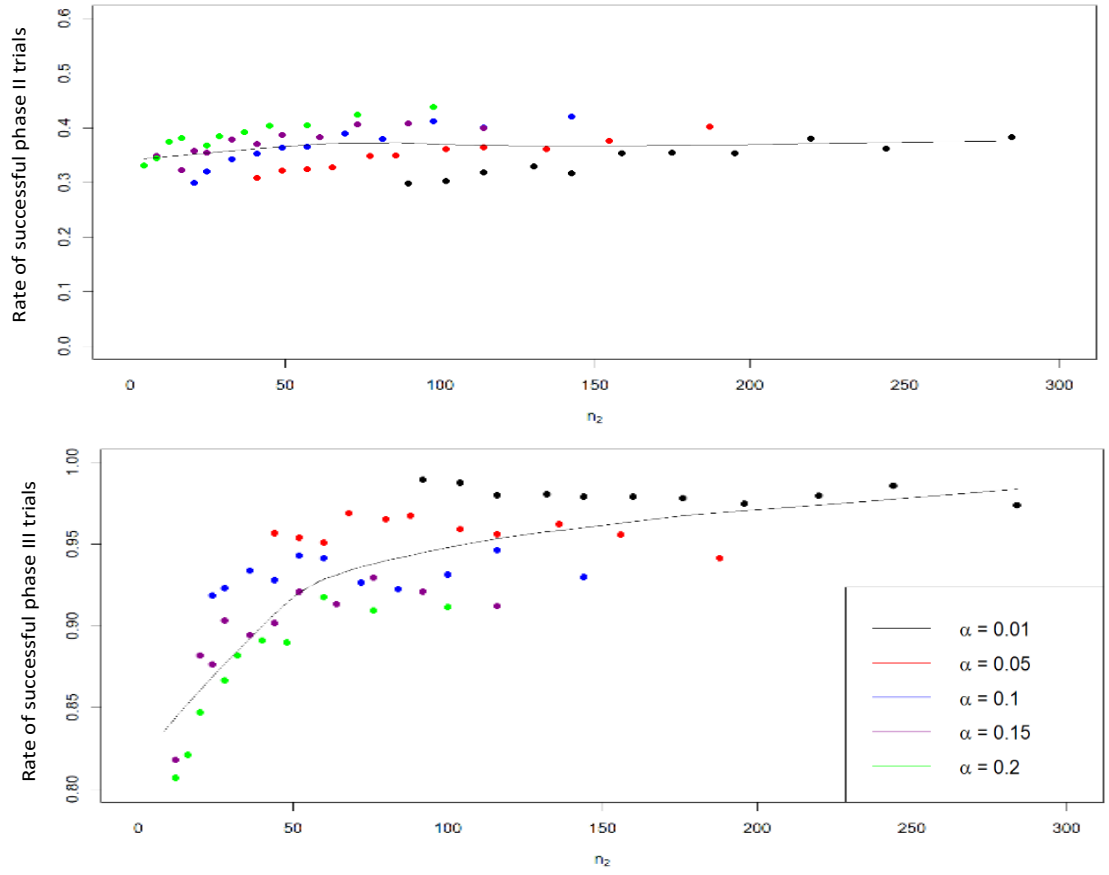


Figure 7.9 The success rate of the randomised phase II trials (top); the success rate of the phase III trials (bottom)

7.3.2.1 Sensitivity analysis for the randomised design

The effect of the sample size of randomised designs in phase II trials was assessed in a sensitivity analysis, where the mean of the treatment effects available was assumed to be either positive or negative, rather than no effect, on average. In Figure 7.10 the effect of changing the mean of the true treatment effect is highlighted: it is clear that when there is a negative treatment effect

there is more of an increase in the number of successful phase III trials than the negative situation initially, and therefore the peak is more prominent in the negative scenario.

Overall, the number of phase III successes is also reduced between the two scenarios. The optimal total sample size differs very slightly between the two scenarios: in the positive treatment effect scenario, the optimal sample size is between $10 - 25$, while in the negative scenario, the total sample size is most efficient when it is between $14 - 35$. This is also only slightly different to the no effect scenario where the range was between $14 - 30$. The most efficient designs are those with an $\alpha = 0.1, 0.15 \ \& \ 0.2$, in all scenarios.

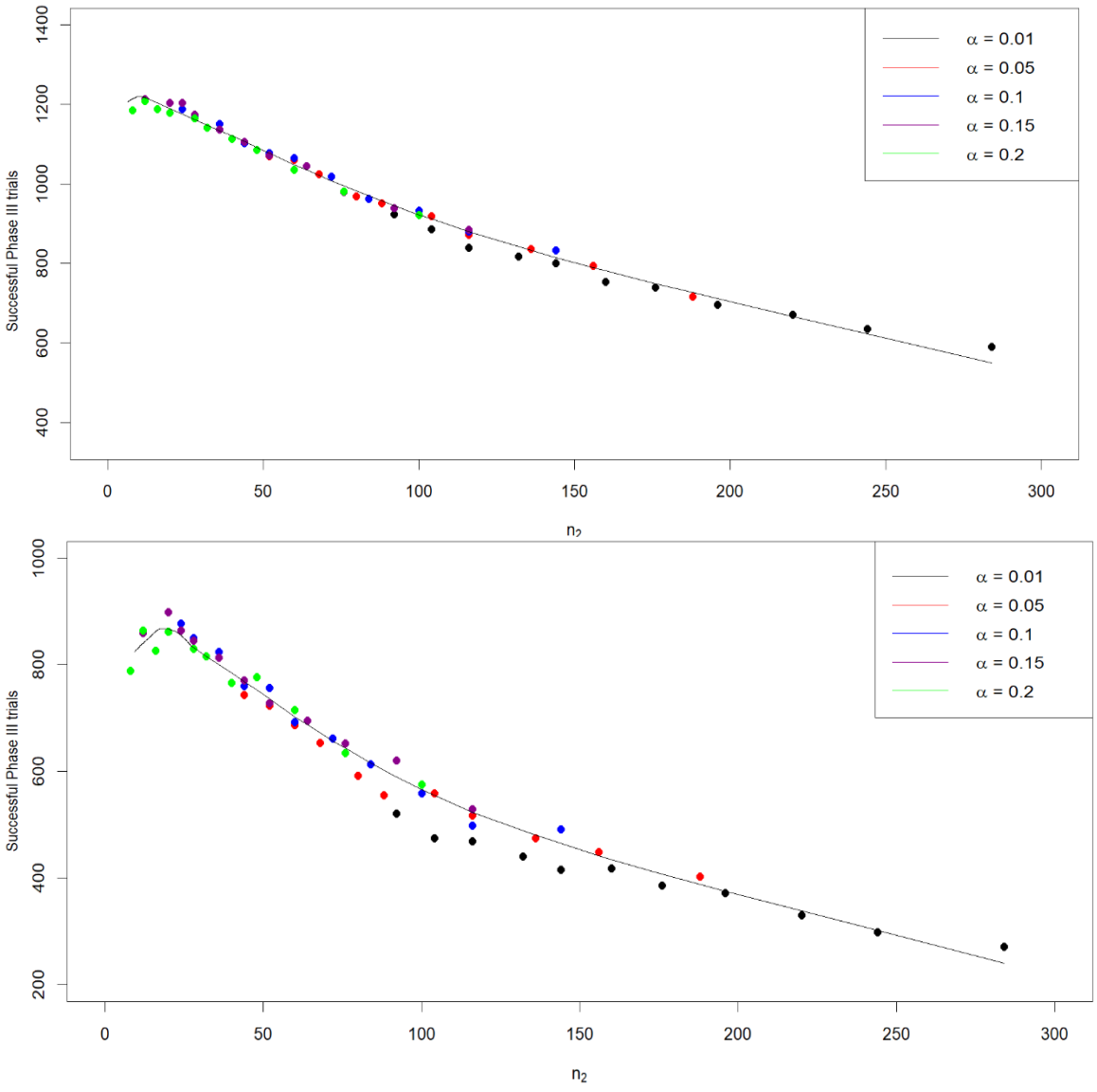


Figure 7.10 The total number of successful phase III trials depending on the number of patients in randomised phase II trials differentiated by the value of α; positive treatment effect $\mu = 0.5$ (top); negative treatment effect $\mu = 0.5$ (bottom)

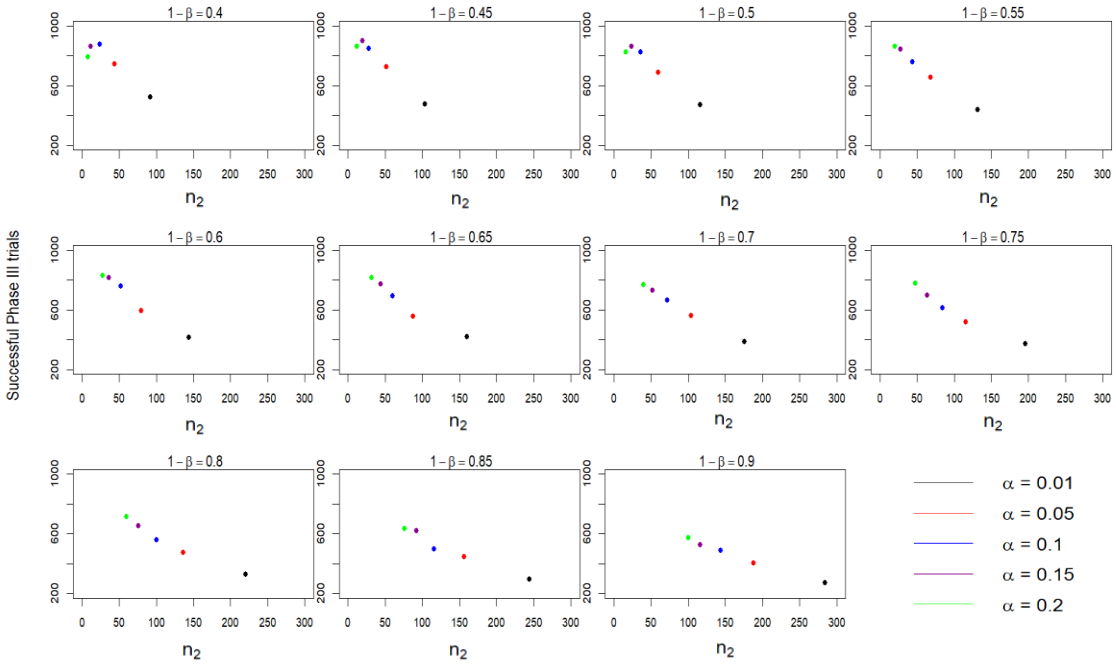


Figure 7.11 The total number of successful phase III trials, shown by the power of each randomised phase II trial with the corresponding value of α for the negative treatment effect scenario

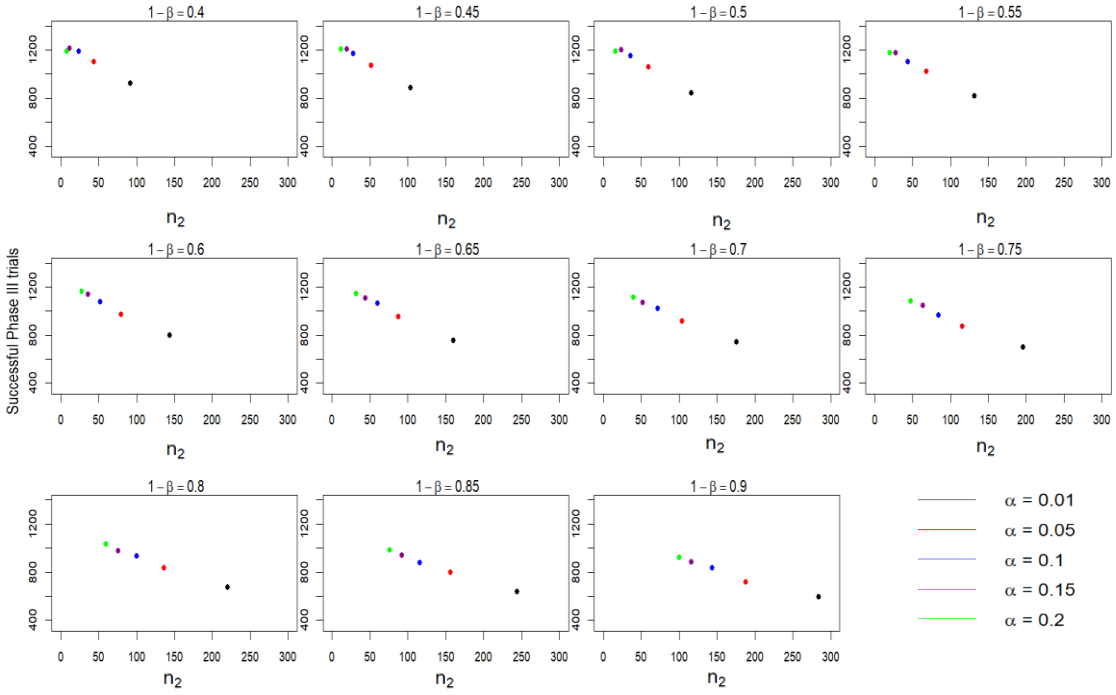


Figure 7.12 The total number of successful phase III trials, shown by the power of each randomised phase II trial with the corresponding value of α for the positive treatment effect scenario

The effect of the power in the negative and positive scenarios are shown in Figure 7.11 and Figure 7.12, respectively. Both figures show that the lower the power the more efficient the phase II trial. The optimum range of power, in either scenario, is between $0.4 - 0.6$. It is also clear that the effect of α becomes

more prominent as the power increases, as indicated by the fact the differences in phase III successes differ greatly between the $\alpha's$ investigated, particularly as the power increases. Therefore, with higher power less stringent $\alpha's$ are more efficient than using stringent values. This was the case for all values of the mean of the available treatment effects investigated, hence the findings for the sample size were robust.

7.4 Discussion

The purpose of this chapter was to assess the impact of the sample size of phase II trials on their ability to successfully screen new treatments in phase III trials. The sample size of any trial is an important element and can be a contributing factor to the trial's success. Too small a sample in phase II means that the treatment effect is potentially missed, while too large a sample size means that resources could be wasted. The aim of this chapter was to find out if the efficiency of phase II trials always increases with an increase in its sample size or whether there is a point where increasing the sample size adds no value to the drug development process. The measure of efficiency was defined as the number of successful phase III trials, given a large, fixed population size, so that the interpretation of the results apply to the scenario where multiple testing in phase II and III trials occurs, such as in large pharmaceutical companies.

Two designs of phase II trials were chosen, specifically Simon's two-stage single-arm design (63) and the randomised single-stage design with an with 1:1 ratio in each arm. These were chosen based on the findings in Chapter 6, where they were found to have the most advantages over the other designs investigated. It should be noted that the purpose of choosing two designs to investigate their sample size was not to compare their performance (this was discussed in detail in Chapter 6), rather it was to provide a profile of their efficiency, so that trialists may use this information to guide the process of designing phase II trials that better fulfil their purpose.

Both the two-stage single-arm and randomised phase II trials with the highest number of successful phase III trials were those that require a small number of patients. The sample sizes yielding the highest number of successful phase III trials was found to be between $5 - 15$ patients, in total for Simon's design, while the highest number of patients required in a randomised design was between $14 - 30$ patients, in total. Since the randomised phase II trials were assumed to have 1:1 randomisation ratio, only $7 - 15$ patients are required in each of the control and experimental arm. It was also concluded that total sample sizes larger than around 20 patients using Simon's design and 50 patients using the randomised design lead to a dramatic decrease in the number of successful

phase III trials, over a long period of time. Thus, anything higher than these totals would reduce the efficiency of phase II trials and, as a result, negatively impact the drug development process. The two-stage single-arm phase II trial recommendations made here contradict Schoenfeld(116), who suggested that 25 patients are required for single-arm single-stage phase II trial. This could be attributable to the different designs used here and by Schoenfeld (116). However, Julious's (117) recommendation of a minimum of 12 patients per arm in phase II studies coincides with the randomised sample size range recommended here, though my findings suggest randomised phase II trials could be as low as 7 patients per arm. Stallard (39) also recommends a sample size of 35 in a randomised phase II trial designed to detect an effect size of 0.2. This also coincides with the findings reported in this chapter. The sample sizes recommended here are not what is typically used in phase II trials. The systematic review (Chapter 2) showed that, of the phase II trials investigated, 65.6% of them recruited more than 50 patients in total and more than half of those (34.3%) accrued more than 100 patients. Such sizes of phase II trials would lead to inefficiencies over a long period of time as reported in this chapter.

The effects of the type I and II errors, which determine the sample size, were explored in this chapter. The one-sided type I error rate which yielded the most efficient Simon's two-stage single-arm (63) and randomised phase II designs was found to range from $0.1 - 0.2$, of those that were investigated. Higher type I error rates resulted in more efficient phase II trials, set up using both Simon's two-stage single-arm and randomised designs. This coincides with Stallard's (39) recommendation of 0.2 for the value of the type I error. Sharma, Karrison et al. (71) also suggest that a one-sided significance level of 0.1 is consistent with the more exploratory nature of a phase II study, and therefore is more appropriate, which is demonstrated in this chapter. Ratain & Sargent (96) also hold this opinion as they believe that the sample size of a phase II trial should be as small as possible.

My findings also showed that stringent type I error rates ($\alpha \leq 0.05$) resulted in a decrease in the number of successful phase III trials, using both Simon's and randomised designs, compared with the other type I error values. Such stringent type I error rates resulted in high success rates in phase III but low success rates in phase II trials. This is due to the fact that the smaller the value of α the harder it is to get through to phase III, so with a small α fewer treatments get through, but those that do are truly efficacious, so are therefore more likely to be successful in the subsequent phase III trial. High success rates using stringent α, combined with the fact that they result in a lower number of

successful phase III trials, implies that, the short-term benefits (when looking at the success rates) of using a stringent type I error is more profound than using a more relaxed type I error. Comparatively, less stringent type I errors should be used in the long term where there are a plethora of phase II and III trials and sufficient funds and resources to carry them out. Pond and Abbasi (83) also found that stringent type I error rates increases the success rate of phase III trials, and they, therefore, advocate stringent type I error rates. However, it should be noted that Pond and Abbasi (83) only explored the success rate of the phase III trial, and did not incorporate the long term number of successful phase III trials. With so many drugs available for testing in the current era, it is not the success rates of phase II or III trials that determine the efficiency of phase II trials, where success rate is defined as the proportion of successful trials out of the trials which are carried out; rather it is the number of successful phase III trials that need to be increased, which is the measure reported in this thesis. Further justification for the choice of the measure of efficiency is provided in Chapter 8, Section 8.3.

The effect of the power, which is the complement of the type II error rate, on the efficiency of phase II trials was found to have a profound impact. The results of efficiency of Simon's and randomised designs of phase II trials indicated that as the power increases the number of successful phase III trials decreases, indicating that phase II trials are more efficient with smaller power. It was also found that, for all the designs investigated, an acceptable range to use for power was between $0.4 - 0.6$, with 0.4 yielding the most efficiency. Using this range with any value of α $(= 0.05, 0.1, 0.15, 0.2)$, except 0.01, would not jeopardise the efficiency of phase II trials. However, beyond this power $(1 - \beta > 0.6)$, the value of α needs to be carefully selected: with higher power, less stringent α's $(= 0.1, 0.15, 0.2)$ are better for the efficiency of phase II trials. The optimal combination of α and $1 - \beta$ was found be 0.15 and 0.4, respectively. This is in line with Stallard's (39) recommendations of a power of 0.4 and is very close to the recommendation he made for the value of α $(= 0.2$, as mentioned earlier). However, these recommendations are not in line with what is, currently, used in practice. The systematic review (Chapter 2) revealed that the type I error is very rarely larger than 0.1 (only 2.3% of the phase II trials investigated). In addition, the value for the power of 0.4 was never used in the phase II trials, reviewed. Only two trials used a power smaller than 0.8: one used 0.75 and the other used 0.7. The recommendations in this chapter highlight the need to use much smaller powers and high α's in order to maximise the efficiency of phase II trials.

It is clear that many assumptions have been made in order to reach the conclusions in this chapter. It was assumed that there is an unlimited number of treatments available and that their effect sizes were assumed to come from a known distribution that is independently and identically distributed from a standard normal distribution. While there are currently many treatments available for testing, the number of treatments is not unlimited. However, as technology enhances and as we strive to provide better treatments for cancer patients, this assumption is optimistic but seems reasonable.

It is acknowledged that using a standard normal distribution for the effects of the treatments available for testing is likely to affect the conclusions made in this chapter. As such, a sensitivity analysis, adjusting the distribution, was conducted in order to assess the robustness of my findings. Two further scenarios were investigated: when there is a positive treatment effect and when there is a negative treatment effect. Differences in these two scenarios arose in the number of successful phase III trials found in all scenarios investigated however, the conclusions regarding the sample size, and as a result of this, the operating characteristics, α and β remained the same. Therefore, the conclusions made are robust to this assumption.

Another assumption that may have had an influence on the conclusions made is the choice for the value of the effect size, $\delta = 0.2$ and the choice of $p_1 = 0.25$. It is acknowledged that the effect size varies from trial to trial and that the choice of δ affects the sample size, however, the chosen value was selected based on the fact that it was the average effect size used in the phase II trials evaluated in the systematic review (Chapter 2). The choice of p_1 was also based on previous literature, where Taylor et al. (70) used values of 0.1 and 0.3 to evaluate the benefits of single-arm and randomised phase II trials. With the same aim, Pond and Abbasi (83) ranged their p_1 value from 0.05 to 0.75 by increments of 0.05. Combining these two articles, it was decided that 0.25 is a reasonable value. It is also clear that previous authors varied these values to assess their impact and this could be a potential avenue that could be taken to enhance the work reported.

In the simulations, it was also assumed that there are a large number of patients available for the drug development process, including phase II and III trials. This is a reasonable assumption as the number of cancer patients is constantly increasing (118). Since some conclusions were found to be sensitive to these assumptions, exploring the effect the different distributions for the prior treatment effects would be of benefit. Furthermore, assuming that a small number of patients is available would allow us to explore the effects of the sample size on the efficiency of phase II trials in rare disease sites.

Another strong assumption that was made was the fact that the effect sizes of the endpoints in phase II and phase III trials were strongly correlated. Ensuring a strong correlation increased the number of successful phase III trials in comparison to the situation where the treatment effects of the endpoints in phase II and III trials were uncorrelated. In reality this assumption is unlikely to be met, however, this assumption was made in order to assess the impact of the sample size without other factors affecting the efficiency of phase II trials. Furthermore, this assumption was made as a result of the findings in Chapter 5, where the phase II trial efficiency was enhanced with a strong correlation between treatment effects of the trial endpoints, therefore this is an important design parameter that needs to be incorporated in the design process of phase II trials.

In conclusion, I have found that both Simon's two-stage single-arm design and the randomised design for phase II trials are the most efficient with small sample sizes. It was also found that the choice of the type I and II errors has a profound impact on the efficiency of phase II trials, and as such, it is recommended that a less stringent one-sided $\alpha = 0.1, 0.15$ and 0.2 is used with small power ranging between $0.4 - 0.6$. This is not what is typically used in practice however, in the long run, designing phase II trials with the recommended values would be more efficient.

Chapter 8 Discussion

Over the years, the design of phase II trials has evolved, mainly due to the increase in the number and variety of the treatments available for testing (35). When the number of treatments were limited, the gold standard design for oncology phase II trials was Gehan's single-arm two-stage design (119), where 14 patients are recruited to ascertain whether the true response rate was less than 20% with 95% confidence (35). If one response was found, a second stage would be initiated to estimate the true response rate. With the increase in the number of treatments available for testing and the demand for a higher standard of evidence of potential clinical benefit (35), Simon's single-arm two-stage design (63), evaluating tumour response rate of novel treatments, became popular (12). In the current era, where there is a plethora of treatments that need to be tested in phase II trials, coupled with the emergence of molecularly targeted agents that may be more likely to be cytostatic, rather than cytotoxic, the most efficient design of the oncology phase II trial is again open to much debate. In addition to this, the depressing statistics that are prevalent in oncology phase III trials, where they are reported to fail more than 50% of the time (10) has meant that the efficiency of phase II trial designs need to be evaluated.

As such, the aim of this thesis was to explore the effects of phase II design parameters in terms of their ability to successfully screen new treatments. With the knowledge of the findings presented here, researchers are fully aware of the implications of their chosen design. This is an important tool for researchers as it may inform decisions about the design parameters that will increase the current efficiency of phase II trials, and therefore improve the efficiency of the drug development process as a whole. Specific recommendations about the most efficient choices for the design parameters investigated are also provided, as a result of the findings.

The design parameters that were investigated were based on the results of the systematic review (Chapter 2). The correlation between phase II and III endpoints was explored. Here, the aim was to reveal the effects of correlation between the endpoints on the efficiency of phase II trials. In addition, the model used included other parameters, such as the variance of the treatment effect, that affected the efficiency of phase II trials, and thus allowed the exploration of their effect on phase II trial efficiency. Another design parameter explored was the decision to use a randomised design and the decision to use two-stage

designs or their counter parts. The goal was to reveal which design is the most efficient. The final design parameter explored was the sample size of phase II trials, and as a result, the trial's operating characteristics. Here, the aim was to find out if increasing the sample size of phase II trials would increase efficiency or whether there is a point where recruiting patients beyond this point would lead to inefficiencies. In addition, the effect of the operating characteristics on efficiency, namely the type I and II errors of the phase II trials was explored. These parameters were investigated given a pre-specified set of underlying assumptions regarding the effects of the available treatments and the targeted treatment effect in each of the phase II and III trials, with efficiency defined based on the number of patients required to observe a positive phase III trial, and evaluation of treatments over a prolonged period of time. Given these assumptions, the conclusions made, in this thesis, regarding optimising the efficiency of phase II trial design was that the relationship between the treatment effects of the phase II and III endpoints should be correlated. In addition, the most efficient phase II design was Simon's two-stage single-arm design. Finally, the sample size of the phase II trial is at its most efficient when the type I error is less stringent and when the power is small. Interestingly, the phase II trials do not increase in efficiency as the sample size increases.

In this chapter, a critical evaluation of the methods used in this research is presented and potential areas where further work may be warranted is highlighted, in addition to a discussion of how this research can impact the future design of phase II trials.

8.1 Phase II design parameters

The phase II design parameters investigated in this thesis were selected based on the findings in the systematic review (Chapter 2). The main aim of this chapter was to reveal how phase II trials are designed in the current era. Consequently, this would reveal whether there is a consensus among researchers about the design of phase II trials. It is acknowledged that the systematic review was limited to three high impact journals and only looked at phase II trials published during two years. The high impact journals were chosen to evaluate how the most impactful phase II trials are designed, and therefore potentially the most likely to influence future decision making, whilst being pragmatic about the size and scope of the review. It also was an update of the systematic reviews that were previously conducted by other researchers and therefore some methods were adapted from them: Langrand-Escure et al. (9) reviewed phase II trials published in three high impact journals between the years 2010-2015. Mariani and Marubini (20) also reviewed phase II trials but

only those that were published in 1997. Instead of reviewing phase II trials published in one year, Chapter 2 reports findings from phase II trials published during two years, and since Langrand-Escure's et al. (9) review spanned across five years, the two years reviewed were selected to be four years apart to capture any potential changes to phase II designs during that time. The studies included in the systematic review (Chapter 2) were selected using the same keyword search as Mariani and Marubini (20).

Using only three journals over two years to reach the conclusions limited the findings, possibly resulting in a lack or representation of the use of novel treatments that may be more commonly used in higher impact journals, than the ones used. In addition, the choice of journals was based on Langrand-Escure's et al. (9), where they chose three high impact journals. The impact of their chosen journals may have changed since their review was published, which may have further resulted in the lack of representation of novel designs. Another limitation of the systematic review was the use of a single database, namely, Medline. While it can be very useful, using one database can also restrict the findings and run the risk of excluding relevant studies. However, the systematic review helped identify the design parameters that are in contention among researchers, therefore it fulfilled its purpose, efficiently (using only three journals).

Ultimately the design parameters that were selected for evaluation in this thesis were the correlation between phase II and III trial (i.e., the choice of the phase II endpoint that determines the correlation), the design choice, specifically incorporating randomisation or two-stage designs, and the sample size (and as a result the operating characteristics) of phase II trials. These parameters were chosen due to the fact that decisions about them can be controlled by the researcher. Other parameters such as population wide shifts in outcomes, caused by improved treatments (78) can influence phase II trial efficiency. However, while this parameter is an important consideration for the researcher to take into account during the planning of a trial, it is not a parameter that they can influence directly.

Two endpoint types were used for the phase II trials, throughout the evaluations. A continuous endpoint was assumed when the correlation between the treatment effects of the phase II and III endpoint was investigated. It is acknowledged that phase II trials rarely incorporate continuous outcomes, in fact in the systematic review, none of the included phase II trials used a continuous outcome. However, this was chosen due to the mathematical ease of using a normal distribution to quantify the effect of the treatments on both the phase II and III endpoints. This endpoint was then changed to a binary endpoint

when simulations were used. A binary endpoint was selected based on the fact that it is the most commonly used endpoint in phase II trials, as demonstrated in the systematic review (Chapter 2) where 79.7% of phase II trials utilised a binary outcome. Future work could investigate the impact of phase II trials that use a time-to-event outcome since they were used in 23.4% of the trials included in the review.

The focus of the phase II trials in this research was to determine whether the treatment was efficacious. However, many phase II trials fail to proceed to phase III trials due to the treatment's toxicity. There are designs that incorporate both endpoints into their design, such as the Bryant and Day design (120). Therefore, an extension to this research project could be the explorations of such designs, in terms of their effect on phase II trial efficiency.

Other designs could have also been included in the evaluations such as Fleming's two-stage design (64) or even the use of Bayesian designs in phase II trials. Berry et al. (121) looked at Bayesian designs of phase II trials in oncology. They investigated Simon's two-stage design (63) and a Bayesian adaptive design with frequent interim analyses and a Bayesian adaptive design with frequent interim analysis and hierarchical modelling across patient groups. They defined this as "borrowing" information from one group to estimate the treatment effect of another (121). They found that phase II trials that use a Bayesian hierarchical design is more powerful than other designs investigated, i.e., trial comes to the correct conclusion. Fleming's two-stage design and Bayesian designs were not included in this thesis as they were not found to be popular in the systematic review (Chapter 2). Fleming's design was only used 2.3% and a Bayesian multi-stage design was used in one phase II trial (0.78% of the trials investigated).

When designing a phase II trial, the choice between the designs available are broad and may be affected by external factors. For example, the choice between randomisation and single-arm designs may be affected by the rarity of a disease, the seriousness of the condition or high unmet medical need (122). The value of historical controls may be of importance and regulators have accepted the use of historical controls in these cases (123). One such method is to apply a Bayesian (informative) prior to the historical data (123) and analysing the data using frequentist methods. This is known as a hybrid approach as it combines elements from both Bayesian and frequentist methods (92). Such designs and methodologies demonstrate the breadth of the choices available to researchers when designing phase II trials. It is acknowledged that the results and recommendations do not cover all options available and that these were outside the scope of this research. However, the parameter choices

were based on the findings of the systematic review outlining what is currently used in phase II trials.

Throughout the parameter evaluations, the design of phase II trials was specific to a single scenario where the targeted treatment effect was set to 0.2. In reality the targeted treatment effect in phase II trials can be variable, which would affect the sample size, and hence the design and outcome of the phase II trial. A difference of 0.2 is characterised by Cohen's d as a small difference (124, 125) and is more difficult to find than a larger value. This was chosen as a conservative value and also based on the systematic review where this difference was the most common choice on average (see Chapter 4). This effect size was extracted only from those phase II trials which reported it or reported their hypothesis and used a binary endpoint.

8.2 The perspective of running multiple trials

The evaluations of the impact of phase II design parameters in this thesis were carried out under the assumption that phase II and III trials can occur in a sequential manner, spanning several years. This assumption gives rise to three consequential assumptions that are needed. The first is that there is a large number of patients available for testing, in the simulations this was set to five hundred thousand. Given the fact that the trials are assumed to run over several years this assumption is realistic, particularly that there is increasing number of people developing cancer, with about an average of 1000 new cases every day in the years between 2016 to 2018 (97).

The second assumption is that there are a large number of treatments available for testing. This is not so farfetched given the fact that in the current era where there is a rise in treatments available for testing. Of course, with a large number of treatments to test, the resources required to launch and run these trials are great and therefore the application of my findings are well suited to pharmaceutical companies where they have the capacity to run multiple phase II and III trials, sequentially. As a result, the findings do not have a direct application to individuals designing a single phase II trial.

The third consequential assumption that arose from running multiple trials is that the treatments are assumed to follow some distribution. Throughout the thesis this has been referred to as the true treatment effect distribution. This was chosen to follow a standard normal distribution. A normal distribution was chosen as it was thought to be a close reflection of reality, given the central limit theorem which states that if a sufficiently large sample is selected from a population, in this case treatments available, then the sample means follow a

normal distribution (126). The choice of the mean for the treatment effect distribution was zero which is interpreted as the treatments available having no effect on average. The proportion of phase III success, seen in the results in earlier chapters was quite high (>90%) but in reality, such numbers are not seen. Harrison (47) indicated that phase III trials fail 32% of the time. This indicates that the distribution of the treatments available is far more pessimistic in cancer than the scenarios investigated.

A sensitivity analysis was conducted to assess the robustness of the assumption of a mean of no effect in the distribution of the treatments available. Two scenarios were considered: a positive treatment effect and a negative one. It was concluded that the phase II trial design recommendations do not change with a different mean in the treatment effects available, indicating that the sensitivity analysis have led to establishing that the findings in this research are robust. However, what does change is the number of patients required to lead to the first successful phase III trial and over the long run the number of successful phase III trials. In the positive mean treatment effect scenario, the number of patients required decreased, phase III trial successes increased and therefore the efficiency of phase II trials increased, while in the negative mean treatment effect scenario, the opposite happened. Other sensitivity analyses could have been conducted, namely, investigating changes to the variance of the normal distribution. This was only done in Chapter 5 to capture the effect of the correlation between the treatment effects of phase II and III endpoint. However, further work is warranted to explore what would be concluded about the best design and sample size, given changes to the variance. In addition, instead of a normal distribution, the effect of the true treatment effect distribution, could take a more flexible form such as a beta distribution, so that a variety of shapes can be used, which can incorporate the pessimistic outlook in cancer.

8.3 Measure of phase II trial efficiency

The measure of phase II trial efficiency was chosen by reviewing the literature to evaluate the methods used by other researchers. A set criteria were used in order to identify the most appropriate measure that can be used to measure the impact of the phase II trial parameters. The most important criteria was that the measure needed to incorporate the long term benefits of a phase II trial, given the fact the perspective is running multiple trials spanning several years. Many measures have been used by other authors, those that used utility functions were ruled out due to the fact that they included costs, which are difficult to estimate and differ between academic and pharmaceutical trials both nationally

and internationally. An alternative measure used was the proportion of successful phase III trials out of the total number of phase III trials conducted in a limited population.

With so many treatments available for testing, it is not the success rate of phase II or III trials that determine the efficiency of phase II trials, where success rate is defined as the proportion of successful trials out of the trials which are carried out; rather it is the number of successful phase III trials that need to be increased, where there is truly a positive treatment effect. Therefore, maximising the number of successful phase III trials results in the availability of more beneficial treatments for patients. Consequently, the most appropriate measure of efficiency of phase II trials is the number of patients required to lead to the first successful phase III trial. This measure was used by several authors previously: Stallard (39) minimised the number of patients required to lead to the first successful phase III trial to obtain the optimal phase II sample size. Yao et al. (87) minimised the number of patients required before the first promising treatment was identified in a phase II trial. They stated that this was appropriate as there are no real limitations on the number of patients available for enrolment into trials or on the number of treatments worthy of screening. Over a long run of experimentation, with several phase II and III trials, the endpoint used by Stallard (39) is equivalent to the total number of successful phase III trials, which was used in Chapters 6 and 7.

8.4 Phase III design

The design of the phase III trial was fixed throughout the evaluations of the parameters in this thesis. It was assumed that the phase III trials were randomised with two arms with a 1:1 allocation ratio measuring a continuous outcome. It is acknowledged that phase III trials are not typically conducted with a continuous outcome, rather a time-to-event endpoint is used, such as overall survival (127). However, the test statistics based on both time-to-event outcomes and continuous ones are asymptotically normal, in other words, the correlations are between these normal test statistics (108). Therefore, it was deemed appropriate to use this outcome. It also simplified the analytical evaluations conducted in Chapter 5 and the simulations in Chapters 6 and 7.

The operating characteristics used in phase III trials may also have had an impact on the findings presented in this thesis. In phase III trials the significance level was set at 0.05 and was two-sided, while the power was set to 80%. These values are typical choices made in phase III trials (98). However, other significance levels and powers could be chosen, which may increase or decrease the number of successful phase III trials found. However, since this

was fixed for all designs, it does not have a bearing on the conclusions made. Further research may be warranted to explore the effect of different phase III designs on the efficiency of the drug development process. This would be of great value particularly that the usage of MAMS trials are becoming more popular (128).

8.5 Implications of findings

In this research, I explored the effects of the key parameters used in phase II trials. For this reason, the findings presented are transferable to future phase II clinical trials. Applying these findings to the planning and design stage of phase II trials will improve their efficiency, and by extension the efficiency of the drug development process. Benefits to both patients and drug companies will thus be seen. However, this research cannot be used on its own; rather experience can also guide researchers to inform phase II trial design decisions.

Another area where this research can be used to improve matters is in rare cancers. Therapies for rare diseases are often not commercially beneficial and, despite efficacious drugs, these are not developed further as profit will not be large enough to sustain its production. Knowing the effect of the phase II design parameters on the efficiency of the drug development process can be particularly helpful in these cases, as researchers can design efficient phase II trials with more assurance of patient and commercial benefit. Perhaps a difficulty of applying these methods to rare cancer trials is the assumption of running multiple trials, implying that a large number of patients is available for testing. However, due to the setup of the simulations, where the population value was specified, an extension of the work presented here could easily be implemented by reducing the population size to evaluate the effect of each of the phase II design parameters where the population is limited. This extension would supplement existing research such as those presented by Hee and Stallard (91) and Stallard et al. (129) where they evaluate the optimal design for conducting clinical trials under the assumption of a limited population.

The implications of my findings can also extend to novel designs, such as, MAMS, platform trials, basket trials and umbrella trials. Platform trials are defined by the broad objective of finding the most efficient treatment for a disease by investigating multiple treatments, within one master protocol where there are multiple treatments available in quick succession of each other (130). Basket trials are used to establish the efficiency of a treatment for a mutation, found in several cancer types, where patients with a specific mutation are accrued into multiple arms of the trial (131). In contrast, umbrella trials accrue patients with the same cancer type (but different mutations), into one trial with

multiple arms, each of which is testing a different treatment (131). It is thought that these types of designs can help make the drug development process more efficient and are thought to be particularly useful within oncology, since it is a medical condition that has multiple therapies worth investigating (128). This is due to the fact that knowledge and understanding of cancer biology is constantly developing and it is being recognised that there are treatment options that target cancer cells, not just at a tissue level but at a molecular level. Multiple phase II trials are set up within a basket trial design and if one or more treatments show efficacy then they are tested in phase III confirmatory trials (130). This structure is similar to the assumption of running multiple trials, used in this research, except rather than simultaneous evaluations of treatments in basket trials, the multiple trial assumption lends itself to consecutive testing. However, both methods depend on the need to have multiple treatments and multiple patients available for accrual. Although in simultaneous trials having an exceptionally large number of patients available at any one time would be difficult, however, the recommendations here is that the sample size of phase II trials should be small (for the scenarios investigated), meaning that the large availability of patients is not necessary. Therefore, the recommendations made in this thesis can be applied to such designs.

Another advancement that is occurring in oncology drug development is the use of biomarkers in phase II trials (9). Before a biomarker can be used in a phase II trial it needs to be validated as a predictive marker of efficacy. In addition, historical data is not available for these biomarkers and therefore randomisation would need to be utilised in such trials. As seen in Chapter 6, the use of randomisation was found to be less efficient than using Simon's single-arm design, however, as stated in Section 6.4, when the historical control is not robust or is inaccurate, randomisation in phase II becomes the preferred and more efficient design. However, this advantage was not observed, as the uncertainty about the historical control rate was not a parameter incorporated in the simulations. Taylor et al. (70) and Pond and Abbasi (83) investigated the short-term effects of randomised and single-arm trials, where the accuracy of the historical control rate was incorporated. The authors recommended two-arm phase II trials when the uncertainty in the historical control rate is large or when a large number of patients are available. However, when these two conditions are not met, the authors advocated the use of single-arm phase II trials. Further work could investigate the long-term impact (over a series of multiple trials) of the estimate of the historical control rate comparing randomised and single-arm phase II trials.

While the focus throughout this thesis has been improving the efficiency of cancer phase II trials, the methods and subsequent results are generalisable and can be applied to different disease areas where phase II design can be improved due to the high attrition rates seen in subsequent phase III trials. Vaduganathan et al. (131) review phase II trials conducted in heart failure in order to identify the reasons for the disconnect between phase II and III trials. They found that discrepancies lie in the concept, design, execution and interpretation of phase II trials, and despite this phase III trials are initiated. Recommendations made in this thesis regarding the design of phase II trials, namely, a high correlation between endpoints, using Simon's single-arm two-stage design and a less stringent type I error and large type II error (equivalent to small power) to reduce the sample size, is relevant to other disease areas.

While these recommendations are not currently used in oncology phase II trials, as seen in the systematic review (Chapter 2), this research suggests that over a long period of experimentation, continuing with the same methods of designing phase II trials may mean that a great deal of resources are wasted, that could otherwise be used elsewhere.

As the cancer clinical trial environment constantly changes, it is important that considerations to completely novel designs, such as biomarker-based designs which incorporate multiple targeted therapies (includes umbrella trials and basket trials), in addition to traditional means, should be made. This ensures that the best decisions for phase II trial designs may be used for specific situations, to improve the efficiency of phase II trials, and in doing so we ensure the drug development process in cancer will also improve.

References

1. Tonkens R. An overview of the drug development process. Physician executive. 2005;31(3):48.
2. FDA. Step 3: Clinical Research [Internet] [updated 01/04/2018. Available from: https://www.fda.gov/patients/drug-development-process/step-3-clinical-research.
3. ICH, editor General considerations for clinical trials E8. International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use; 1997.
4. Torjesen I. Drug development: the journey of a medicine from lab to shelf. Pharmaceutical Journal. 2015.
5. Schaefer S, Kolkhof P. Failure is an option: learning from unsuccessful proof-of-concept trials. Drug Discovery Today. 2008;13(21-22):913-6.
6. Kaitin KI. Deconstructing the drug development process: the new face of innovation. Clinical Pharmacology & Therapeutics. 2010;87(3):356-61.
7. Hutchinson L, Kirk R. High drug attrition rates—where are we going wrong? Nature reviews Clinical oncology. 2011;8(4):189-90.
8. Begley C, Ellis L. Drug development: Raise standards for preclinical cancer research. Nature [Online]. 2012(483 (7391)).
9. Langrand-Escure J, Rivoirard R, Oriol M, Tinquaut F, Rancoule C, Chauvin F, et al. Quality of reporting in oncology phase II trials: A 5-year assessment through systematic review. PLoS One. 2017;12(12):e0185536.
10. An M-W, Mandrekar SJ, Branda ME, Hillman SL, Adjei AA, Pitot HC, et al. Comparison of continuous versus categorical tumor measurement–based metrics to predict overall survival in cancer treatment trials. Clinical Cancer Research. 2011;17(20):6592-9.
11. Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. Nature biotechnology. 2014;32(1):40-51.
12. Rubinstein L. Phase II design: history and evolution. Chinese Clinical Oncology. 2014;3(4):48.
13. Gray R, Manola J, Saxman S, Wright J, Dutcher J, Atkins M, et al. Phase II clinical trial design: methods in translational research from the Genitourinary Committee at the Eastern Cooperative Oncology Group. Clinical cancer research. 2006;12(7):1966-9.
14. Mandrekar SJ, Sargent DJ. Randomized phase II trials: time for a new era in clinical trial design. Journal of Thoracic Oncology. 2010;5(7):932-4.
15. Korn EL, Arbuck SG, Pluda JM, Simon R, Kaplan RS, Christian MC. Clinical trial designs for cytostatic agents: are new approaches needed? Journal of Clinical Oncology. 2001;19(1):265-72.
16. Arruebo M, Vilaboa N, Sáez-Gutierrez B, Lambea J, Tres A, Valladares M, et al. Assessment of the evolution of cancer treatment therapies. Cancers. 2011;3(3):3279-330.
17. Booth CM, Calvert AH, Giaccone G, Lobbezoo MW, Eisenhauer EA, Seymour LK, et al. Design and conduct of phase II studies of targeted anticancer therapy: recommendations from the task force on methodology for the development of innovative cancer therapies (MDICT). European Journal of Cancer. 2008;44(1):25-9.
18. Giaccone G, Herbst R, Manegold C, Scagliotti GV, Rosell R, Miller V, et al. Gefitinib in combination with gemcitabine and cisplatin in advanced non-

small-cell lung cancer: a phase III trial--INTACT 1. Journal of Clinical Oncology. 2004.
19. Demetri GD, Von Mehren M, Blanke CD, Van den Abbeele AD, Eisenberg B, Roberts PJ, et al. Efficacy and safety of imatinib mesylate in advanced gastrointestinal stromal tumors. New England Journal of Medicine. 2002;347(7):472-80.
20. Mariani L, Marubini E. Content and quality of currently published phase II cancer trials. Journal of clinical oncology. 2000;18(2):429-.
21. Rubinstein L, Crowley J, Ivy P, LeBlanc M, Sargent D. Randomized Phase II Designs. Clinical Cancer Research. 2009;15(6):1883-90.
22. Brown S, Gregory W, Twelves C, Buyse M, Collinson F, Parmar M, et al. Designing phase II trials in cancer: a systematic review and guidance. British journal of cancer. 2011;105(2):194-9.
23. Therasse P, Arbuck SG, Eisenhauer EA, Wanders J, Kaplan RS, Rubinstein L, et al. New guidelines to evaluate the response to treatment in solid tumors. Journal of the National Cancer Institute. 2000;92(3):205-16.
24. Abou-Alfa GK, Schwartz L, Ricci S, Amadori D, Santoro A, Figer A, et al. Phase II study of sorafenib in patients with advanced hepatocellular carcinoma. Journal of clinical oncology. 2006;24(26):4293-300.
25. Ratain MJ, Eisen T, Stadler WM, Flaherty KT, Kaye SB, Rosner GL, et al. Phase II placebo-controlled randomized discontinuation trial of sorafenib in patients with metastatic renal cell carcinoma. Journal of Clinical Oncology. 2006;24(16):2505-12.
26. Escudier B, Eisen T, Stadler WM, Szczylik C, Oudard S, Siebels M, et al. Sorafenib in advanced clear-cell renal-cell carcinoma. New England Journal of Medicine. 2007;356(2):125-34.
27. Llovet JM, Ricci S, Mazzaferro V, Hilgard P, Gane E, Blanc J-F, et al. Sorafenib in advanced hepatocellular carcinoma. New England journal of medicine. 2008;359(4):378-90.
28. Ang M-K, Tan S-B, Lim W-T. Phase II clinical trials in oncology: are we hitting the target? Expert review of anticancer therapy. 2010;10(3):427-38.
29. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). European journal of cancer. 2009;45(2):228-47.
30. Cannistra SA. Phase II trials in journal of clinical oncology. Journal of Clinical Oncology. 2009;27(19):3073-6.
31. Gan HK, Grothey A, Pond GR, Moore MJ, Siu LL, Sargent D. Randomized phase II trials: inevitable or inadvisable? Journal of Clinical Oncology. 2010;28(15):2641-7.
32. Seymour L, Ivy SP, Sargent D, Spriggs D, Baker L, Rubinstein L, et al. The design of phase II clinical trials testing cancer therapeutics: consensus recommendations from the clinical trial design task force of the national cancer institute investigational drug steering committee. Clinical Cancer Research. 2010;16(6):1764-9.
33. Delgado A, Guddati AK. Clinical endpoints in oncology-a primer. American Journal of Cancer Research. 2021;11(4):1121.
34. Dhani N, Tu D, Sargent DJ, Seymour L, Moore MJ. Alternate endpoints for screening phase II studies. Clinical Cancer Research. 2009;15(6):1873-82.
35. Grayling MJ, Dimairo M, Mander AP, Jaki TF. A review of perspectives on the use of randomization in phase II oncology trials. JNCI: Journal of the National Cancer Institute. 2019;111(12):1255-62.

36. Jung S-H. Statistical issues for design and analysis of single-arm multi-stage phase II cancer clinical trials. Contemporary clinical trials. 2015;42:9-17.
37. Buyse M. Phase III design: principles. Chinese Clinical Oncology. 2016;5(1):10-.
38. Woodcock J, LaVange LM. Master protocols to study multiple therapies, multiple diseases, or both. New England Journal of Medicine. 2017;377(1):62-70.
39. Stallard N. Optimal sample sizes for phase II clinical trials and pilot studies. Statistics in medicine. 2012;31(11-12):1031-42.
40. Bogaerts J, Sydes MR, Keat N, McConnell A, Benson A, Ho A, et al. Clinical trial designs for rare diseases: studies developed and discussed by the International Rare Cancers Initiative. European journal of cancer. 2015;51(3):271-81.
41. Chevret S. Bayesian adaptive clinical trials: a dream for statisticians only? Statistics in medicine. 2012;31(11-12):1002-13.
42. Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? Nature reviews Drug discovery. 2004;3(8):711-6.
43. Maitland ML, Hudoba C, Snider KL, Ratain MJ. Analysis of the yield of phase II combination therapy trials in medical oncology. Clinical Cancer Research. 2010;16(21):5296-302.
44. Zia MI, Siu LL, Pond GR, Chen EX. Comparison of outcomes of phase II studies and subsequent randomized control studies using identical chemotherapeutic regimens. Journal of Clinical Oncology. 2005;23(28):6982-91.
45. Shi Q, Sargent DJ. Key Statistical Concepts in Cancer Research. Clinical advances in hematology & oncology: H&O. 2015;13(3):180-5.
46. Moreno L, Pearson AD. How can attrition rates be reduced in cancer drug discovery? Expert opinion on drug discovery. 2013;8(4):363-8.
47. Harrison RK. Phase II and phase III failures: 2013–2015. Nature reviews Drug discovery. 2016;15(12):817.
48. Grellety T, Petit-Monéger A, Diallo A, Mathoulin-Pelissier S, Italiano A. Quality of reporting of phase II trials: a focus on highly ranked oncology journals. Annals of oncology. 2014;25(2):536-41.
49. Lee JJ, Feng L. Randomized phase II designs in cancer clinical trials: current status and future directions. Journal of Clinical Oncology. 2005;23(19):4450-7.
50. Agency BC. Drug Index: Provincial Health Services Authority; 2021 [Available from: http://www.bccancer.bc.ca/health-professionals/clinical-resources/cancer-drug-manual/drug-index.
51. Chemocare. Chemotherapy Drugs and Drugs often Used During Chemotherapy 2021 [Available from: http://chemocare.com/chemotherapy/drug-info/default.aspx.
52. UK CR. Cancer drugs A to Z list 2020 [Available from: https://www.cancerresearchuk.org/about-cancer/cancer-in-general/treatment/cancer-drugs/drugs.
53. Team R. RStudio: Integrated Development Environment for R. RStudio: PBC; 2020.
54. Ansell SM, Minnema MC, Johnson P, Timmerman JM, Armand P, Shipp MA, et al. Nivolumab for Relapsed/Refractory Diffuse Large B-Cell Lymphoma in Patients Ineligible for or Having Failed Autologous Transplantation: A Single-Arm, Phase II Study. Journal of Clinical Oncology. 2019;37(6):481-9.

55. Coen JJ, Zhang P, Saylor PJ, Lee CT, Wu CL, Parker W, et al. Bladder Preservation With Twice-a-Day Radiation Plus Fluorouracil/Cisplatin or Once Daily Radiation Plus Gemcitabine for Muscle-Invasive Bladder Cancer: NRG/RTOG 0712-A Randomized Phase II Trial. Journal of Clinical Oncology. 2019;37(1):44-51.
56. Ma DJ, Price KA, Moore EJ, Patel SH, Hinni ML, Garcia JJ, et al. Phase II Evaluation of Aggressive Dose De-Escalation for Adjuvant Chemoradiotherapy in Human Papillomavirus-Associated Oropharynx Squamous Cell Carcinoma. Journal of Clinical Oncology. 2019;37(22):1909-18.
57. Chung HC, Ros W, Delord JP, Perets R, Italiano A, Shapira-Frommer R, et al. Efficacy and Safety of Pembrolizumab in Previously Treated Advanced Cervical Cancer: Results From the Phase II KEYNOTE-158 Study. Journal of Clinical Oncology. 2019;37(17):1470-8.
58. Lopez-Chavez A, Thomas A, Rajan A, Raffeld M, Morrow B, Kelly R, et al. Molecular profiling and targeted therapy for advanced thoracic malignancies: a biomarker-derived, multiarm, multihistology phase II basket trial. Journal of clinical oncology. 2015;33(9):1000.
59. Freedman RA, Gelman RS, Anders CK, Melisko ME, Parsons HA, Cropp AM, et al. TBCRC 022: A Phase II Trial of Neratinib and Capecitabine for Patients With Human Epidermal Growth Factor Receptor 2-Positive Breast Cancer and Brain Metastases. Journal of Clinical Oncology. 2019;37(13):1081-9.
60. Ramchandren R, Domingo-Domenech E, Rueda A, Trneny M, Feldman TA, Lee HJ, et al. Nivolumab for Newly Diagnosed Advanced-Stage Classic Hodgkin Lymphoma: Safety and Efficacy in the Phase II CheckMate 205 Study. Journal of Clinical Oncology. 2019;37(23):1997-2007.
61. Assenat E, Pageaux GP, Thezenas S, Peron JM, Becouarn Y, Seitz JF, et al. Sorafenib alone vs. sorafenib plus GEMOX as 1st-line treatment for advanced HCC: the phase II randomised PRODIGE 10 trial. British Journal of Cancer. 2019;120(9):896-902.
62. Modest DP, Martens UM, Riera-Knorrenschild J, Greeve J, Florschutz A, Wessendorf S, et al. FOLFOXIRI Plus Panitumumab As First-Line Treatment of RAS Wild-Type Metastatic Colorectal Cancer: The Randomized, Open-Label, Phase II VOLFI Study (AIO KRK0109). Journal of Clinical Oncology. 2019;37(35):3401-11.
63. Simon R. Optimal two-stage designs for phase II clinical trials. Controlled clinical trials. 1989;10(1):1-10.
64. Fleming TR. One-sample multiple testing procedure for phase II clinical trials. Biometrics. 1982:143-51.
65. A'hern R. Sample size tables for exact single-stage phase II designs. Statistics in medicine. 2001;20(6):859-66.
66. Bompas E, Le Cesne A, Tresch-Bruneel E, Lebellec L, Laurence V, Collard O, et al. Sorafenib in patients with locally advanced and metastatic chordomas: a phase II trial of the French Sarcoma Group (GSF/GETO). Annals of Oncology. 2015;26(10):2168-73.
67. Takahashi H, Tada Y, Saotome T, Akazawa K, Ojiri H, Fushimi C, et al. Phase II Trial of Trastuzumab and Docetaxel in Patients With Human Epidermal Growth Factor Receptor 2-Positive Salivary Duct Carcinoma. Journal of Clinical Oncology. 2019;37(2):125-34.
68. Slomovitz BM, Jiang Y, Yates MS, Soliman PT, Johnston T, Nowakowski M, et al. Phase II study of everolimus and letrozole in patients with recurrent endometrial carcinoma. Journal of Clinical Oncology. 2015;33(8):930-6.

69. Massarelli E, Lin H, Ginsberg LE, Tran HT, Lee JJ, Canales JR, et al. Phase II trial of everolimus and erlotinib in patients with platinum-resistant recurrent and/or metastatic head and neck squamous cell carcinoma. Annals of Oncology. 2015;26(7):1476-80.
70. Taylor JM, Braun TM, Li Z. Comparing an experimental agent to a standard agent: relative merits of a one-arm or randomized two-arm phase II design. Clinical Trials. 2006;3(4):335-48.
71. Sharma MR, Karrison TG, Jin Y, Bies RR, Maitland ML, Stadler WM, et al. Resampling phase III data to assess phase II trial designs and endpoints. Clinical Cancer Research. 2012;18(8):2309-15.
72. Sharma MR, Gray E, Goldberg RM, Sargent DJ, Karrison TG. Resampling the N9741 trial to compare tumor dynamic versus conventional end points in randomized phase II trials. Journal of Clinical Oncology. 2015;33(1):36.
73. Fridlyand J, Kaiser LD, Fyfe G. Analysis of tumor burden versus progression-free survival for Phase II decision making. Contemporary clinical trials. 2011;32(3):446-52.
74. An M-W, Han Y, Meyers JP, Bogaerts J, Sargent DJ, Mandrekar SJ. Clinical utility of metrics based on tumor measurements in phase II trials to predict overall survival outcomes in phase III trials by using resampling methods. Journal of Clinical Oncology. 2015;33(34):4048.
75. Ayanlowo A, Redden D. Stochastically curtailed phase II clinical trials. Statistics in medicine. 2007;26(7):1462-72.
76. Cellamare M, Sambucini V. A randomized two-stage design for phase II clinical trials based on a Bayesian predictive approach. Statistics in medicine. 2015;34(6):1059-78.
77. Chen Y, Chen Z, Mori M. A new statistical decision rule for single-arm phase II oncology trials. Statistical methods in medical research. 2016;25(1):118-32.
78. Tang H, Foster NR, Grothey A, Ansell SM, Goldberg RM, Sargent DJ. Comparison of error rates in single-arm versus randomized phase II cancer clinical trials. Journal of Clinical Oncology. 2010;28(11):1936.
79. Sambucini V. Comparison of single-arm vs. randomized phase II clinical trials: a Bayesian approach. Journal of biopharmaceutical statistics. 2015;25(3):474-89.
80. Chen C, Sun L, Li C-L. Evaluation of early efficacy endpoints for proof-of-concept trials. Journal of Biopharmaceutical Statistics. 2013;23(2):413-24.
81. Preussler S, Kieser M, Kirchner M. Optimal sample size allocation and go/no-go decision rules for phase II/III programs where several phase III trials are performed. Biometrical Journal. 2019;61(2):357-78.
82. Götte H, Schüler A, Kirchner M, Kieser M. Sample size planning for phase II trials based on success probabilities for phase III. Pharmaceutical statistics. 2015;14(6):515-24.
83. Pond GR, Abbasi S. Quantitative evaluation of single-arm versus randomized phase II cancer clinical trials. Clinical Trials. 2011;8(3):260-9.
84. Marchenko O, Miller J, Parke T, Perevozskaya I, Qian J, Wang Y. Improving oncology clinical programs by use of innovative designs and comparing them via simulations. Therapeutic innovation & regulatory science. 2013;47(5):602-12.
85. Parke T, Marchenko O, Anisimov V, Ivanova A, Jennison C, Perevozskaya I, et al. Comparing oncology clinical programs by use of innovative designs and expected net present value optimization: Which

adaptive approach leads to the best result? Journal of biopharmaceutical statistics. 2017;27(3):457-76.
86. Ding M, Rosner GL, Müller P. Bayesian optimal design for phase II screening trials. Biometrics. 2008;64(3):886-94.
87. Yao T-J, Begg CB, Livingston PO. Optimal sample size for a series of pilot trials of new agents. Biometrics. 1996:992-1001.
88. Stallard N. Sample size determination for phase II clinical trials based on Bayesian decision theory. Biometrics. 1998:279-94.
89. Leung DHY, Wang YG. A Bayesian decision approach for sample size determination in phase II trials. Biometrics. 2001;57(1):309-12.
90. Kirchner M, Kieser M, Götte H, Schüler A. Utility-based optimization of phase II/III programs. Statistics in medicine. 2016;35(2):305-16.
91. Hee SW, Stallard N. Designing a series of decision-theoretic phase II trials in a small population. Statistics in medicine. 2012;31(30):4337-51.
92. Kieser M, Kirchner M, Dölger E, Götte H. Optimal planning of phase II/III programs for clinical trials with multiple endpoints. Pharmaceutical statistics. 2018;17(5):437-57.
93. Jung SH. Randomized phase II trials with a prospective control. Statistics in medicine. 2008;27(4):568-83.
94. Wang Y-G, Leung DH-Y. An optimal design for screening trials. Biometrics. 1998:243-50.
95. Sargent DJ, Taylor JM. Current issues in oncology drug development, with a focus on phase II trials. Journal of biopharmaceutical statistics. 2009;19(3):556-62.
96. Ratain MJ, Sargent DJ. Optimising the design of phase II oncology trials: the importance of randomisation. European Journal of Cancer. 2009;45(2):275-80.
97. CRUK. Cancer incidence statistics [internet].2022 [Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/incidence#heading-Zero.
98. Charles P, Giraudeau B, Dechartres A, Baron G, Ravaud P. Reporting of sample size calculation in randomised controlled trials. Bmj. 2009;338.
99. Atkinson Jr AJ, Colburn WA, DeGruttola VG, DeMets DL, Downing GJ, Hoth DF, et al. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clinical pharmacology & therapeutics. 2001;69(3):89-95.
100. Tomasz Burzykowski GM, Marc Buyse The Evaluation of Surrogate Endpoints. M.Gail KK, J.Samet, A.Tsiatis, W.Wong, editor: Springer; 2005. 408 p.
101. Kaiser LD. Tumor burden modeling versus progression-free survival for phase II decision making. Clinical Cancer Research. 2013;19(2):314-9.
102. Stone A, Wheeler C, Barge A. Improving the design of phase II trials of cytostatic anticancer agents. J Contemporary Clinical Trials. 2007;28(2):138-45.
103. McLeod C, Norman R, Litton E, Saville BR, Webb S, Snelling TL. Choosing primary endpoints for clinical trials of health care interventions. Contemporary clinical trials communications. 2019;16:100486.
104. Burzykowski T, Buyse M. Surrogate threshold effect: an alternative measure for meta-analytic surrogate endpoint validation. Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry. 2006;5(3):173-86.
105. Team RDC. R: A Language and Environment for Statistical Computing. Vienna, Austria: R foundation for Statistical Computing; 2009.

106. Alan Genz FB, Tetsuhisa Miwa, Xuefei Mi, Friedrich Leisch, Fabian Scheipl, Bjoern Bornkamp, Martin Maechler, Torsten Hothorn. Multivariate Normal and t Distributions version 1.0-6. 2017.
107. Sharma MR, Maitland ML, Ratain MJ. RECIST: no longer the sharpest tool in the oncology clinical trials toolbox—point. Cancer research. 2012;72(20):5145-9.
108. Rosenbaum S. Moments of a truncated bivariate normal distribution. Journal of the Royal Statistical Society: Series B (Methodological). 1961;23(2):405-8.
109. Mandrekar SJ, Qi Y, Hillman SL, Ziegler KLA, Reuter NF, Rowland Jr KM, et al. Endpoints in phase II trials for advanced non-small cell lung cancer. Journal of thoracic oncology. 2010;5(1):3-9.
110. Sharma MR, Stadler WM, Ratain MJ. Randomized phase II trials: a long-term investment with promising returns. Journal of the National Cancer Institute. 2011;103(14):1093-100.
111. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. Introduction to meta-analysis: John Wiley & Sons; 2021.
112. Khan I, Sarker S, Hackshaw A. Smaller sample sizes for phase II trials based on exact tests with actual error rates by trading-off their nominal levels of significance and power. British journal of cancer. 2012;107(11):1801-9.
113. U.S. Food and Drug Administration. The Drug Development Process [Available from: http://www.fda.gov/ForPatients/Approvals/Drugs/ucm405622.htm.
114. Schwartz LH, Litière S, De Vries E, Ford R, Gwyther S, Mandrekar S, et al. RECIST 1.1—Update and clarification: From the RECIST committee. European journal of cancer. 2016;62:132-7.
115. Seshan VE, Whiting K. clinfun: Clinical Trial Design and Data Analysis Functions. R package version 1.1.0: https://CRAN.R-project.org/package=clinfun; 2022.
116. Schoenfeld D. Statistical considerations for pilot studies. International Journal of Radiation Oncology* Biology* Physics. 1980;6(3):371-4.
117. Julious SA. Sample size of 12 per group rule of thumb for a pilot study. Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry. 2005;4(4):287-91.
118. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2018;68(6):394-424.
119. Gehan EA. The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. Journal of chronic diseases. 1961;13(4):346-53.
120. Bryant J, Day R. Incorporating toxicity considerations into the design of two-stage phase II clinical trials. Biometrics. 1995;51(4):1372-83.
121. Berry SM, Broglio KR, Groshen S, Berry DA. Bayesian hierarchical modeling of patient subpopulations: efficient designs of phase II oncology clinical trials. Clinical Trials. 2013;10(5):720-34.
122. Jahanshahi M, Gregg K, Davis G, Ndu A, Miller V, Vockley J, et al. The use of external controls in FDA regulatory decision making. Therapeutic Innovation & Regulatory Science. 2021;55(5):1019-35.
123. Lim J, Walley R, Yuan J, Liu J, Dabral A, Best N, et al. Minimizing patient burden through the use of historical subject-level data in innovative confirmatory

clinical trials: review of methods and opportunities. Therapeutic innovation & regulatory science. 2018;52(5):546-59.
124. Cohen J. Statistical power analysis for the behavioral sciences: Routledge; 2013.
125. Rice ME, Harris GT. Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. Law and human behavior. 2005;29(5):615-20.
126. Kwak SG, Kim JH. Central limit theorem: the cornerstone of modern statistics. Korean journal of anesthesiology. 2017;70(2):144-56.
127. Fiteni F, Westeel V, Pivot X, Borg C, Vernerey D, Bonnetain F. Endpoints in cancer clinical trials. Journal of visceral surgery. 2014;151(1):17-22.
128. Noor NM, Love SB, Isaacs T, Kaplan R, Parmar MK, Sydes MR. Uptake of the multi-arm multi-stage (MAMS) adaptive platform approach: a trial-registry review of late-phase randomised clinical trials. BMJ open. 2022;12(3):e055615.
129. Stallard N, Miller F, Day S, Hee SW, Madan J, Zohar S, et al. Determination of the optimal sample size for a clinical trial accounting for the population size. Biometrical Journal. 2017;59(4):609-25.
130. Berry SM, Connor JT, Lewis RJ. The platform trial: an efficient strategy for evaluating multiple treatments. Jama. 2015;313(16):1619-20.
131. Vaduganathan M, Greene SJ, Ambrosy AP, Gheorghiade M, Butler J. The disconnect between phase II and phase III trials of drugs for heart failure. Nature Reviews Cardiology. 2013;10(2):85-97.

Appendix A Proof of $QS = I$

The variance-covariance matrix S was confirmed by multiplying it by its' inverse, Q to give the identity matrix,I, where Q is:

$$\frac{1}{4(1-\rho^2)\gamma\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2}}\begin{pmatrix} (1-\rho^2)\gamma\frac{\sigma_2^2}{n_2} & 0 & -(1-\rho^2)\gamma\frac{\sigma_2^2}{n_2} & 0 \\ 0 & (1-\rho^2)\gamma\frac{\sigma_1^2}{n_1} & 0 & -(1-\rho^2)\gamma\frac{\sigma_1^2}{n_1} \\ -(1-\rho^2)\gamma\frac{\sigma_2^2}{n_2} & 0 & 4\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2}+(1-\rho^2)\gamma\frac{\sigma_2^2}{n_2} & -4\rho\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2} \\ 0 & -(1-\rho^2)\gamma\frac{\sigma_1^2}{n_1} & -4\rho\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2} & 4\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2}+(1-\rho^2)\gamma\frac{\sigma_1^2}{n_1} \end{pmatrix}$$

and

$$S = \begin{pmatrix} \frac{4\sigma_1^2}{n_1}+\sigma^2+\tau^2 & \sigma^2 & \sigma^2+\tau^2 & \sigma^2 \\ \sigma^2 & \frac{4\sigma_2^2}{n_2}+\sigma^2+\tau^2 & \sigma^2 & \sigma^2+\tau^2 \\ \sigma^2+\tau^2 & \sigma^2 & \sigma^2+\tau^2 & \sigma^2 \\ \sigma^2 & \sigma^2+\tau^2 & \sigma^2 & \sigma^2+\tau^2 \end{pmatrix}$$

Let $\rho = \frac{\sigma^2}{\sigma^2+\tau^2} \Rightarrow (1-\rho^2) = \frac{(\sigma^2+\tau^2)-\sigma^2}{(\sigma^2+\tau^2)^2}$ and $\gamma = \sigma^2+\tau^2$; where ρ is the correlation between the true treatment effects associated with the phase II and III trials, θ_1 andθ_2, and γ is the distribution's variance.

<u>Terms in row 1 of QS</u>:

$$QS_{11} = (1-\rho^2)\gamma\frac{\sigma_2^2}{n_2}\left(\frac{4\sigma_1^2}{n_1}+\gamma\right)-(1-\rho^2)\gamma\frac{\sigma_2^2}{n_2}$$

$$QS_{11} = 4(1-\rho^2)\gamma\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2}\times\frac{1}{4(1-\rho^2)\gamma\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2}} = 1$$

$$QS_{12} = (1-\rho^2)\gamma\frac{\sigma_2^2}{n_2}\sigma^2-(1-\rho^2)\gamma\frac{\sigma_2^2}{n_2}\sigma^2 = 0$$

$$QS_{13} = (1-\rho^2)\gamma\frac{\sigma_2^2}{n_2}(\sigma^2+\tau^2)-(1-\rho^2)\gamma\frac{\sigma_2^2}{n_2}(\sigma^2+\tau^2) = 0$$

$$QS_{14} = (1-\rho^2)\gamma\frac{\sigma_2^2}{n_2}\sigma^2-(1-\rho^2)\gamma\frac{\sigma_2^2}{n_2}\sigma^2 = 0$$

Terms in row 2 of QS:

$$QS_{21} = (1-\rho^2)\gamma\frac{\sigma_1^2}{n_1}\sigma^2 - (1-\rho^2)\gamma\frac{\sigma_1^2}{n_1}\sigma^2 = 0$$

$$QS_{22} = (1-\rho^2)\gamma\frac{\sigma_1^2}{n_1}\left(\frac{4\sigma_2^2}{n_2}+\gamma\right) - (1-\rho^2)\gamma^2\frac{\sigma_1^2}{n_1}$$

$$QS_{22} = 4(1-\rho^2)\gamma\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2}\times\frac{1}{4(1-\rho^2)\gamma\frac{\sigma_1^2\sigma_2^2}{n_1n_2}} = 1$$

$$QS_{23} = (1-\rho^2)\gamma\frac{\sigma_1^2}{n_1}\sigma^2 - (1-\rho^2)\gamma\frac{\sigma_1^2}{n_1}\sigma^2 = 0$$

$$QS_{24} = (1-\rho^2)\gamma^2\frac{\sigma_1^2}{n_1} - (1-\rho^2)\gamma^2\frac{\sigma_1^2}{n_1} = 0$$

Terms in row 3 of QS:

$$QS_{31} = -(1-\rho^2)\gamma\frac{\sigma_2^2}{n_2}\left(\frac{4\sigma_1^2}{n_1}+\gamma\right) + 4\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2}\gamma + (1-\rho^2)\gamma^2\frac{\sigma_2^2}{n_2} - 4\rho\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2}\sigma^2$$

$$QS_{31} = -4(1-\rho^2)\gamma\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2} + 4\gamma\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2} - 4\rho\sigma^2\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2}$$

$$QS_{31} = 4\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2}\rho(\rho\gamma-\sigma^2); \rho\gamma = \sigma^2$$

$$QS_{31} = 4\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2}\rho(\sigma^2-\sigma^2) = 0$$

$$QS_{32} = -(1-\rho^2)\gamma\frac{\sigma_2^2}{n_2}\sigma^2 + \left(4\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2} + (1-\rho^2)\gamma\frac{\sigma_2^2}{n_2}\right)\sigma^2 - 4\rho\gamma\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2}$$

$$QS_{32} = 4\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2}(\rho\gamma-\sigma^2); \rho\gamma = \sigma^2$$

$$QS_{32} = 4\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2}(\sigma^2-\sigma^2) = 0$$

$$QS_{33} = -(1-\rho^2)\gamma^2\frac{\sigma_2^2}{n_2} + \left(4\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2} + (1-\rho^2)\gamma\frac{\sigma_2^2}{n_2}\right)\gamma - 4\rho\sigma^2\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2}$$

$$QS_{33} = 4\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2}(\gamma-\rho\sigma^2); \gamma = \frac{\sigma^2}{\rho}$$

$$QS_{33} = 4\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2}\gamma(1-\rho^2)\times\frac{1}{4(1-\rho^2)\gamma\frac{\sigma_1^2\sigma_2^2}{n_1n_2}} = 1$$

$$QS_{34} = -(1-\rho^2)\gamma\frac{\sigma_2^2}{n_2}\sigma^2 + \left(4\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2} + (1-\rho^2)\gamma\frac{\sigma_2^2}{n_2}\right)\sigma^2 - 4\rho\gamma\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2}$$

$$QS_{34} = 4\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2}(\sigma^2-\sigma^2) = 0$$

Terms in row 4 of QS:

$$QS_{41} = -(1-\rho^2)\gamma\frac{\sigma_1^2}{n_1}\sigma^2 - 4\rho\gamma\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2} + \left(4\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2} + (1-\rho^2)\gamma\frac{\sigma_1^2}{n_1}\right)\sigma^2$$

$$QS_{41} = 4\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2}(\sigma^2 - \sigma^2) = 0$$

$$QS_{42} = -(1-\rho^2)\gamma\frac{\sigma_1^2}{n_1}\left(\frac{4\sigma_2^2}{n_2} + \gamma\right) + 4\rho\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2}\sigma^2 + \left(4\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2} + (1-\rho^2)\gamma\frac{\sigma_1^2}{n_1}\right)\gamma$$

$$QS_{42} = 4\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2}\rho(\sigma^2 - \sigma^2) = 0$$

$$QS_{43} = -(1-\rho^2)\gamma\frac{\sigma_1^2}{n_1}\sigma^2 - 4\rho\gamma\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2} + \left(4\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2} + (1-\rho^2)\gamma\frac{\sigma_1^2}{n_1}\right)\sigma^2$$

$$QS_{43} = 4\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2}(\sigma^2 - \sigma^2) = 0$$

$$QS_{44} = -(1-\rho^2)\gamma^2\frac{\sigma_1^2}{n_1} - 4\rho\gamma\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2} + \left(4\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2} + (1-\rho^2)\gamma\frac{\sigma_1^2}{n_1}\right)\gamma$$

$$QS_{44} = 4\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2}(\gamma - \rho\sigma^2); \gamma = \frac{\sigma^2}{\rho}$$

$$QS_{44} = 4\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2}\gamma(1-\rho^2) \times \frac{1}{4(1-\rho^2)\gamma\frac{\sigma_1^2\sigma_2^2}{n_1 n_2}} = 1$$

So off-diagonal terms are all 0 and the leading diagonal terms are all equal to 1; hence the variance-covariance S is verified as the inverse of Q.

Appendix B R code for the results presented in Chapter 5 – investigating the relationship between the phase II and III endpoints

```
# loading the necessary packages

library(mvtnorm)

#setting the variance of the true treatment effect of
available treatments

sigma<-seq(0.1,10,0.3)

Lsig<-length(sigma) # number of scenarios to be explored

diff1<-0.3 # clinically significant difference between the
# means of the two arms in phase II

diff2<-0.3 # clinically significant difference between the
#means of the two arms in phase III

# treatments tested in the phase II/ phase III trial

sigma1<-1 # sigma1/sigma2 part of the variance of
#difference in treatments in phase II/III

sigma2<-1 # needs the sqrt(n1) to be the variance

mu<- 0 # mean of the underlying treatment effect


alpha1<-0.05 # significance level for phase II

alpha2<-0.05 # significance level for phase III

power1<-0.8 # Power phase II

power2<-0.8 # Power phase III


get.n1 <- function(alpha1,power1,diff1,sigma1)

{# randomised phase II design with continuous outcome

  # one-sided sig. level

  n1 = round(4*sigma1^2 * (qnorm(1-alpha1) +
qnorm(power1))^2/(diff1^2)/2)*2

  n1 # total sample size of phase II trial

}
```

```
get.n2 <- function(alpha2,power2,diff2,sigma2)
{# randomised phase III design with continuous outcome
  # two-sided sig. level
  n2 = round(4*sigma2^2 * (qnorm(1-alpha2/2) +
qnorm(power2))^2/(diff2^2)/2)*2
  n2 # total sample size of phase III trial
}
#obtain phase II sample size
n1<- get.n1(alpha1,power1,diff1,sigma1)

# obtain phase III sample size
n2<-get.n2(alpha2,power2,diff2,sigma2)

# critical value for the phase II
k1<-qnorm(1-alpha1)*(2*sigma1/sqrt(n1))

# critical value for the phase III
k2<-qnorm(1-alpha2/2)*(2*sigma2/sqrt(n2))

#--------#
# tau==0 #
#--------#
tau<-0 # variance of the true response to treatment; same
#for phase II
#and phase III
# when tau =0 means the endpoint treatment effects have a
perfect correlation
# in other words they are the same endpoint in both
rho0<- matrix(nrow=Lsig, ncol=1,dimnames = list(c(1:Lsig)))
gam0<- matrix(nrow=Lsig, ncol=1,dimnames = list(c(1:Lsig)))
```

```
for (i in 1:Lsig ) {
  # using sigma and tau we can calculate the correlation
rho
  rho0[i,1]<- (sigma[i]^2)/(sigma[i]^2 +tau^2)
  gam0[i,1]<- sigma[i]^2 +tau^2
}
# empty array for S matrix
s110<- matrix(nrow=Lsig, ncol=1,dimnames = list(c(1:Lsig)))
s220<- matrix(nrow=Lsig, ncol=1,dimnames = list(c(1:Lsig)))
s120<- matrix(nrow=Lsig, ncol=1,dimnames = list(c(1:Lsig)))
s210<- matrix(nrow=Lsig, ncol=1,dimnames = list(c(1:Lsig)))

for (i in 1:Lsig)
{ #matrix The marginal distribution to calculate the
  #probability of success in phase II and III trials
  s110[i,1]<- (4*sigma1^2/n1)+ gam0[i]
  s220[i,1]<- (4*sigma2^2/n2)+ gam0[i]
  s120[i,1] = s210[i,1] <- sigma[i]^2
 }
fx1x2vc0<-  array(1:2*2*Lsig, dim=c(2,2,Lsig))
for (i in 1:Lsig){
  # f(x1,x2)
  fx1x2vc0[,,i]<-
matrix(c(s110[i],s120[i],s210[i],s220[i]) , nrow = 2, ncol=
2)
}
#probability of successful phase II and III when tau=0
probsucph2and30<- matrix(nrow=Lsig, ncol=1,dimnames =
list(c(1:Lsig)))
```

```
for (i in 1:Lsig){
  probsucph2and30[i,1] <-
pmvnorm(lower=c(k1,k2),upper=c(Inf,Inf),mean=c(mu,mu),sigma
=fx1x2vc0[,,i])
}
# probability of successful phase II
probsucph20<- matrix(nrow=Lsig, ncol=1,dimnames =
list(c(1:Lsig)))
for (i in 1:Lsig){
  probsucph20[i,1]<- pnorm(k1, mean = mu, sd =
sqrt(4*sigma1^2/n1 + sigma[i]^2 + tau^2) , lower.tail =
FALSE)
}
# total number of patients required to lead to the first
#successful phase III
totalexp.n0<- matrix(nrow=Lsig, ncol=1,dimnames =
list(c(1:Lsig)))
for (i in 1:Lsig) {
  totalexp.n0[i,1]<-(n1 + n2*probsucph20[i]) /
probsucph2and30[i]
}
# collating all the results into the dataset
alldata0 <-
cbind(gam0,sigma,rho0,probsucph20,probsucph2and30,totalexp.
n0)
alldata0<-data.frame(alldata0)
sigma<-data.frame(sigma)
names(alldata0)
names(alldata0)[names(alldata0)=="V1"] <- "gam0"
names(alldata0)[names(alldata0)=="V3"] <- "rho0"
names(alldata0)[names(alldata0)=="V4"] <- "probsucph20"
names(alldata0)[names(alldata0)=="V5"] <- "probsucph2and30"
```

```
names(alldata0)[names(alldata0)=="V6"] <- "totalexp.n0"
# the same methods used for tau=0.5, 1, 1.5 & 2
#----------#
# tau==0.5 #
#----------#
sigma<-seq(0.1,10,0.3)
tau<-0.5 # variance of the true response to treatment; same
#for phase II and phase III
# when tau =0.5 means the endpoint treatment effects have a
#perfect correlation in other words they are the same
#endpoint in both
rho<- matrix(nrow=Lsig, ncol=1,dimnames = list(c(1:Lsig)))
gam<- matrix(nrow=Lsig, ncol=1,dimnames = list(c(1:Lsig)))
for (i in 1:Lsig ) {
# using sigma and tau we can calculate the correlation rho
  rho[i,1]<- (sigma[i]^2)/(sigma[i]^2 +tau^2)
  gam[i,1]<- sigma[i]^2 +tau^2
}
s11<- matrix(nrow=Lsig, ncol=1,dimnames = list(c(1:Lsig)))
s22<- matrix(nrow=Lsig, ncol=1,dimnames = list(c(1:Lsig)))
s12<- matrix(nrow=Lsig, ncol=1,dimnames = list(c(1:Lsig)))
s21<- matrix(nrow=Lsig, ncol=1,dimnames = list(c(1:Lsig)))
for (i in 1:Lsig)
{#matrix The marginal distribution to calculate the
# probability of success in phase II and III trials
  s11[i,1]<- (4*sigma1^2/n1)+ gam[i]
  s22[i,1]<- (4*sigma2^2/n2)+ gam[i]
  s12[i,1] = s21[i,1] <- sigma[i]^2
}
fx1x2vc<-  array(1:2*2*Lsig, dim=c(2,2,Lsig))
for (i in 1:Lsig){
```

```
# f(x1,x2)

  fx1x2vc[,,i]<-  matrix(c(s11[i],s12[i],s21[i],s22[i]) ,
nrow = 2, ncol= 2)

}

#probability of successful phase II and III when tau=0.5

probsucph2and3<- matrix(nrow=Lsig, ncol=1,dimnames =
list(c(1:Lsig)))

for (i in 1:Lsig){

  probsucph2and3[i,1] <-
pmvnorm(lower=c(k1,k2),upper=c(Inf,Inf),mean=c(mu,mu),sigma
=fx1x2vc[,,i])

}

# probability of successful phase II

probsucph2<- matrix(nrow=Lsig, ncol=1,dimnames =
list(c(1:Lsig)))

for (i in 1:Lsig){

# total number of patients required to lead to the first
successful phase III

  probsucph2[i,1]<- pnorm(k1, mean = mu, sd =
sqrt(4*sigma1^2/n1 + sigma[i]^2 + tau^2) , lower.tail =
FALSE)

}

totalexp.n<- matrix(nrow=Lsig, ncol=1,dimnames =
list(c(1:Lsig)))

for (i in 1:Lsig) {

  totalexp.n[i,1]<-(n1 + n2*probsucph2[i]) /
probsucph2and3[i]

}

# collating results into the dataset

alldata <-
cbind(gam,sigma,rho,probsucph2,probsucph2and3,totalexp.n)

alldata<-data.frame(alldata)

sigma<-data.frame(sigma)
```

```
names(alldata)
names(alldata)[names(alldata)=="V1"] <- "gam"
names(alldata)[names(alldata)=="V3"] <- "rho"
names(alldata)[names(alldata)=="V4"] <- "probsucph2"
names(alldata)[names(alldata)=="V5"] <- "probsucph2and3"
names(alldata)[names(alldata)=="V6"] <- "totalexp.n"
#--------#
# tau==1 #
#--------#
sigma<-seq(0.1,10,0.3)
tau<-1 # variance of the true response to treatment; same
#for phase II and phase III
#when tau =1 means the endpoint treatment effects have a
#perfect correlation in other words they are the same
endpoint in both
rho1<- matrix(nrow=Lsig, ncol=1,dimnames = list(c(1:Lsig)))
gam1<- matrix(nrow=Lsig, ncol=1,dimnames = list(c(1:Lsig)))
for (i in 1:Lsig ) {
# using sigma and tau we can calculate the correlation rho
  rho1[i,1]<- (sigma[i]^2)/(sigma[i]^2 +tau^2)
  gam1[i,1]<- sigma[i]^2 +tau^2
}
s111<- matrix(nrow=Lsig, ncol=1,dimnames = list(c(1:Lsig)))
s221<- matrix(nrow=Lsig, ncol=1,dimnames = list(c(1:Lsig)))
s121<- matrix(nrow=Lsig, ncol=1,dimnames = list(c(1:Lsig)))
s211<- matrix(nrow=Lsig, ncol=1,dimnames = list(c(1:Lsig)))

for (i in 1:Lsig)
{#matrix - the marginal distribution to calculate the
  # probability of success in phase II and III trials
  s111[i,1]<- (4*sigma1^2/n1)+ gam1[i]
```

```
  s221[i,1]<- (4*sigma2^2/n2)+ gam1[i]
  s121[i,1] = s211[i,1] <- sigma[i]^2
}
fx1x2vc1<-  array(1:2*2*Lsig, dim=c(2,2,Lsig))
for (i in 1:Lsig){
# f(x1,x2)
  fx1x2vc1[,,i]<-
matrix(c(s111[i],s121[i],s211[i],s221[i]) , nrow = 2, ncol=
2)
}
#probability of successful phase II and III when tau=1
probsucph2and31<- matrix(nrow=Lsig, ncol=1,dimnames =
list(c(1:Lsig)))
for (i in 1:Lsig){
  probsucph2and31[i,1] <-
pmvnorm(lower=c(k1,k2),upper=c(Inf,Inf),mean=c(mu,mu),sigma
=fx1x2vc1[,,i])
}
# probability of successful phase II
probsucph21<- matrix(nrow=Lsig, ncol=1,dimnames =
list(c(1:Lsig)))
for (i in 1:Lsig){
  probsucph21[i,1]<- pnorm(k1, mean = mu, sd =
sqrt(4*sigma1^2/n1 + sigma[i]^2 + tau^2) , lower.tail =
FALSE)
}
# total number of patients required to lead to the first
#successful phase
totalexp.n1<- matrix(nrow=Lsig, ncol=1,dimnames =
list(c(1:Lsig)))
for (i in 1:Lsig) {
  totalexp.n1[i,1]<-(n1 + n2*probsucph21[i]) /
probsucph2and31[i]
```

```
}
# collating all the results into the dataset
alldata1 <-
cbind(gam1,sigma,rho1,probsucph21,probsucph2and31,totalexp.
n1)
alldata1<-data.frame(alldata1)
sigma<-data.frame(sigma)
names(alldata1)
names(alldata1)[names(alldata1)=="V1"] <- "gam1"
names(alldata1)[names(alldata1)=="V3"] <- "rho1"
names(alldata1)[names(alldata1)=="V4"] <- "probsucph21"
names(alldata1)[names(alldata1)=="V5"] <- "probsucph2and31"
names(alldata1)[names(alldata1)=="V6"] <- "totalexp.n1"
#----------#
# tau==1.5 #
#----------#
sigma<-seq(0.1,10,0.3)
tau<-1.5 # variance of the true response to treatment; same
#for phase II and phase III when tau =1.5 means the
endpoint #treatment effects have a perfect correlation in
other #words they are the same endpoint in both
rho1.5<- matrix(nrow=Lsig, ncol=1,dimnames =
list(c(1:Lsig)))
gam1.5<- matrix(nrow=Lsig, ncol=1,dimnames =
list(c(1:Lsig)))
for (i in 1:Lsig ) {
# using sigma and tau we can calculate the correlation rho
  rho1.5[i,1]<- (sigma[i]^2)/(sigma[i]^2 +tau^2)
  gam1.5[i,1]<- sigma[i]^2 +tau^2
}
s111.5<- matrix(nrow=Lsig, ncol=1,dimnames =
list(c(1:Lsig)))
```

```
s221.5<- matrix(nrow=Lsig, ncol=1,dimnames =
list(c(1:Lsig)))

s121.5<- matrix(nrow=Lsig, ncol=1,dimnames =
list(c(1:Lsig)))

s211.5<- matrix(nrow=Lsig, ncol=1,dimnames =
list(c(1:Lsig)))


for (i in 1:Lsig)

{ #matrix the marginal distribution to calculate the

  # probability of success in phase II and III trials

  s111.5[i,1]<- (4*sigma1^2/n1)+ gam1.5[i]

  s221.5[i,1]<- (4*sigma2^2/n2)+ gam1.5[i]

  s121.5[i,1] = s211.5[i,1] <- sigma[i]^2

}

fx1x2vc1.5<-  array(1:2*2*Lsig, dim=c(2,2,Lsig))

for (i in 1:Lsig){

# f(x1,x2)

  fx1x2vc1.5[,,i]<-
matrix(c(s111.5[i],s121.5[i],s211.5[i],s221.5[i]) , nrow =
2, ncol= 2)


}

#probability of successful phase II and III when tau=1.5

probsucph2and31.5<- matrix(nrow=Lsig, ncol=1,dimnames =
list(c(1:Lsig)))

for (i in 1:Lsig){

  probsucph2and31.5[i,1] <-
pmvnorm(lower=c(k1,k2),upper=c(Inf,Inf),mean=c(mu,mu),sigma
=fx1x2vc1.5[,,i])

}

# probability of successful phase II
```

```
probsucph21.5<- matrix(nrow=Lsig, ncol=1,dimnames =
list(c(1:Lsig)))

for (i in 1:Lsig){

  probsucph21.5[i,1]<- pnorm(k1, mean = mu, sd =
sqrt(4*sigma1^2/n1 + sigma[i]^2 + tau^2) , lower.tail =
FALSE)

}

# total number of patients required to lead to the first
#successful phase

totalexp.n1.5<- matrix(nrow=Lsig, ncol=1,dimnames =
list(c(1:Lsig)))

for (i in 1:Lsig) {

  totalexp.n1.5[i,1]<-(n1 + n2*probsucph21.5[i]) /
probsucph2and31.5[i]

}

# collating all the results into the dataset

alldata1.5 <-
cbind(gam1.5,sigma,rho1.5,probsucph21.5,probsucph2and31.5,t
otalexp.n1.5)

alldata1.5<-data.frame(alldata1.5)

sigma<-data.frame(sigma)

names(alldata1.5)

names(alldata1.5)[names(alldata1.5)=="V1"] <- "gam1.5"

names(alldata1.5)[names(alldata1.5)=="V3"] <- "rho1.5"

names(alldata1.5)[names(alldata1.5)=="V4"] <-
"probsucph21.5"

names(alldata1.5)[names(alldata1.5)=="V5"] <-
"probsucph2and31.5"

names(alldata1.5)[names(alldata1.5)=="V6"] <-
"totalexp.n1.5"

#--------#

# tau==2 #

#--------#
```

```
sigma<-seq(0.1,10,0.3)
tau<-2 # variance of the true response to treatment; same
#for phase II and phase III when tau =0 means the endpoint
#treatment effects have a perfect correlation in other
#words they are the same endpoint in both
rho2<- matrix(nrow=Lsig, ncol=1,dimnames = list(c(1:Lsig)))
gam2<- matrix(nrow=Lsig, ncol=1,dimnames = list(c(1:Lsig)))

for (i in 1:Lsig ) {
# using sigma and tau we can calculate the correlation rho
  rho2[i,1]<- (sigma[i]^2)/(sigma[i]^2 +tau^2)
  gam2[i,1]<- sigma[i]^2 +tau^2
}
s112<- matrix(nrow=Lsig, ncol=1,dimnames = list(c(1:Lsig)))
s222<- matrix(nrow=Lsig, ncol=1,dimnames = list(c(1:Lsig)))
s122<- matrix(nrow=Lsig, ncol=1,dimnames = list(c(1:Lsig)))
s212<- matrix(nrow=Lsig, ncol=1,dimnames = list(c(1:Lsig)))

for (i in 1:Lsig)
{#matrix The marginal distribution to calculate the
# probability of success in phase II and III #trials
  s112[i,1]<- (4*sigma1^2/n1)+ gam2[i]
  s222[i,1]<- (4*sigma2^2/n2)+ gam2[i]
  s122[i,1] = s212[i,1] <- sigma[i]^2
}
fx1x2vc2<-  array(1:2*2*Lsig, dim=c(2,2,Lsig))
for (i in 1:Lsig){
# f(x1,x2)
  fx1x2vc2[,,i]<-
matrix(c(s112[i],s122[i],s212[i],s222[i]) , nrow = 2, ncol=
2)
```

```
}
#probability of successful phase II and III when tau=0
probsucph2and32<- matrix(nrow=Lsig, ncol=1,dimnames =
list(c(1:Lsig)))
for (i in 1:Lsig){
  probsucph2and32[i,1] <-
pmvnorm(lower=c(k1,k2),upper=c(Inf,Inf),mean=c(mu,mu),sigma
=fx1x2vc2[,,i])
}
# probability of successful phase II
probsucph22<- matrix(nrow=Lsig, ncol=1,dimnames =
list(c(1:Lsig)))
for (i in 1:Lsig){
  probsucph22[i,1]<- pnorm(k1, mean = mu, sd =
sqrt(4*sigma1^2/n1 + sigma[i]^2 + tau^2) , lower.tail =
FALSE)
}
# total number of patients required to lead to the first
successful phase
totalexp.n2<- matrix(nrow=Lsig, ncol=1,dimnames =
list(c(1:Lsig)))
for (i in 1:Lsig) {
  totalexp.n2[i,1]<-(n1 + n2*probsucph22[i]) /
probsucph2and32[i]
}
# collating all the results into the dataset
alldata2 <-
cbind(gam2,sigma,rho2,probsucph22,probsucph2and32,totalexp.
n2)
alldata2<-data.frame(alldata2)
sigma<-data.frame(sigma)
names(alldata2)
names(alldata2)[names(alldata2)=="V1"] <- "gam2"
```

```
names(alldata2)[names(alldata2)=="V3"] <- "rho2"
names(alldata2)[names(alldata2)=="V4"] <- "probsucph22"
names(alldata2)[names(alldata2)=="V5"] <- "probsucph2and32"
names(alldata2)[names(alldata2)=="V6"] <- "totalexp.n2"

############################################################
#plots for the results
#same endpoint

# plot of sigma and number of patients required
plot(c(0,10),c(900,14000),
xlab=expression(italic(sigma)),ylab= "Number of Patients
Per Successful Phase III", main= "Same Endpoints in Phase
II and III Trials", type='n',family='serif')
lines(alldata0$sigma,alldata0$totalexp.n0, col = "red",
lty=1)

#plot comparing the probability of successes of phase II
only and phase II and III trials when both trials use the
same endpoints (tau=0, 0.5, 1, 1.5, 2)
plot(c(0,10),c(0,0.5), xlab=expression(italic(sigma)),ylab=
"Probability of Success", main= "Same Endpoints in Phase II
and III Trials", type='n',family='serif')
lines(alldata0$sigma,alldata0$probsucph20, col = "red",
lty=1)
lines(alldata0$sigma,alldata0$probsucph2and30, col =
"blue", lty=1)
legend("bottomright", legend=c("Phase II","Phase II and
III"),
       col=c("red", "blue"), lty=c(1,1))
# tau=0.5
plot(c(0,10),c(0,0.55),
xlab=expression(italic(sigma)),ylab= "Probability of
```

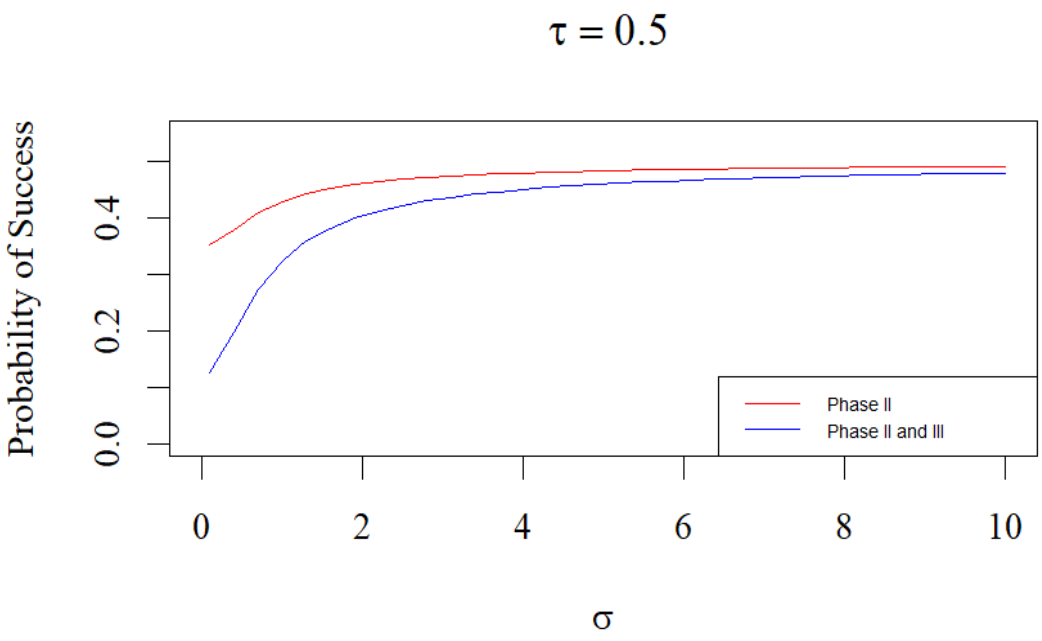
```
Success",  main = bquote(~ tau == 0.5),
type='n',family='serif')

lines(alldata$sigma,alldata$probsucph2, col = "red", lty=1)

lines(alldata$sigma,alldata$probsucph2and3, col = "blue",
lty=1)

legend("bottomright", legend=c("Phase II","Phase II and
III"),

       col=c("red", "blue"), lty=c(1,1))

# tau=1

plot(c(0,10),c(0,0.55),
xlab=expression(italic(sigma)),ylab= "Probability of
Success",  main = bquote(~ tau == 1),
type='n',family='serif')

lines(alldata1$sigma,alldata1$probsucph21, col = "red",
lty=1)

lines(alldata1$sigma,alldata1$probsucph2and31, col =
"blue", lty=1)

legend("bottomright", legend=c("Phase II","Phase II and
III"),

       col=c("red", "blue"), lty=c(1,1))

# tau=1.5

plot(c(0,10),c(0,0.55),
xlab=expression(italic(sigma)),ylab= "Probability of
Success",  main = bquote(~ tau == 1.5),
type='n',family='serif')

lines(alldata1.5$sigma,alldata1.5$probsucph21.5, col =
"red", lty=1)

lines(alldata1.5$sigma,alldata1.5$probsucph2and31.5, col =
"blue", lty=1)

legend("bottomright", legend=c("Phase II","Phase II and
III"),

       col=c("red", "blue"), lty=c(1,1))

# tau=2
```

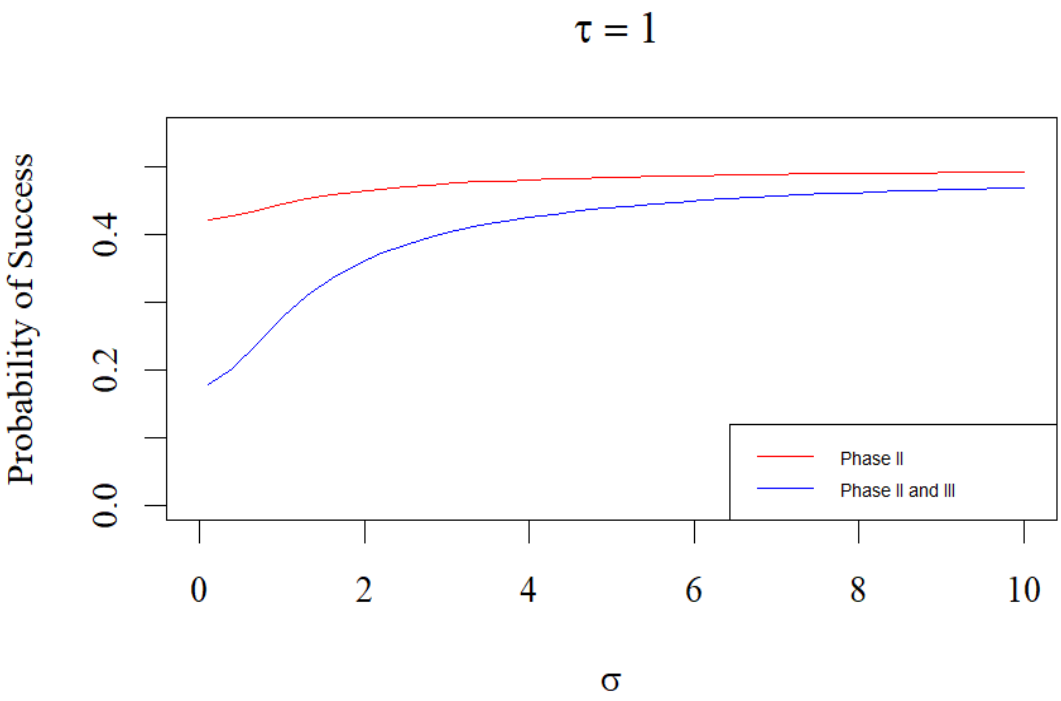
```
plot(c(0,10),c(0,0.55),
xlab=expression(italic(sigma)),ylab= "Probability of
Success",  main = bquote(~ tau == 2),
type='n',family='serif')

lines(alldata2$sigma,alldata2$probsucph22, col = "red",
lty=1)

lines(alldata2$sigma,alldata2$probsucph2and32, col =
"blue", lty=1)

legend("bottomright", legend=c("Phase II","Phase II and
III"),

       col=c("red", "blue"), lty=c(1,1))

##########################################################

#different endpoint scenario


#plot of sigma and number of patients required; comparison
of the effect of having the same endpoint in both trials
and different endpoints

plot(c(0,10),c(900,4000),
xlab=expression(italic(sigma)),ylab= "Number of Patients
Per Successful Phase III", type='n',family='serif')

lines(alldata$sigma,alldata$totalexp.n, col = "orange",
lty=1)

lines(alldata1$sigma,alldata1$totalexp.n1, col = "blue",
lty=1)

lines(alldata1.5$sigma,alldata1.5$totalexp.n1.5, col =
"purple", lty=1)

lines(alldata2$sigma,alldata2$totalexp.n2, col = "black",
lty=1)

legend("topright", legend = c(#expression(paste(tau, " = ",
0)), expression(paste(tau, " = ", 0.5)),
expression(paste(tau, " = ", 1)),
expression(paste(tau, " = ", 1.5)),
expression(paste(tau, " = ", 2))),col=c("orange",
"blue","purple","black"),lty = c(1,1,1,1))
```

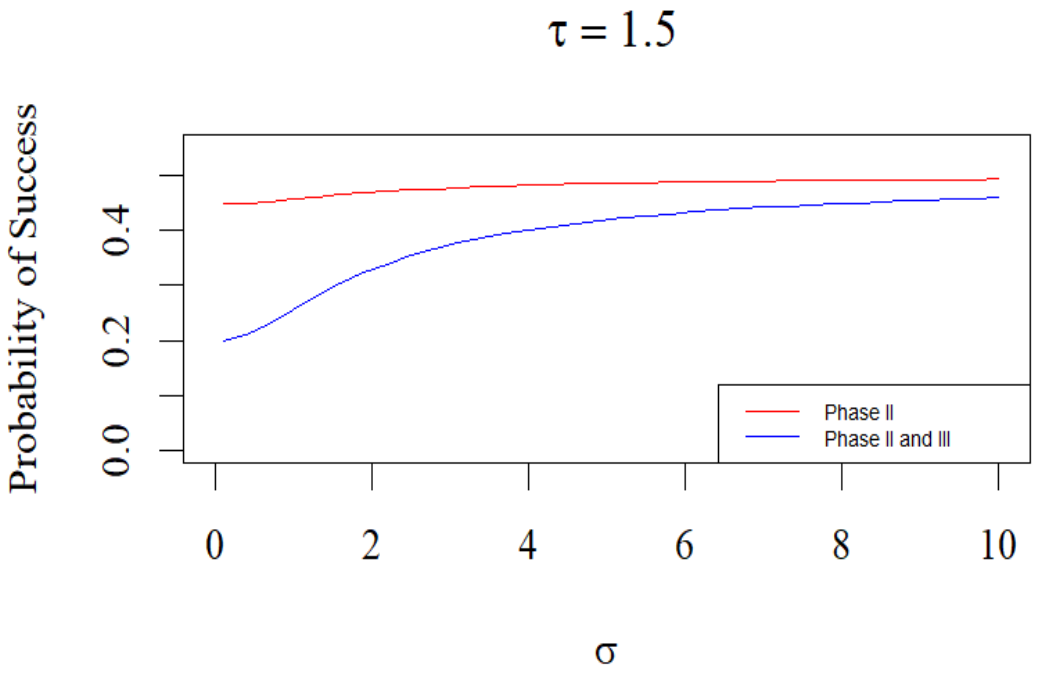
```
#plot showing the effect of correlation between endpoints
and phase II efficiency

plot(c(0,1),c(900,4000), xlab=expression(italic(rho)),ylab=
"Number of Patients Per Successful Phase III", main=
"Different Endpoint in Phase II and III
trials",type='n',family='serif')

lines(alldata$rho,alldata$totalexp.n, col = "orange",
lty=1)

lines(alldata1$rho1,alldata1$totalexp.n1, col = "blue",
lty=1)

lines(alldata1.5$rho1.5,alldata1.5$totalexp.n1.5, col =
"purple", lty=1)

lines(alldata2$rho2,alldata2$totalexp.n2, col = "black",
lty=1)

legend("topright", legend = c(expression(paste(tau, " = ",
0.5)), expression(paste(tau, " = ", 1)),
expression(paste(tau, " = ", 1.5)),
expression(paste(tau, " = ", 2))),col=c("orange",
"blue","purple","black"),lty = c(1,1,1,1))

############################################################

###################### CONTOUR PLOTS ######################

############################################################

# Contour plot NOT NEEDED for TAU=0 as RHO=1 so

# expected number of patients required to reach the first
successful phase III trial

# only ranges with sigma

# when tau>0 rho and sigma both have a range of values and
therefore both affect the

# expected number of patients required to lead to the first
successful phase III trial

install.packages("plotly") # ggplot need for this

library(plotly)

p <- plot_ly(data = alldata, x=~sigma,y=~rho,
z=~totalexp.n, type = "contour", colorscale='Viridis')
```

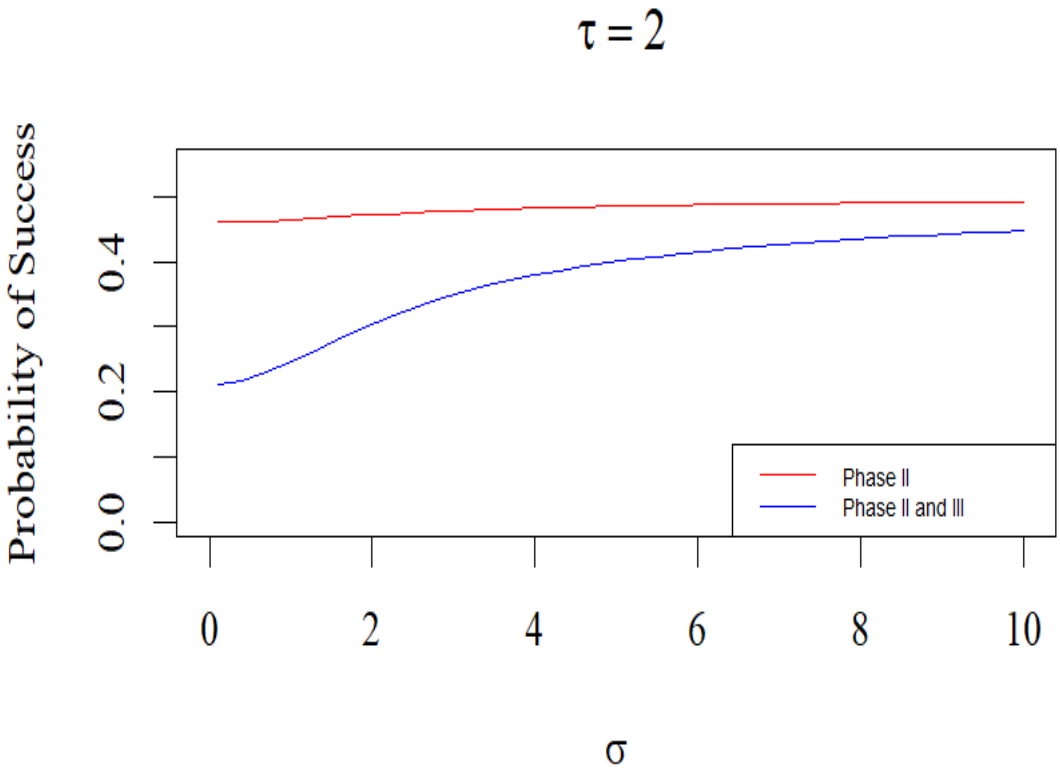
```
p # tau=0.5


p1 <- plot_ly(data = alldata1, x=~sigma,y=~rho1,
z=~totalexp.n1, type = "contour", colorscale='Viridis')
p1 # tau=1


p1.5 <- plot_ly(data = alldata1.5, x=~sigma,y=~rho1.5,
z=~totalexp.n1.5, type = "contour", colorscale='Viridis')
p1.5 # tau=1.5


p2 <- plot_ly(data = alldata2, x=~sigma,y=~rho2,
z=~totalexp.n2, type = "contour", colorscale='Viridis')
p2 # tau=2
```

Appendix C Graphs for the probability of success in phase II and probability of success in phase II and III trials for different values of τ






τ = 1.5
Probability of Success
0.0
0.2
0.4
0
2
4
6
8
10
σ
Phase II
Phase II and III


τ = 2
Probability of Success
0.0
0.2
0.4
0
2
4
6
8
10
σ
Phase II
Phase II and III

Appendix D R code for the design evaluations

D.1 Randomised single-stage design evaluations

```
# loading needed packages

library("plyr")

# fixing a seed to obtain reproducible results

set.seed(2212022)

#true treatment effects

truedelta<- rnorm(10000, 0, 1)

logdelta<- truedelta*(pi/sqrt(3))

hist(truedelta,prob=TRUE, breaks=20, main = "True Treatment
Effect", xlab = expression(Delta))

curve(dnorm(x, mean(truedelta), sd(truedelta)), add=TRUE,
col="darkblue", lwd=2)


p.1<- 0.25 # control arm's probability of success


# corresponding true phase II treatment effects

c <- log(p.1/(1-p.1)) # log odds ratio of the control arm's
#probability of success

p.2 <- ((exp(truedelta*pi/sqrt(3))*p.1
))/((p.1*(exp(truedelta*pi/sqrt(3))  - 1)  + 1))

# the inverse log of the normally distributed treatment
effects

# how the true delta is correlated to p.2

hist(truedelta,prob=TRUE, breaks=20, main = "True Treatment
Effect", xlab = expression(mu[2]))

plot(truedelta, p.2, xlim = c(-4,4), ylim = c(0,1),xlab =
expression(mu[2]), ylab =expression(p[2]) )

Deltatrue<- data.frame(p.2,truedelta)
```

```
get.sample.sizeph2 <- function(alpha,power,p.1, delta2)
  # function to do standard sample size calculation for a
#1:1 randomised ph2 trial
  # N.B. alpha is one-sided error rate and sample size is
#total for two arms
{
 round((2*(p.1*(1 - p.1) + (p.1+delta2)*(1 -
(p.1+delta2)))*((qnorm(1-alpha)+qnorm(power))^2)/
(delta2^2)/2)*2)
}

get.sample.sizeph3 <- function(sigma,alpha,power,delta1)
  # function to do standard sample size calculation
#assuming sigma = 1
  # N.B. alpha is two-sided error rate and sample size is
#total for two arms
{
  round((2 * 2 *sigma*sigma*(qnorm(1-
(alpha/2))+qnorm(power))^2/(delta1^2))/2)*2
}
alpha=0.05 # type I error for phase II and III trials
power=0.8 # 1-type II error for phase II and III trials
delta1=0.3 # delta1 is the clinically significant
#difference we wish to detect in phase III
delta2=0.2# delta2 is the csd we wish to detect in phase II
sigma<-1 # sd for underlying treatment effect

n3<-get.sample.sizeph3(sigma, alpha, power, delta1)
# sample size of the phase III trial depends on alpha,
#power, sigma and delta1 which is the treatment effect we
#wish to detect
```

```
n2<-get.sample.sizeph2(alpha,power, p.1, delta2)

# sample size of the phase II trial depends on alpha,
#power, p.1 and delta1 which is the treatment effect we
#wish to detect

mu1<-0 # control arm mean

sigma1<- sigma2 <-1 # variance of phase II and III trial
#responses

l<-15

control3<- rnorm(10000,mu1,sigma1) # patients available for
control arm in phase III

mu2<-list()

experimental3<-list()

ph3samplecontrol <- list()

ph3sampleexp <- list()

ph3test<-list()

pv3<-list()

control2<-rbinom(10000,1,p.1) # patients available for
control arm in phase II

ph2delta<-c()

experimental2<-list()

ph2samplecontrol<-list()

ph2sampleexp<-list()

total<- list()

ph2test<-list()

pv2<-c()

ph2tabexp<-list()

ph2tab<-list()

ph2tabcont<-list()

ph2vecs<-vector()

ph2grp<-vector()

total<- list()
```

```
ph2test<-list()
pv2<-c()
ph2tabexp<-list()
ph2tabcont<-list()
x<-c()
n<-c()
phase_III_ready <- FALSE
phase_II_ready<- TRUE
r2<- lapply(1:l, matrix, data=NA,nrow=1, ncol=1)
r3<- lapply(1:l, matrix, data=NA,nrow=1, ncol=1)
result<-lapply(1:l, matrix, data=NA,nrow=1, ncol=1)
td2<- lapply(1:l, matrix, data=NA,nrow=1, ncol=1)
td3<- lapply(1:l, matrix, data=NA,nrow=1, ncol=1)

for (i in 1:l){
# loop to start the simulation
  pop <- 500000 # patients in the population
  while( pop > (n3)){
    # continues until there's enough patients for the phase
#III trial
    if(phase_III_ready & phase_II_ready){
      #run phase III trial
      mu2<- (sqrt(3)*log(((p.1*ph2delta)-
ph2delta)/(p.1*(ph2delta-1))))/pi
      # mu2 is the treatment effect tested in phase 3 and
#corresponds to p.2
      experimental3<-rnorm(10000,mu2,sigma2)
      # patients available to be entered in the
#experimental arm
      ph3samplecontrol<- sample(control3, n3/2,
replace=FALSE) # sample n2/2 patients for the phase III
```

```
      #half patients sampled in control arm
      ph3sampleexp<- sample(experimental3, n3/2,
replace=FALSE) # sample the remaining patients for the
#phase III
      #half patients sampled in experimental arm
      ph3test<-t.test(ph3samplecontrol,ph3sampleexp,
var.equal= TRUE, mu= 0, alternative = "t")
#running the two-sample t-test:
    pv3 <- ph3test$p.value # pvalue extracted
      if (pv3 <= alpha & mu2>0){
        # ph3 successful if pvalue<=alpha and the treatment
#effect>0
        td3[[i]]<- c(td3[[i]], mu2)
        r3[[i]]<- c(r3[[i]],1)
        pop <- pop - n3
      }
      else {
        # ph3 unsuccessful
        td3[[i]]<- c(td3[[i]], mu2)
        r3[[i]]<-c(r3[[i]],0)
        pop<- pop - n3
      }

      phase_III_ready <- FALSE

    } else {
      if(pop> n3 + n2) {
# enough patients in population to run both ph. II and III
        phase_II_ready<- TRUE
        #phase II simulations
        ph2delta<- sample(p.2,1,replace = FALSE)
```

```
        experimental2<-rbinom(10000,1,ph2delta)
        ph2samplecontrol <- sample(control2, n2/2,
replace=FALSE) # sampled 1/2 n in control
        ph2sampleexp <- sample(experimental2,n2/2,
replace=FALSE) # sampled 1/2 n in experimental
        ph2samplecontrol<- factor(ph2samplecontrol,c(0,1),
labels = c('fail','success'))
        ph2sampleexp<- factor(ph2sampleexp,c(0,1), labels =
c('fail','success'))
        ph2tabexp<-table(ph2sampleexp)
        ph2tabcont<-table(ph2samplecontrol)
        x<-c(ph2tabcont[2],ph2tabexp[2])
        n<-c(ph2tabcont[2]+ph2tabcont[1],
ph2tabexp[2]+ph2tabexp[1])
        ph2test<-prop.test(x, n, alternative = c("l"),
conf.level = (1- alpha), correct = FALSE) # run ph2 trial
        pv2 <- ph2test$p.value # extract the p-value
        if (pv2 <= alpha & !is.na(pv2)){
          #ph2 success
          td2[[i]]<- c(td2[[i]], ph2delta)
          r2[[i]] <- c(r2[[i]],1)
          phase_III_ready <- TRUE
          pop <- pop - n2 # take away patients used
        }
        else {
          # phase II failure
          td2[[i]]<- c(td2[[i]], ph2delta)
          td3[[i]]<- c(td3[[i]], NA)
          r2[[i]] <- c(r2[[i]],0)
          phase_III_ready <- FALSE
          r3[[i]] <- c(r3[[i]], 3)
```

```
          pop <- pop - n2
        }
      } else {break
}
    }
   }
  }
tot.ph2<-list()
tot.ph3succ<- list()
tot.ph3fail<- list()
tot.ALL<- list()
tot.ph3<-list()
tot.ph2succ<-list()
tot.ph2fail<- list()
b <-list()
r3.f<-list()
r2.f<-list()
a<-list()
for (i in 1:l){
  # collating results
  result[[i]]<- cbind(r2[[i]],r3[[i]], td2[[i]], td3[[i]])
  result[[i]]<-data.frame(result[[i]])
  result[[i]]<-result[[i]][-1,]
  names(result[[i]])<- c("ph2out", "ph3out", "p2", "m2")
  tot.ph2[[i]]<-nrow(result[[i]])
  r3.f[[i]] <- factor(result[[i]][,2], levels = c(1,0,3),
labels = c("success", "fail", "not run"))
  r2.f[[i]] <- factor(result[[i]][,1], levels =
c(1,0),labels = c("success", "fail"))
```

```
  b[[i]]<- data.frame(table(r3.f[[i]]))
  a[[i]]<- data.frame(table(r2.f[[i]])
  tot.ph3fail[[i]]<- b[[i]][2,2] # no. phase 3 fails
  tot.ph3succ[[i]]<-b[[i]][1,2] # no. phase 3 successes
  tot.ph2fail[[i]]<- a[[i]][2,2] # no. phase 2 fails
  tot.ph2succ[[i]]<-a[[i]][1,2] # no. phase 2 success
  tot.ph3[[i]]<- b[[i]][1,2] + b[[i]][2,2] # total phase
#III run
# collating results
ph2<-unlist(tot.ph2)
ph3<-unlist(tot.ph3)
ph2fail<-unlist(tot.ph2fail)
ph3fail<-unlist(tot.ph3fail)
ph2succ<- unlist(tot.ph2succ)
ph3succ<- unlist(tot.ph3succ)
randss.res<- data.frame(ph2)
randss.res<-cbind(randss.res, ph3)
randss.res<-cbind(randss.res, ph2fail)
randss.res<-cbind(randss.res, ph3fail)
randss.res<-cbind(randss.res, ph2succ)
randss.res<-cbind(randss.res, ph3succ)
# descriptive stats
summary(randss.res$ph3succ) # phase 3 success
summary(randss.res$ph3fail) # phase 3 fails
summary(randss.res$ph3) # total phase 3
summary(randss.res$ph2) # total phase 2
summary(randss.res$ph2succ) # phase 2 success
summary(randss.res$ph2fail) # phase 2 fails
```

D.2 Randomised two-stage (Jung's) design evaluations

```
# loading needed packages

library("plyr")

# fixing a seed to obtain reproducible results

set.seed(2512022)

#true treatment effects

truedelta<- rnorm(10000, 0, 1)

# histogram of true treatment effect

hist(truedelta,prob=TRUE, breaks=20, main = "True Treatment
Effect", xlab = expression(Delta))

curve(dnorm(x, mean(truedelta), sd(truedelta)), add=TRUE,
col="darkblue", lwd=2)

p.1<- 0.25 # control arm's probability of success

# corresponding true phase II treatment effects

c <- log(p.1/(1-p.1)) # log odds ratio of the control arm's
#probability of success

p.2 <- ((exp(truedelta*pi/sqrt(3))*p.1
))/((p.1*(exp(truedelta*pi/sqrt(3))  - 1)  + 1))  # the
#inverse log of the normally distributed treatment effects

# how the true delta is correlated to p.2

hist(truedelta,prob=TRUE, breaks=20, main = "True Treatment
Effect", xlab = expression(mu[2]))

plot(truedelta, p.2, xlim = c(-4,4), ylim = c(0,1),xlab =
expression(mu[2]), ylab =expression(p[2]) )

Deltatrue<- data.frame(p.2,truedelta)


get.sample.sizeph3 <- function(sigma,alpha,power,delta1)

  # function to do standard sample size calculation
#assuming sigma = 1

  # N.B. alpha is two-sided error rate and sample size is
#total for two arms

{
```

```
  round((2 * 2 *sigma*sigma*(qnorm(1-
(alpha/2))+qnorm(power))^2/(delta1^2))/2)*2
}

alpha=0.05 # type I error for phase III trials
power=0.8 # 1-type II error for phase III trials
delta1=0.3 # delta1 is the clinically significant
#difference we wish to detect in phase III
delta2=0.2 #delta2 is the csd we wish to detect in phase II
sigma<-1 # sd for underlying treatment effect

n3<-get.sample.sizeph3(sigma, alpha, power, delta1)
# sample size of the phase III trial depends on alpha,
#power, sigma and delta1 which is the treatment effect we
#wish to detect
# get the sample size for the two-stage phase II trials
# sample size and critical values obtained from Fortran
code for Jung's design
# OC for this design is alpha=0.05, beta=0.2, delta2=0.2
and p0=0.25 therefore p1=0.45
n2 <- 112 # total sample size for a randomised two-stage
design
n2.1 <- 13 # number of patients in each arm in the first
stage
n2.2<- 43 # number of patients in each arm in the second
stage
a1 <- 2 # critical value in the 1st stage
a <- 8 # cumulative critical value of 1st and 2nd stage
mu1<-0 # control arm mean
sigma1<- sigma2 <-1
l=15
```

```
control3<- rnorm(10000,mu1,sigma1) # patients available for
control arm in phase III

mu2<-list()

experimental3<-list()

ph3samplecontrol <- list()

ph3sampleexp <- list()

ph3test<-list()

pv3<-list()

control2<-rbinom(n=10000,size=1, prob=p.1) # patients
available for control arm in phase II

ph2delta<-c()

experimental2<-list()

ph2sample.1<-list()

ph2sample.2<-list()

ph2sample2.1<- list()

ph2sample2.2<- list()

ph2table2.2<- list()

ph2table2.1<- list()

ph2vecs<-vector()

ph2grp<-vector()

total<- list()

ph2test<-list()

pv2<-c()

ph2table<-list()

response.1<- list()

response.2<- list()

response<- list()

ph2table.1<- list()

ph2table.2<- list()

x1<- list()
```

```
y1<- list()
obs_a1<- list()
x2<- list()
y2<- list()
obs_y<- list()
obs_x<- list()
obs_a<- list()

x<-c()
n<-c()
phase_III_ready <- FALSE
phase_II_2ndstage_ready<- FALSE
phase_II_1ststage_ready<- TRUE

r2.1<- lapply(1:l, matrix, data=NA,nrow=1, ncol=1)
r2.2<- lapply(1:l, matrix, data=NA,nrow=1, ncol=1)

r3<- lapply(1:l, matrix, data=NA,nrow=1, ncol=1)
result<-lapply(1:l, matrix, data=NA,nrow=1, ncol=1)
td2.1<- lapply(1:l, matrix, data=NA,nrow=1, ncol=1)
td2.2<- lapply(1:l, matrix, data=NA,nrow=1, ncol=1)
td3<- lapply(1:l, matrix, data=NA,nrow=1, ncol=1)
rr<- lapply(1:l, matrix, data=NA,nrow=1, ncol=1)

for (i in 1:l){
  pop <- 500000 #population available
  while( pop > (n3)){
    # run loop until there isn't enough patients to run
#phase III trial
```

```
    if(phase_III_ready & phase_II_1ststage_ready &
phase_II_2ndstage_ready){
      #run phase III trial
      mu2<- (sqrt(3)*log(((p.1*ph2delta)-
ph2delta)/(p.1*(ph2delta-1))))/pi
      # mu2 is the treatment effect tested in phase 3 and
#corresponds to p.2
      experimental3<-rnorm(10000,mu2,sigma2)
      # patients available to be entered in the
#experimental arm
      ph3samplecontrol<- sample(control3, n3/2,
replace=FALSE) # sample n3/2 patients for the phase III
      #half patients sampled in control arm
      ph3sampleexp<- sample(experimental3, n3/2,
replace=FALSE) # sample the remaining patients for the
#phase III
      #half patients sampled in experimental arm
      ph3test<-t.test(ph3samplecontrol,ph3sampleexp,
var.equal= TRUE, mu= 0, alternative = "t")
#running the two-sample t-test:
       pv3 <- ph3test$p.value
      if (pv3 <= alpha & mu2>0){
        #successful phase III
        td3[[i]]<- c(td3[[i]], mu2)
        r3[[i]]<- c(r3[[i]],1)
        td2.1[[i]]<- c(td2.1[[i]], ph2delta)
        pop <- pop - n3
      }
      else {
        # failed phase III
        td3[[i]]<- c(td3[[i]], mu2)
        r3[[i]]<-c(r3[[i]],0)
```

```
        td2.1[[i]]<- c(td2.1[[i]], ph2delta)
        pop<- pop - n3 # take out patients from the
population
      }
      phase_III_ready <- FALSE
      } else {
      if(pop> n3 + n2) {
        phase_II_1ststage_ready<- TRUE
        #first stage phase II simulations
        ph2delta<- sample(p.2,1,replace = FALSE)
        experimental2<-rbinom(10000,1,ph2delta)
        ph2sample.1<- sample(experimental2, n2.1,
replace=FALSE) # sample the patients for the phase II
        # trial from the experimental patients distribution
        ph2sample.1 <- factor(ph2sample.1,levels = c(0,1),
labels = c("0", "1"))
        ph2table.1<-table(ph2sample.1)
        x1<-table(ph2sample.1)["1"]
        ph2sample.2<- sample(control2, n2.1, replace=FALSE)
# sample the patients for the phase II
        # trial from the control patients distribution
        ph2sample.2 <- factor(ph2sample.2,levels = c(0,1),
labels = c("0", "1"))
        ph2table.2<-table(ph2sample.2)
        y1<-table(ph2sample.2)["1"]
        obs_a1<- x1 - y1
        if (obs_a1 >= a1){
          # successful 1st stage of phase II
          phase_II_2ndstage_ready <- TRUE
          pop <- pop - (n2.1*2)
        }
```

```
        else {
          # failed 1st stage phase II
          td2.2[[i]]<- c(td2.2[[i]], NA)
          td3[[i]]<- c(td3[[i]], NA)
          r2.2[[i]] <- c(r2.2[[i]],3)
# 2nd stage didn't run
          r2.1[[i]] <- c(r2.1[[i]],0)
# 1st stage unsuccessful
          phase_II_2ndstage_ready<- FALSE
          phase_III_ready <- FALSE
          phase_II_1ststage_ready<- TRUE
          r3[[i]] <- c(r3[[i]], 3) # phase III didn't run
          td2.1[[i]]<- c(td2.1[[i]], ph2delta)
          pop <- pop - (n2.1*2) # take out patients used
#first stage
        }
        if (phase_II_2ndstage_ready == TRUE) {
          # run 2nd stage
          ph2sample2.1<- sample(experimental2, n2.2,
replace=FALSE) # sample the patients for the phase II
          # trial from the experimental patients
distribution
          ph2sample2.1 <- factor(ph2sample2.1,levels =
c(0,1), labels = c("0", "1"))
          ph2table2.1<-table(ph2sample2.1)
          x2<-table(ph2sample2.1)["1"]
          obs_x<- x1 + x2
         ph2sample2.2<- sample(control2, n2.2,
replace=FALSE) # sample the patients for the phase II
          # trial from the control patients distribution
```

```
          ph2sample2.2 <- factor(ph2sample2.2,levels =
c(0,1), labels = c("0", "1"))
          ph2table2.2<-table(ph2sample2.2)
          y2<-table(ph2sample2.2)["1"]
          obs_y<- y1 + y2
          obs_a<-obs_x - obs_y
         if (obs_a >= a){
            # phase II success
            td2.2[[i]]<- c(td2.2[[i]], ph2delta)
            r2.1[[i]] <- c(r2.1[[i]],1)
            r2.2[[i]] <- c(r2.2[[i]],1)
            phase_III_ready <- TRUE
            pop <- pop - (n2.2*2)# take out patients used
in second stage
          }
         else {
            # failed second stage
            td2.2[[i]]<- c(td2.2[[i]], ph2delta)
            td3[[i]]<- c(td3[[i]], NA)
            r2.1[[i]] <- c(r2.1[[i]],1)
            r2.2[[i]] <- c(r2.2[[i]],0)
            phase_III_ready <- FALSE
            r3[[i]] <- c(r3[[i]], 3)
            td2.1[[i]]<- c(td2.1[[i]], ph2delta)
            phase_II_1ststage_ready<- TRUE
            phase_II_2ndstage_ready<- FALSE
            pop <- pop - (n2.2*2)# take out patients used
in second stage
          }
        }
      }
```

```
else {break
}
   }
  }
}
tot.ph2<-list()
tot.ph3succ<- list()
tot.ph3fail<- list()
tot.ALL<- list()
tot.ph3<-list()
tot.ph2succ<-list()
tot.ph2fail<- list()
b <-list()
r3.f<-list()
r2.f.1<-list()
r2.f.2<-list()
a<-list()
c<- list()
#collating results
for (i in 1:l){
  result[[i]]<- cbind(r2.1[[i]], r2.2[[i]],r3[[i]],
td2.1[[i]], td2.2[[i]], td3[[i]])
  result[[i]]<-data.frame(result[[i]])
  result[[i]]<-result[[i]][-1,]
  names(result[[i]])<- c("ph2out.1","ph2out.2", "ph3out",
"p2", "p2", "m2")
  tot.ph2[[i]]<-nrow(result[[i]])
  r3.f[[i]] <- factor(result[[i]][,3], levels = c(1,0,3),
labels = c("success", "fail", "not run"))
  r2.f.2[[i]] <- factor(result[[i]][,2], levels = c(1,0,3),
labels = c("success", "fail", "not run"))
```

```
  r2.f.1[[i]] <- factor(result[[i]][,1], levels =
c(1,0),labels = c("success", "fail"))

  c[[i]]<- data.frame(table(r3.f[[i]])) # table of
successful/unsuccessful phase III

  a[[i]]<- data.frame(table(r2.f.1[[i]])) # table of
successful/unsuccessful phase II st1

  b[[i]]<- data.frame(table(r2.f.2[[i]])) # table of
successful/unsuccessful phase II st2

tot.ph3fail[[i]]<- c[[i]][2,2]# no. phase 3 fails

tot.ph3succ[[i]]<-c[[i]][1,2] # no. phase 3 successes

tot.ph2fail[[i]]<- a[[i]][2,2] + b[[i]][2,2]# no. phase 2
#fails

tot.ph2succ[[i]]<-b[[i]][1,2]# no. phase 2 success

tot.ph3[[i]]<- c[[i]][1,2] + c[[i]][2,2]# total phase
#III run

# collating results

ph2<-unlist(tot.ph2)

ph3<-unlist(tot.ph3)

ph2fail<-unlist(tot.ph2fail)

ph3fail<-unlist(tot.ph3fail)

ph2succ<- unlist(tot.ph2succ)

ph3succ<- unlist(tot.ph3succ)


randts.res<- data.frame(ph2)

randts.res<-cbind(randts.res, ph3)

randts.res<-cbind(randts.res, ph2fail)

randts.res<-cbind(randts.res, ph3fail)

randts.res<-cbind(randts.res, ph2succ)

randts.res<-cbind(randts.res, ph3succ)

# descriptive stats

summary(randts.res$ph3succ) # phase 3 success
```

```
summary(randts.res$ph3fail) # phase 3 fails
summary(randts.res$ph3) # total phase 3
summary(randts.res$ph2) # total phase 2
summary(randts.res$ph2succ) # phase 2 success
summary(randts.res$ph2fail) # phase 2 fails
```

D.3 Single-arm single-stage (A'hern's) design evaluations

```
# loading needed packages

library("plyr")

# fixing a seed to obtain reproducible results

set.seed(1212022)

#true treatment effects

truedelta<- rnorm(10000, 0, 1)

logdelta<- truedelta*(pi/sqrt(3))

# histogram of distribution of treatment effect of
#available treatments

hist(truedelta,prob=TRUE, breaks=20, main = "True Treatment
Effect", xlab = expression(Delta))

curve(dnorm(x, mean(truedelta), sd(truedelta)), add=TRUE,
col="darkblue", lwd=2)

p.1<- 0.25 # control arm's probability of success

# corresponding true phase II treatment effects

c <- log(p.1/(1-p.1)) # log odds ratio of the control arm's
#probability of success

p.2 <- ((exp(truedelta*pi/sqrt(3))*p.1
))/((p.1*(exp(truedelta*pi/sqrt(3))  - 1)  + 1))  # the
#inverse log of the normally distributed treatment effects


# how the true delta is correlated to p.2

hist(truedelta,prob=TRUE, breaks=20, main = "True Treatment
Effect", xlab = expression(mu[2]))

plot(truedelta, p.2, xlim = c(-4,4), ylim = c(0,1),xlab =
expression(mu[2]), ylab =expression(p[2]) )

Deltatrue<- data.frame(p.2,truedelta)

get.sample.sizeph3 <- function(sigma,alpha,power,delta1)

  # function to do standard sample size calculation
#assuming sigma = 1
```

```
  # N.B. alpha is two-sided error rate and sample size is
#total for two arms
{
 round((2 * 2 *sigma*sigma*(qnorm(1-
(alpha/2))+qnorm(power))^2/(delta1^2))/2)*2
}
alpha=0.05 # type I error for phase II and III trials
power=0.8 # 1-type II error for phase II and III trials
delta1=0.3 # delta1 is the clinically significant
difference we wish to detect in phase III
delta2=0.2# delta2 is the csd we wish to detect in phase II
sigma<-1 # sd for underlying treatment effect

n3<-get.sample.sizeph3(sigma, alpha, power, delta1)
# sample size of the phase III trial depends on alpha,
power, sigma
# and delta1 which is the treatment effect we wish to
detect
# sample size of the phase II trial depends on alpha,
power, p.1
# and delta1 which is the treatment effect we wish to
detect.
# A'hern exact single-stage design is used to obtain the
sample size
# and the success cut off value.
# from the paper n2=36 and r =14
n2<-36 # total sample size
r<-14 # more than or equal to 14 then phase II is a success
mu1<-0 # control arm mean
sigma1<- sigma2 <-1 # variance of phase II and III trial
#responses
l<-15
```

```
control3<- rnorm(10000,mu1,sigma1) # patients available for
control arm in phase III

mu2<-list()

experimental3<-list()

ph3samplecontrol <- list()

ph3sampleexp <- list()

ph3test<-list()

pv3<-list()

control2<-rbinom(10000,1,p.1) # patients available for
control arm in phase II

ph2delta<-c()

experimental2<-list()

ph2samplecontrol<-list()

ph2sampleexp<-list()

total<- list()

ph2test<-list()

pv2<-c()

ph2tabexp<-list()

ph2tab<-list()

ph2tabcont<-list()

response<- list()


phase_III_ready <- FALSE

phase_II_ready<- TRUE


r2<- lapply(1:l, matrix, data=NA,nrow=1, ncol=1)

r3<- lapply(1:l, matrix, data=NA,nrow=1, ncol=1)

result<-lapply(1:l, matrix, data=NA,nrow=1, ncol=1)

td2<- lapply(1:l, matrix, data=NA,nrow=1, ncol=1)

td3<- lapply(1:l, matrix, data=NA,nrow=1, ncol=1)
```

```
rr<- lapply(1:l, matrix, data=NA,nrow=1, ncol=1)
for (i in 1:l){
  pop <- 500000
# population of patients available
  while( pop > (n3)){
    #loop continues until not enough patients for phase III
    if(phase_III_ready & phase_II_ready){
      #run phase III trial
      mu2<- (sqrt(3)*log(((p.1*ph2delta)-
ph2delta)/(p.1*(ph2delta-1))))/pi
      experimental3<-rnorm(10000,mu2,sigma2)


      ph3samplecontrol<- sample(control3, n3/2,
replace=FALSE) # sample n2/2 patients for the phase III
      ph3sampleexp<- sample(experimental3, n3/2,
replace=FALSE) # sample the remaining patients for the
phase III
      ph3test<-t.test(ph3samplecontrol,ph3sampleexp,
var.equal= TRUE, mu= 0, alternative = "t")
#running the two-sample t-test:
      pv3 <- ph3test$p.value


      if (pv3 <= alpha & mu2>0){
# phase III successful
        td3[[i]]<- c(td3[[i]], mu2)
        r3[[i]]<- c(r3[[i]],1)
        pop <- pop - n3 #take out used patients in trial
      }
      else {
# phase III fail
        td3[[i]]<- c(td3[[i]], mu2)
```

```
        r3[[i]]<-c(r3[[i]],0)
        pop<- pop - n3 #take out used patients in trial
      }
      phase_III_ready <- FALSE
    } else {
# if there's enough patients for phase II and III run loop
      if(pop> n3 + n2) {
        phase_II_ready<- TRUE
        #phase II simulations
        ph2delta<- sample(p.2,1,replace = FALSE)
        experimental2<-rbinom(10000,1,ph2delta)
       ph2sample<- sample(experimental2, n2, replace=FALSE)
# sample the patients for the phase II
       # trial from the experimental patients distribution
       ph2sample <- factor(ph2sample,levels = c(0,1),
labels = c("0", "1"))
       ph2expdata<-data.frame(ph2sample)
       names(ph2expdata)[names(ph2expdata) ==
"ph2sampleexp"] <- "ph2"
       ph2table<-table(ph2sample)
       response<-table(ph2sample)["1"]
        if (response >= r){
          td2[[i]]<- c(td2[[i]], ph2delta)
          r2[[i]] <- c(r2[[i]],1)
          phase_III_ready <- TRUE
          pop <- pop - n2 #take out used patients in trial
        }
        else {
          td2[[i]]<- c(td2[[i]], ph2delta)
          td3[[i]]<- c(td3[[i]], NA)
          r2[[i]] <- c(r2[[i]],0)
```

```
          phase_III_ready <- FALSE
          r3[[i]] <- c(r3[[i]], 3)
          pop <- pop - n2 #take out used patients in trial
        }
      } else {break
}
    }
  }
}
tot.ph2<-list()
tot.ph3succ<- list()
tot.ph3fail<- list()
tot.ALL<- list()
tot.ph3<-list()
tot.ph2succ<-list()
tot.ph2fail<- list()
b <-list()
r3.f<-list()
r2.f<-list()
a<-list()
# collating results
for (i in 1:l){
  result[[i]]<- cbind(r2[[i]],r3[[i]], td2[[i]], td3[[i]])
  result[[i]]<-data.frame(result[[i]])
  result[[i]]<-result[[i]][-1,]
  names(result[[i]])<- c("ph2out", "ph3out", "p2", "m2")
  tot.ph2[[i]]<-nrow(result[[i]])
  r3.f[[i]] <- factor(result[[i]][,2], levels = c(1,0,3),
labels = c("success", "fail", "not run"))
```

```
  r2.f[[i]] <- factor(result[[i]][,1], levels =
c(1,0),labels = c("success", "fail"))


  b[[i]]<- data.frame(table(r3.f[[i]]))
  a[[i]]<- data.frame(table(r2.f[[i]]))
  tot.ph3fail[[i]]<- b[[i]][2,2] # no. phase 3 fails
  tot.ph3succ[[i]]<-b[[i]][1,2] # no. phase 3 successes
  tot.ph2fail[[i]]<- a[[i]][2,2] # no. phase 2 fails
  tot.ph2succ[[i]]<-a[[i]][1,2] # no. phase 2 success
  tot.ph3[[i]]<- b[[i]][1,2] + b[[i]][2,2] # total phase
#III run
# collating results
ph2<-unlist(tot.ph2)
ph3<-unlist(tot.ph3)
ph2fail<-unlist(tot.ph2fail)
ph3fail<-unlist(tot.ph3fail)
ph2succ<- unlist(tot.ph2succ)
ph3succ<- unlist(tot.ph3succ)
singless.res<- data.frame(ph2)
singless.res<-cbind(singless.res, ph3)
singless.res<-cbind(singless.res, ph2fail)
singless.res<-cbind(singless.res, ph3fail)
singless.res<-cbind(singless.res, ph2succ)
singless.res<-cbind(singless.res, ph3succ)
# descriptive stats
summary(singless.res$ph3succ) # phase 3 success
summary(singless.res$ph3fail) # phase 3 fails
summary(singless.res$ph3) # total phase 3
summary(singless.res$ph2) # total phase 2
summary(singless.res$ph2succ) # phase 2 success
```

```
summary(singless.res$ph2fail) # phase 2 fails
```

D.4 Single-arm two-stage design (Simon's minimax) design

```
# loading needed packages

library("plyr")

install.packages("clinfun")

library('clinfun')

# fixing a seed to obtain reproducible results

set.seed(2622022)

#true treatment effects

truedelta<- rnorm(10000, 0, 1)

# histogram of distribution for treatment effects available

hist(truedelta,prob=TRUE, breaks=20, main = "True Treatment
Effect", xlab = expression(Delta))

curve(dnorm(x, mean(truedelta), sd(truedelta)), add=TRUE,
col="darkblue", lwd=2)

p.1<- 0.25 # control arm's probability of success

# corresponding true phase II treatment effects

c <- log(p.1/(1-p.1)) # log odds ratio of the control arm's
#probability of success

p.2 <- ((exp(truedelta*pi/sqrt(3))*p.1
))/((p.1*(exp(truedelta*pi/sqrt(3))  - 1)  + 1))  # the
#inverse log of the normally distributed treatment effects

# how the true delta is correlated to p.2

hist(truedelta,prob=TRUE, breaks=20, main = "True Treatment
Effect", xlab = expression(mu[2]))

plot(truedelta, p.2, xlim = c(-4,4), ylim = c(0,1),xlab =
expression(mu[2]), ylab =expression(p[2]) )

Deltatrue<- data.frame(p.2,truedelta)

get.sample.sizeph3 <- function(sigma,alpha,power,delta1)

  # function to do standard sample size calculation
assuming sigma = 1

  # N.B. alpha is one-sided error rate and sample size is
total for two arms
```

```
{
  round((2 * 2 *sigma*sigma*(qnorm(1-
(alpha/2))+qnorm(power))^2/(delta1^2))/2)*2
}
alpha=0.05 # type I error for phase II and III trials
power=0.8 # 1-type II error for phase II and III trials
delta1=0.3 # delta1 is the clinically significant
#difference we wish to detect in phase III
delta2=0.2# delta2 is the csd we wish to detect in phase II
sigma<-1 # sd for underlying treatment effect

n3<-get.sample.sizeph3(sigma, alpha, power, delta1)
# sample size of the phase III trial depends on alpha,
power, sigma
# and delta1 which is the treatment effect we wish to
detect

# get the sample size for the two-stage phase II trials
minimax
ph2simon(p.1, p.1+delta2, 0.05, 0.2)
n2<-36 # total
n2.1<-17 # stage 1 sample size 17
r1<- 4 #need to see more than 4 responses to proceed to 2nd
#stage
# if we see more than 4 responses proceed to stage 2 and
proceed to phase III with the novel treatment
n2.2<-36 - 17 # stage 2 sample size is 41-17=24
r<-13 #need to see more than 14 responses in total
mu1<-0 # control arm mean
sigma1<- sigma2 <-1
l=15
```

```
control3<- rnorm(10000,mu1,sigma1) # patients available for
#control arm in phase III
mu2<-list()
experimental3<-list()
ph3samplecontrol <- list()
ph3sampleexp <- list()
ph3test<-list()
pv3<-list()
control2<-p.1 # patients available for control arm in phase
II
ph2delta<-c()
experimental2<-list()
ph2sample.1<-list()
ph2sample.2<-list()
ph2vecs<-vector()
ph2grp<-vector()
total<- list()
ph2test<-list()
ph2table<-list()
response.1<- list()
response.2<- list()
response<- list()
phase_III_ready <- FALSE
phase_II_2ndstage_ready<- FALSE
phase_II_1ststage_ready<- TRUE

r2.1<- lapply(1:l, matrix, data=NA,nrow=1, ncol=1)
r2.2<- lapply(1:l, matrix, data=NA,nrow=1, ncol=1)
r3<- lapply(1:l, matrix, data=NA,nrow=1, ncol=1)
result<-lapply(1:l, matrix, data=NA,nrow=1, ncol=1)
```

```
td2.1<- lapply(1:l, matrix, data=NA,nrow=1, ncol=1)
td2.2<- lapply(1:l, matrix, data=NA,nrow=1, ncol=1)
td3<- lapply(1:l, matrix, data=NA,nrow=1, ncol=1)
rr<- lapply(1:l, matrix, data=NA,nrow=1, ncol=1)

for (i in 1:l){
# population of patients available
  pop <- 500000
  while( pop > (n3)){
 #loop only runs if enough patients exist for phase III
    if(phase_III_ready & phase_II_1ststage_ready &
phase_II_2ndstage_ready){
      #run phase III trial
      mu2<- (sqrt(3)*log(((p.1*ph2delta)-
ph2delta)/(p.1*(ph2delta-1))))/pi
  experimental3<-rnorm(10000,mu2,sigma2)
  ph3samplecontrol<- sample(control3, n3/2, replace=FALSE)
# sample n2/2 patients for the phase III
      ph3sampleexp<- sample(experimental3, n3/2,
replace=FALSE) # sample the remaining patients for the
#phase III
ph3test<-t.test(ph3samplecontrol,ph3sampleexp, var.equal=
TRUE, mu= 0, alternative = "t") #running the two-sample t-
test:
      pv3 <- ph3test$p.value # extract p-value
      if (pv3 <= alpha & mu2>0){
        #successful phase III
        td3[[i]]<- c(td3[[i]], mu2)
        r3[[i]]<- c(r3[[i]],1)
        td2.1[[i]]<- c(td2.1[[i]], ph2delta)
        pop <- pop - n3 #take out used patients in trial
      }
```

```
      else {
        # failed phase III
        td3[[i]]<- c(td3[[i]], mu2)
        r3[[i]]<-c(r3[[i]],0)
        td2.1[[i]]<- c(td2.1[[i]], ph2delta)
        pop<- pop - n3 #take out used patients in trial
      }
      phase_III_ready <- FALSE
    } else {
# run if enough patient in population for phase II and III
      if(pop> n3 + n2) {
        phase_II_1ststage_ready<- TRUE
        #first stage phase II simulations
        ph2delta<- sample(p.2,1,replace = FALSE)
        experimental2<-rbinom(10000,1,ph2delta)
        ph2sample.1<- sample(experimental2, n2.1,
replace=FALSE) # sample the patients for the phase II
        # trial from the experimental patients distribution
        ph2sample.1 <- factor(ph2sample.1,levels = c(0,1),
labels = c("0", "1"))
        ph2expdata.1<-data.frame(ph2sample.1)
        names(ph2expdata.1)[names(ph2expdata.1) ==
"ph2sampleexp.1"] <- "ph2.1"
        ph2table.1<-table(ph2sample.1)
        response.1<-table(ph2sample.1)["1"]
        if (response.1 > r1){
          # successful 1st stage of phase II
          phase_II_2ndstage_ready <- TRUE
    pop <- pop - n2.1#take out used patients in first stage
        }
        else {
```

```
          # failed 1st stage phase II
          td2.2[[i]]<- c(td2.2[[i]], NA)
          td3[[i]]<- c(td3[[i]], NA)
          r2.2[[i]] <- c(r2.2[[i]],3) # 2nd stage didn't
run
          r2.1[[i]] <- c(r2.1[[i]],0) # 1st stage
unsuccessful
          phase_II_2ndstage_ready<- FALSE
          phase_III_ready <- FALSE
          phase_II_1ststage_ready<- TRUE
       r3[[i]] <- c(r3[[i]], 3) # phase III also didn't run
          td2.1[[i]]<- c(td2.1[[i]], ph2delta)
pop <- pop - n2.1 #take out used patients in first stage
      }
        if (phase_II_2ndstage_ready == TRUE) {
          # run 2nd stage
          ph2sample.2<- sample(experimental2, n2.2,
replace=FALSE) # sample the patients for the phase II
       # trial from the experimental patients distribution
          ph2sample.2<- factor(ph2sample.2,levels = c(0,1),
labels = c("0", "1"))
          ph2table.2<-table(ph2sample.2)
          response.2<-table(ph2sample.2)["1"]
          response<- response.1 + response.2
        if (response > r){
          # phase II success
          td2.2[[i]]<- c(td2.2[[i]], ph2delta)
          r2.1[[i]] <- c(r2.1[[i]],1)
          r2.2[[i]] <- c(r2.2[[i]],1)
          phase_III_ready <- TRUE
     pop <- pop - n2.2 #take out used patients in 2nd stage
```

```
        }
        else {
          # failed second stage
          td2.2[[i]]<- c(td2.2[[i]], ph2delta)
          td3[[i]]<- c(td3[[i]], NA)
          r2.1[[i]] <- c(r2.1[[i]],1)
          r2.2[[i]] <- c(r2.2[[i]],0)
          phase_III_ready <- FALSE
          r3[[i]] <- c(r3[[i]], 3)
          td2.1[[i]]<- c(td2.1[[i]], ph2delta)
          phase_II_1ststage_ready<- TRUE
          phase_II_2ndstage_ready<- FALSE
     pop <- pop - n2.2#take out used patients in 2nd stage
        }
      }
    }else {break
}
  }
 }
}

tot.ph2<-list()
tot.ph3succ<- list()
tot.ph3fail<- list()
tot.ALL<- list()
tot.ph3<-list()
tot.ph2succ<-list()
tot.ph2fail<- list()
b <-list()
r3.f<-list()
```

```
r2.f.1<-list()

r2.f.2<-list()

a<-list()

c<- list()

for (i in 1:l){

  result[[i]]<- cbind(r2.1[[i]], r2.2[[i]],r3[[i]],
td2.1[[i]], td2.2[[i]], td3[[i]])

  result[[i]]<-data.frame(result[[i]])

  result[[i]]<-result[[i]][-1,]

  names(result[[i]])<- c("ph2out.1","ph2out.2", "ph3out",
"p2", "p2", "m2")

  tot.ph2[[i]]<-nrow(result[[i]])

  r3.f[[i]] <- factor(result[[i]][,3], levels = c(1,0,3),
labels = c("success", "fail", "not run"))

  r2.f.2[[i]] <- factor(result[[i]][,2], levels = c(1,0,3),
labels = c("success", "fail", "not run"))

  r2.f.1[[i]] <- factor(result[[i]][,1], levels =
c(1,0),labels = c("success", "fail"))


  c[[i]]<- data.frame(table(r3.f[[i]])) # table of
successful/unsuccessful phase III

  a[[i]]<- data.frame(table(r2.f.1[[i]])) # table of
successful/unsuccessful phase II st1

  b[[i]]<- data.frame(table(r2.f.2[[i]])) # table of
successful/unsuccessful phase II st2

  tot.ph3fail[[i]]<- c[[i]][2,2] # ph3 fail

  tot.ph3succ[[i]]<-c[[i]][1,2] # ph3 success

  tot.ph2fail[[i]]<- a[[i]][2,2] + b[[i]][2,2] # ph2 fail

  tot.ph2succ[[i]]<-b[[i]][1,2] # ph2 success

  tot.ph3[[i]]<- c[[i]][1,2] + c[[i]][2,2] # total phase 3

#collating results

ph2<-unlist(tot.ph2)
```

```
ph3<-unlist(tot.ph3)
ph2fail<-unlist(tot.ph2fail)
ph3fail<-unlist(tot.ph3fail)
ph2succ<- unlist(tot.ph2succ)
ph3succ<- unlist(tot.ph3succ)

singlets.res<- data.frame(ph2)
singlets.res<-cbind(singlets.res, ph3)
singlets.res<-cbind(singlets.res, ph2fail)
singlets.res<-cbind(singlets.res, ph3fail)
singlets.res<-cbind(singlets.res, ph2succ)
singlets.res<-cbind(singlets.res, ph3succ)
#descriptive statistics
summary(singlets.res$ph3succ) # phase 3 success
summary(singlets.res$ph3fail) # phase 3 fail
summary(singlets.res$ph3) # total phase 3
summary(singlets.res$ph2) # total phase 2
summary(singlets.res$ph2succ) # phase 2 success
summary(singlets.res$ph2fail) # phase 2 fail
```

Appendix E Sample sizes of Simon's single-arm two-stage minimax design

No of designs	alpha	beta	n2.1	r1	n2.2	r	n2
1	0.01	0.1	66	23	5	26	71
2	0.05	0.1	26	6	23	17	49
3	0.1	0.1	23	5	16	13	39
4	0.15	0.1	18	4	14	10	32
5	0.2	0.1	12	2	15	8	27
6	0.01	0.15	30	8	34	24	64
7	0.05	0.15	25	6	17	15	42
8	0.1	0.15	22	6	8	10	30
9	0.15	0.15	12	2	13	8	25
10	0.2	0.15	8	1	15	7	23
11	0.01	0.2	24	6	31	21	55
12	0.05	0.2	17	4	19	13	36
13	0.1	0.2	15	3	11	9	26
14	0.15	0.2	15	3	6	7	21
15	0.2	0.2	12	2	4	5	16
16	0.01	0.25	47	18	2	19	49
17	0.05	0.25	14	3	16	11	30
18	0.1	0.25	13	3	10	8	23
19	0.15	0.25	10	2	8	6	18
20	0.2	0.25	5	0	8	4	13
21	0.01	0.3	20	6	26	18	46
22	0.05	0.3	10	2	17	10	27
23	0.1	0.3	9	2	11	7	20
24	0.15	0.3	9	2	6	5	15
25	0.2	0.3	6	1	7	4	13
26	0.01	0.35	21	6	21	17	42
27	0.05	0.35	9	2	15	9	24
28	0.1	0.35	6	1	11	6	17
29	0.15	0.35	6	1	6	4	12
30	0.2	0.35	3	0	7	3	10
31	0.01	0.4	17	5	22	16	39
32	0.05	0.4	6	1	15	8	21
33	0.1	0.4	3	0	11	5	14
34	0.15	0.4	9	2	2	4	11
35	0.2	0.4	6	1	5	3	11
36	0.01	0.45	22	8	12	14	34
37	0.05	0.45	8	2	10	7	18
38	0.1	0.45	6	1	7	5	13
39	0.15	0.45	5	1	4	3	9
40	0.2	0.45	4	0	2	2	6
41	0.01	0.5	8	2	23	13	31
42	0.05	0.5	13	5	2	6	15
43	0.1	0.5	5	1	6	4	11

44	0.15	0.5	3	0	5	3	8
45	0.2	0.5	3	0	3	2	6
46	0.01	0.55	10	3	18	12	28
47	0.05	0.55	7	1	7	6	14
48	0.1	0.55	6	2	2	3	8
49	0.15	0.55	2	0	4	2	6
50	0.2	0.55	2	0	4	2	6
51	0.01	0.6	10	3	15	11	25
52	0.05	0.6	5	1	7	5	12
53	0.1	0.6	2	0	6	3	8
54	0.15	0.6	3	0	2	2	5

Appendix F R code for sample size evaluations

F.1 Simon's two-stage single-arm design

```
# loading needed packages

library("plyr")

# fixing a seed to obtain reproducible results

set.seed(1212022)

# Single-arm phase II trials; sample size effect

# two-sided phase III

#true treatment effects

truedelta<- rnorm(10000, 0, 1)

logdelta<- truedelta*(pi/sqrt(3))

hist(truedelta,prob=TRUE, breaks=20, main = "True Treatment
Effect", xlab = expression(Delta))

curve(dnorm(x, mean(truedelta), sd(truedelta)), add=TRUE,
col="darkblue", lwd=2)

p.1<- 0.25 # control arm's probability of success

# corresponding true phase II treatment effects

c <- log(p.1/(1-p.1)) # log odds ratio of the control arm's
#probability of success

p.2 <- ((exp(truedelta*pi/sqrt(3))*p.1
))/((p.1*(exp(truedelta*pi/sqrt(3))  - 1)  + 1))  # the
#inverse log of the normally distributed treatment effects

# how the true delta is correlated to p.2

hist(truedelta,prob=TRUE, breaks=20, main = "True Treatment
Effect", xlab = expression(mu[2]))

plot(truedelta, p.2, xlim = c(-4,4), ylim = c(0,1),xlab =
expression(mu[2]), ylab =expression(p[2]) )

Deltatrue<- data.frame(p.2,truedelta)


pop <- 500000 # population size for each design (i.e.
#sample size)
```

```
alpha=0.05 # type I error for phase II and III trials

power=0.8 # 1-type II error for phase II and III trials

delta1=0.3 # delta1 is the clinically significant
difference we wish to detect in phase III

delta2=0.2# delta2 is the csd we wish to detect in phase II

sigma<-1 # sd for underlying treatment effect

data.sample <-
read.table("C:/Users/enada/OneDrive/Desktop/Nada Elbeltagi-
PhD 17.01.22/Write up/samplesizeSATS.csv", header=TRUE,
sep=",")# loading the sample sizes from an excel file
(these are from Appendix E)

data.sample$power<-1-data.sample$beta # calculate power

data.sample<-data.sample[!(data.sample$n2>300),] #getting
rid of any combinations where n2 is larger than 300

data.sample<-data.sample[!(data.sample$n2<=2),] #getting
rid of any combinations where n2 is larger than 300


l<- length(data.sample$alpha)

get.sample.sizeph3 <- function(sigma,alpha,power,delta1)

  # function to do standard sample size calculation
#assuming sigma = 1

  # N.B. alpha is two-sided error rate and sample size is
#total for two arms

{

  round((2 * 2 *sigma*sigma*(qnorm(1-
(alpha/2))+qnorm(power))^2/(delta1^2))/2)*2

}

n3<-get.sample.sizeph3(sigma, alpha, power, delta1)

# sample size of the phase III trial depends on alpha,
power, sigma

# and delta1 which is the treatment effect we wish to
detect
```

```
mu1<-0 # control arm mean
sigma1<- sigma2 <-1 # sd of the outcome
control3<- rnorm(10000,mu1,sigma1)
# patients available for control arm in phase III
mu2<-list()
experimental3<-list()
ph3samplecontrol <- list()
ph3sampleexp <- list()
ph3test<-list()
pv3<-list()
control2<-p.1 # patients available for control arm in phase
II
ph2delta<-c()
experimental2<-list()
ph2sample.1<-list()
ph2sample.2<-list()
ph2vecs<-vector()
ph2grp<-vector()
total<- list()
ph2test<-list()
ph2table<-list()
response.1<- list()
response.2<- list()
response<- list()
phase_III_ready <- FALSE
phase_II_2ndstage_ready<- FALSE
phase_II_1ststage_ready<- TRUE
r2.1<- lapply(1:l, matrix, data=NA,nrow=1, ncol=1)
r2.2<- lapply(1:l, matrix, data=NA,nrow=1, ncol=1)
```

```
r3<- lapply(1:l, matrix, data=NA,nrow=1, ncol=1)
result<-lapply(1:l, matrix, data=NA,nrow=1, ncol=1)
td2.1<- lapply(1:l, matrix, data=NA,nrow=1, ncol=1)
td2.2<- lapply(1:l, matrix, data=NA,nrow=1, ncol=1)
td3<- lapply(1:l, matrix, data=NA,nrow=1, ncol=1)
rr<- lapply(1:l, matrix, data=NA,nrow=1, ncol=1)

for (i in 1:l){
# population size for each sample size iteration
  pop <- 500000
  while( pop > (n3)){
# loop continues until not enough patients can enter phase
#III
    if(phase_III_ready & phase_II_1ststage_ready &
phase_II_2ndstage_ready){
      #run phase III trial
      mu2<- (sqrt(3)*log(((p.1*ph2delta)-
ph2delta)/(p.1*(ph2delta-1))))/pi
      experimental3<-rnorm(10000,mu2,sigma2)
      ph3samplecontrol<- sample(control3, n3/2,
replace=FALSE) # sample n2/2 patients for the phase III
      ph3sampleexp<- sample(experimental3, n3/2,
replace=FALSE) # sample the remaining patients for the
#phase III
       ph3test<-t.test(ph3samplecontrol,ph3sampleexp,
var.equal= TRUE, mu= 0, alternative = "t") #running the
two-sample t-test:
      pv3 <- ph3test$p.value
      if (pv3 <= alpha & mu2>0){
        #successful phase III
        td3[[i]]<- c(td3[[i]], mu2)
        r3[[i]]<- c(r3[[i]],1)
```

```
        td2.1[[i]]<- c(td2.1[[i]], ph2delta)
        pop <- pop - n3 # take out patients used in trial
      }
      else {
        # failed phase III
        td3[[i]]<- c(td3[[i]], mu2)
        r3[[i]]<-c(r3[[i]],0)
        td2.1[[i]]<- c(td2.1[[i]], ph2delta)
        pop<- pop - n3
      }
    phase_III_ready <- FALSE
    } else {
      if(pop> n3 + data.sample$n2[i]) {
        phase_II_1ststage_ready<- TRUE
        #first stage phase II simulations
        ph2delta<- sample(p.2,1,replace = FALSE)
        experimental2<-rbinom(10000,1,ph2delta)
        ph2sample.1<- sample(experimental2,
data.sample$n2.1[i], replace=FALSE) # sample the patients
for the phase II
        # trial from the experimental patients distribution
        ph2sample.1 <- factor(ph2sample.1,levels = c(0,1),
labels = c("0", "1"))
        ph2expdata.1<-data.frame(ph2sample.1)
        names(ph2expdata.1)[names(ph2expdata.1) ==
"ph2sampleexp.1"] <- "ph2.1"
        ph2table.1<-table(ph2sample.1)
        response.1<-table(ph2sample.1)["1"]
     if (response.1 > data.sample$r1[i]){
          # successful 1st stage of phase II
          phase_II_2ndstage_ready <- TRUE
```

```
          pop <- pop - data.sample$n2.1[i] #take out phase
#II patients used in stage 1
        }
        else {
          # failed 1st stage phase II
          td2.2[[i]]<- c(td2.2[[i]], NA)
          td3[[i]]<- c(td3[[i]], NA)
          r2.2[[i]] <- c(r2.2[[i]],3) # 2nd stage didn't
run
          r2.1[[i]] <- c(r2.1[[i]],0) # 1st stage
unsuccessful
          phase_II_2ndstage_ready<- FALSE
          phase_III_ready <- FALSE
          phase_II_1ststage_ready<- TRUE
          r3[[i]] <- c(r3[[i]], 3) # phase III also didn't
run
          td2.1[[i]]<- c(td2.1[[i]], ph2delta)
          pop <- pop - data.sample$n2.1[i] #take out phase
#II patients used in stage 1
        }
        if (phase_II_2ndstage_ready == TRUE) {
          # run 2nd stage
          ph2sample.2<- sample(experimental2,
data.sample$n2.2[i], replace=FALSE) # sample the patients
#for the phase II
    # trial from the experimental patients distribution
          ph2sample.2<- factor(ph2sample.2,levels = c(0,1),
labels = c("0", "1"))
          ph2table.2<-table(ph2sample.2)
          response.2<-table(ph2sample.2)["1"]
          response<- response.1 + response.2
        if (response > data.sample$r[i]){
```

```
            # phase II success
            td2.2[[i]]<- c(td2.2[[i]], ph2delta)
            r2.1[[i]] <- c(r2.1[[i]],1)
            r2.2[[i]] <- c(r2.2[[i]],1)
            phase_III_ready <- TRUE
            pop <- pop - data.sample$n2.2[i]#take out phase
#II patients used in stage 2
          }
       else {
            # failed second stage
            td2.2[[i]]<- c(td2.2[[i]], ph2delta)
            td3[[i]]<- c(td3[[i]], NA)
            r2.1[[i]] <- c(r2.1[[i]],1)
            r2.2[[i]] <- c(r2.2[[i]],0)
            phase_III_ready <- FALSE
            r3[[i]] <- c(r3[[i]], 3)
            td2.1[[i]]<- c(td2.1[[i]], ph2delta)
            phase_II_1ststage_ready<- TRUE
            phase_II_2ndstage_ready<- FALSE
            pop <- pop - data.sample$n2.2[i] #take out
phase #II patients used in stage 2
         }
         }
      }else {break
}
    }
  }
}
tot.ph2<-list()
tot.ph3succ<- list()
```

```
tot.ph3fail<- list()

tot.ALL<- list()

tot.ph3<-list()

tot.ph2succ<-list()

tot.ph2fail<- list()

b <-list()

r3.f<-list()

r2.f.1<-list()

r2.f.2<-list()

a<-list()

c<- list()

#collating results

for (i in 1:l){

  result[[i]]<- cbind(r2.1[[i]], r2.2[[i]],r3[[i]],
td2.1[[i]], td2.2[[i]], td3[[i]])

  result[[i]]<-data.frame(result[[i]])

  result[[i]]<-result[[i]][-1,]

  names(result[[i]])<- c("ph2out.1","ph2out.2", "ph3out",
"p2", "p2", "m2")

  tot.ph2[[i]]<-nrow(result[[i]])

  r3.f[[i]] <- factor(result[[i]][,3], levels = c(1,0,3),
labels = c("success", "fail", "not run"))

  r2.f.2[[i]] <- factor(result[[i]][,2], levels = c(1,0,3),
labels = c("success", "fail", "not run"))

  r2.f.1[[i]] <- factor(result[[i]][,1], levels =
c(1,0),labels = c("success", "fail"))


  c[[i]]<- data.frame(table(r3.f[[i]])) # table of
successful/unsuccessful phase III

  a[[i]]<- data.frame(table(r2.f.1[[i]])) # table of
successful/unsuccessful phase II st1
```

```
  b[[i]]<- data.frame(table(r2.f.2[[i]])) # table of
successful/unsuccessful phase II st2
  tot.ph3fail[[i]]<- c[[i]][2,2] # ph3 fail
  tot.ph3succ[[i]]<-c[[i]][1,2] # ph3 success
  tot.ph2fail[[i]]<- a[[i]][2,2] + b[[i]][2,2] # ph2 fail
  tot.ph2succ[[i]]<-b[[i]][1,2] # ph2 success
  tot.ph3[[i]]<- c[[i]][1,2] + c[[i]][2,2]
ph2<-unlist(tot.ph2)
ph3<-unlist(tot.ph3)
ph2fail<-unlist(tot.ph2fail)
ph3fail<-unlist(tot.ph3fail)
ph2succ<- unlist(tot.ph2succ)
ph3succ<- unlist(tot.ph3succ)
data.sample<-cbind(data.sample, ph2)
data.sample<-cbind(data.sample, ph3)
data.sample<-cbind(data.sample, ph2fail)
data.sample<-cbind(data.sample, ph3fail)
data.sample<-cbind(data.sample, ph2succ)
data.sample<-cbind(data.sample, ph3succ)
# calculating the success rates
data.sample$prob3_succ<-ph3succ/ph3
data.sample$prob2_succ<-ph2succ/ph2
########################
# understanding results
#alpha
alph3s0.01<-data.sample$ph3succ[data.sample$alpha == 0.01]
aln2s0.01<-data.sample$n2[data.sample$alpha == 0.01]
alph3s0.05<-data.sample$ph3succ[data.sample$alpha == 0.05]
aln2s0.05<-data.sample$n2[data.sample$alpha == 0.05]
alph3s0.1<-data.sample$ph3succ[data.sample$alpha == 0.1]
```

```
aln2s0.1<-data.sample$n2[data.sample$alpha == 0.1]

alph3s0.15<-data.sample$ph3succ[data.sample$alpha == 0.15]

aln2s0.15<-data.sample$n2[data.sample$alpha == 0.15]

alph3s0.2<-data.sample$ph3succ[data.sample$alpha == 0.2]

aln2s0.2<-data.sample$n2[data.sample$alpha == 0.2]

par(mfrow=c(1,1))

# plot sample size against number of phase 3 successes

plot(c(0,80), c(800,1500), type="n", xlab=expression(n[2]),
ylab="Successful Phase III trials", pch=1)

points(aln2s0.01, alph3s0.01, pch= 16)

points(aln2s0.05, alph3s0.05, pch= 16, col="red")

points(aln2s0.1, alph3s0.1, pch= 16, col="blue")

points(aln2s0.15, alph3s0.15, pch= 16, col="darkmagenta")

points(aln2s0.2, alph3s0.2, pch= 16, col="green")

lines(lowess(data.sample$n2, data.sample$ph3succ))

legend("topright",  legend = c(expression(paste(alpha, " =
", 0.01)),
expression(paste(alpha, " = ", 0.05)),
expression(paste(alpha, " = ", 0.1)),
expression(paste(alpha, " = ", 0.15)),
expression(paste(alpha, " = ", 0.2))),
col=c("black","red", "blue","darkmagenta","green"),lty =
c(1,1,1,1,1))

############################################################

# power and alpha of the phase II trials and their effect
on

#extract the number of ph3 successes with a power of 0.4

#AND ALL ALPHAS

#pow=0.4

powalphph3s0.40.01<-data.sample$ph3succ[data.sample$power
== 0.4 & data.sample$alpha == 0.01]

powalphph3s0.40.05<-data.sample$ph3succ[data.sample$power
== 0.4 & data.sample$alpha == 0.05]
```

```
powalphph3s0.40.1<-data.sample$ph3succ[data.sample$power ==
0.4 & data.sample$alpha == 0.1]

powalphph3s0.40.15<-data.sample$ph3succ[data.sample$power
== 0.4 & data.sample$alpha == 0.15]

powalphph3s0.40.2<-data.sample$ph3succ[data.sample$power ==
0.4 & data.sample$alpha == 0.2]


powalphn2s0.40.01<-data.sample$n2[data.sample$power == 0.4
& data.sample$alpha == 0.01]

powalphn2s0.40.05<-data.sample$n2[data.sample$power == 0.4
& data.sample$alpha == 0.05]

powalphn2s0.40.1<-data.sample$n2[data.sample$power == 0.4 &
data.sample$alpha == 0.1]

powalphn2s0.40.15<-data.sample$n2[data.sample$power == 0.4
& data.sample$alpha == 0.15]

powalphn2s0.40.2<-data.sample$n2[data.sample$power == 0.4 &
data.sample$alpha == 0.2]


#0.45

powalphph3s0.450.01<-data.sample$ph3succ[data.sample$beta
== 0.55 & data.sample$alpha == 0.01]

powalphph3s0.450.05<-data.sample$ph3succ[data.sample$beta
== 0.55 & data.sample$alpha == 0.05]

powalphph3s0.450.1<-data.sample$ph3succ[data.sample$beta ==
0.55 & data.sample$alpha == 0.1]

powalphph3s0.450.15<-data.sample$ph3succ[data.sample$beta
== 0.55 & data.sample$alpha == 0.15]

powalphph3s0.450.2<-data.sample$ph3succ[data.sample$beta ==
0.55 & data.sample$alpha == 0.2]


powalphn2s0.450.01<-data.sample$n2[data.sample$beta == 0.55
& data.sample$alpha == 0.01]

powalphn2s0.450.05<-data.sample$n2[data.sample$beta == 0.55
& data.sample$alpha == 0.05]
```

```
powalphn2s0.450.1<-data.sample$n2[data.sample$beta == 0.55
& data.sample$alpha == 0.1]

powalphn2s0.450.15<-data.sample$n2[data.sample$beta == 0.55
& data.sample$alpha == 0.15]

powalphn2s0.450.2<-data.sample$n2[data.sample$beta == 0.55
& data.sample$alpha == 0.2]


# 0.5

powalphph3s0.50.01<-data.sample$ph3succ[data.sample$power
== 0.5 & data.sample$alpha == 0.01]

powalphph3s0.50.05<-data.sample$ph3succ[data.sample$power
== 0.5 & data.sample$alpha == 0.05]

powalphph3s0.50.1<-data.sample$ph3succ[data.sample$power ==
0.5 & data.sample$alpha == 0.1]

powalphph3s0.50.15<-data.sample$ph3succ[data.sample$power
== 0.5 & data.sample$alpha == 0.15]

powalphph3s0.50.2<-data.sample$ph3succ[data.sample$power ==
0.5 & data.sample$alpha == 0.2]


powalphn2s0.50.01<-data.sample$n2[data.sample$power == 0.5
& data.sample$alpha == 0.01]

powalphn2s0.50.05<-data.sample$n2[data.sample$power == 0.5
& data.sample$alpha == 0.05]

powalphn2s0.50.1<-data.sample$n2[data.sample$power == 0.5 &
data.sample$alpha == 0.1]

powalphn2s0.50.15<-data.sample$n2[data.sample$power == 0.5
& data.sample$alpha == 0.15]

powalphn2s0.50.2<-data.sample$n2[data.sample$power == 0.5 &
data.sample$alpha == 0.2]


# 0.55

powalphph3s0.550.01<-data.sample$ph3succ[data.sample$power
== 0.55 & data.sample$alpha == 0.01]
```

```
powalphph3s0.550.05<-data.sample$ph3succ[data.sample$power
== 0.55 & data.sample$alpha == 0.05]

powalphph3s0.550.1<-data.sample$ph3succ[data.sample$power
== 0.55 & data.sample$alpha == 0.1]

powalphph3s0.550.15<-data.sample$ph3succ[data.sample$power
== 0.55 & data.sample$alpha == 0.15]

powalphph3s0.550.2<-data.sample$ph3succ[data.sample$power
== 0.55 & data.sample$alpha == 0.2]


powalphn2s0.550.01<-data.sample$n2[data.sample$power ==
0.55 & data.sample$alpha == 0.01]

powalphn2s0.550.05<-data.sample$n2[data.sample$power ==
0.55 & data.sample$alpha == 0.05]

powalphn2s0.550.1<-data.sample$n2[data.sample$power == 0.55
& data.sample$alpha == 0.1]

powalphn2s0.550.15<-data.sample$n2[data.sample$power ==
0.55 & data.sample$alpha == 0.15]

powalphn2s0.550.2<-data.sample$n2[data.sample$power == 0.55
& data.sample$alpha == 0.2]


# 0.6

powalphph3s0.60.01<-data.sample$ph3succ[data.sample$power
== 0.6 & data.sample$alpha == 0.01]

powalphph3s0.60.05<-data.sample$ph3succ[data.sample$power
== 0.6 & data.sample$alpha == 0.05]

powalphph3s0.60.1<-data.sample$ph3succ[data.sample$power ==
0.6 & data.sample$alpha == 0.1]

powalphph3s0.60.15<-data.sample$ph3succ[data.sample$power
== 0.6 & data.sample$alpha == 0.15]

powalphph3s0.60.2<-data.sample$ph3succ[data.sample$power ==
0.6 & data.sample$alpha == 0.2]


powalphn2s0.60.01<-data.sample$n2[data.sample$power == 0.6
& data.sample$alpha == 0.01]
```

```
powalphn2s0.60.05<-data.sample$n2[data.sample$power == 0.6
& data.sample$alpha == 0.05]

powalphn2s0.60.1<-data.sample$n2[data.sample$power == 0.6 &
data.sample$alpha == 0.1]

powalphn2s0.60.15<-data.sample$n2[data.sample$power == 0.6
& data.sample$alpha == 0.15]

powalphn2s0.60.2<-data.sample$n2[data.sample$power == 0.6 &
data.sample$alpha == 0.2]


# 0.65

powalphph3s0.650.01<-data.sample$ph3succ[data.sample$power
== 0.65 & data.sample$alpha == 0.01]

powalphph3s0.650.05<-data.sample$ph3succ[data.sample$power
== 0.65 & data.sample$alpha == 0.05]

powalphph3s0.650.1<-data.sample$ph3succ[data.sample$power
== 0.65 & data.sample$alpha == 0.1]

powalphph3s0.650.15<-data.sample$ph3succ[data.sample$power
== 0.65 & data.sample$alpha == 0.15]

powalphph3s0.650.2<-data.sample$ph3succ[data.sample$power
== 0.65 & data.sample$alpha == 0.2]


powalphn2s0.650.01<-data.sample$n2[data.sample$power ==
0.65 & data.sample$alpha == 0.01]

powalphn2s0.650.05<-data.sample$n2[data.sample$power ==
0.65 & data.sample$alpha == 0.05]

powalphn2s0.650.1<-data.sample$n2[data.sample$power == 0.65
& data.sample$alpha == 0.1]

powalphn2s0.650.15<-data.sample$n2[data.sample$power ==
0.65 & data.sample$alpha == 0.15]

powalphn2s0.650.2<-data.sample$n2[data.sample$power == 0.65
& data.sample$alpha == 0.2]


# 0.7
```

```
powalphph3s0.70.01<-data.sample$ph3succ[data.sample$power
== 0.7 & data.sample$alpha == 0.01]

powalphph3s0.70.05<-data.sample$ph3succ[data.sample$power
== 0.7 & data.sample$alpha == 0.05]

powalphph3s0.70.1<-data.sample$ph3succ[data.sample$power ==
0.7 & data.sample$alpha == 0.1]

powalphph3s0.70.15<-data.sample$ph3succ[data.sample$power
== 0.7 & data.sample$alpha == 0.15]

powalphph3s0.70.2<-data.sample$ph3succ[data.sample$power ==
0.7 & data.sample$alpha == 0.2]


powalphn2s0.70.01<-data.sample$n2[data.sample$power == 0.7
& data.sample$alpha == 0.01]

powalphn2s0.70.05<-data.sample$n2[data.sample$power == 0.7
& data.sample$alpha == 0.05]

powalphn2s0.70.1<-data.sample$n2[data.sample$power == 0.7 &
data.sample$alpha == 0.1]

powalphn2s0.70.15<-data.sample$n2[data.sample$power == 0.7
& data.sample$alpha == 0.15]

powalphn2s0.70.2<-data.sample$n2[data.sample$power == 0.7 &
data.sample$alpha == 0.2]


# 0.75

powalphph3s0.750.01<-data.sample$ph3succ[data.sample$power
== 0.75 & data.sample$alpha == 0.01]

powalphph3s0.750.05<-data.sample$ph3succ[data.sample$power
== 0.75 & data.sample$alpha == 0.05]

powalphph3s0.750.1<-data.sample$ph3succ[data.sample$power
== 0.75 & data.sample$alpha == 0.1]

powalphph3s0.750.15<-data.sample$ph3succ[data.sample$power
== 0.75 & data.sample$alpha == 0.15]

powalphph3s0.750.2<-data.sample$ph3succ[data.sample$power
== 0.75 & data.sample$alpha == 0.2]
```

```
powalphn2s0.750.01<-data.sample$n2[data.sample$power ==
0.75 & data.sample$alpha == 0.01]

powalphn2s0.750.05<-data.sample$n2[data.sample$power ==
0.75 & data.sample$alpha == 0.05]

powalphn2s0.750.1<-data.sample$n2[data.sample$power == 0.75
& data.sample$alpha == 0.1]

powalphn2s0.750.15<-data.sample$n2[data.sample$power ==
0.75 & data.sample$alpha == 0.15]

powalphn2s0.750.2<-data.sample$n2[data.sample$power == 0.75
& data.sample$alpha == 0.2]


# 0.8

powalphph3s0.80.01<-data.sample$ph3succ[data.sample$power
== 0.8 & data.sample$alpha == 0.01]

powalphph3s0.80.05<-data.sample$ph3succ[data.sample$power
== 0.8 & data.sample$alpha == 0.05]

powalphph3s0.80.1<-data.sample$ph3succ[data.sample$power ==
0.8 & data.sample$alpha == 0.1]

powalphph3s0.80.15<-data.sample$ph3succ[data.sample$power
== 0.8 & data.sample$alpha == 0.15]

powalphph3s0.80.2<-data.sample$ph3succ[data.sample$power ==
0.8 & data.sample$alpha == 0.2]


powalphn2s0.80.01<-data.sample$n2[data.sample$power == 0.8
& data.sample$alpha == 0.01]

powalphn2s0.80.05<-data.sample$n2[data.sample$power == 0.8
& data.sample$alpha == 0.05]

powalphn2s0.80.1<-data.sample$n2[data.sample$power == 0.8 &
data.sample$alpha == 0.1]

powalphn2s0.80.15<-data.sample$n2[data.sample$power == 0.8
& data.sample$alpha == 0.15]

powalphn2s0.80.2<-data.sample$n2[data.sample$power == 0.8 &
data.sample$alpha == 0.2]
```

```
# 0.85

powalphph3s0.850.01<-data.sample$ph3succ[data.sample$power
== 0.85 & data.sample$alpha == 0.01]

powalphph3s0.850.05<-data.sample$ph3succ[data.sample$power
== 0.85 & data.sample$alpha == 0.05]

powalphph3s0.850.1<-data.sample$ph3succ[data.sample$power
== 0.85 & data.sample$alpha == 0.1]

powalphph3s0.850.15<-data.sample$ph3succ[data.sample$power
== 0.85 & data.sample$alpha == 0.15]

powalphph3s0.850.2<-data.sample$ph3succ[data.sample$power
== 0.85 & data.sample$alpha == 0.2]


powalphn2s0.850.01<-data.sample$n2[data.sample$power ==
0.85 & data.sample$alpha == 0.01]

powalphn2s0.850.05<-data.sample$n2[data.sample$power ==
0.85 & data.sample$alpha == 0.05]

powalphn2s0.850.1<-data.sample$n2[data.sample$power == 0.85
& data.sample$alpha == 0.1]

powalphn2s0.850.15<-data.sample$n2[data.sample$power ==
0.85 & data.sample$alpha == 0.15]

powalphn2s0.850.2<-data.sample$n2[data.sample$power == 0.85
& data.sample$alpha == 0.2]


# 0.9

powalphph3s0.90.01<-data.sample$ph3succ[data.sample$power
== 0.9 & data.sample$alpha == 0.01]

powalphph3s0.90.05<-data.sample$ph3succ[data.sample$power
== 0.9 & data.sample$alpha == 0.05]

powalphph3s0.90.1<-data.sample$ph3succ[data.sample$power ==
0.9 & data.sample$alpha == 0.1]

powalphph3s0.90.15<-data.sample$ph3succ[data.sample$power
== 0.9 & data.sample$alpha == 0.15]

powalphph3s0.90.2<-data.sample$ph3succ[data.sample$power ==
0.9 & data.sample$alpha == 0.2]
```

```
powalphn2s0.90.01<-data.sample$n2[data.sample$power == 0.9
& data.sample$alpha == 0.01]

powalphn2s0.90.05<-data.sample$n2[data.sample$power == 0.9
& data.sample$alpha == 0.05]

powalphn2s0.90.1<-data.sample$n2[data.sample$power == 0.9 &
data.sample$alpha == 0.1]

powalphn2s0.90.15<-data.sample$n2[data.sample$power == 0.9
& data.sample$alpha == 0.15]

powalphn2s0.90.2<-data.sample$n2[data.sample$power == 0.9 &
data.sample$alpha == 0.2]

# panel plot of the effect of power (0.4-0.9 increments of
#0.5) and alpha (0.01,0.05,0.1,0.15,0.2)

par(mar=c(5.1,1,1,1))

par(oma=c(1,1,1,1))

par(mfrow=c(3,4))

# 0.4 plot all alpha

plot(c(0,80), c(800,1400), type="n", xlab=expression(n[2]),
ylab = "Successful Phase III", main = bquote( ~ 1-beta ==
0.4), pch = 1)

points(powalphn2s0.40.01, powalphph3s0.40.01,pch= 16,
col="black")

points(powalphn2s0.40.05, powalphph3s0.40.05,pch= 16,
col="red")

points(powalphn2s0.40.1, powalphph3s0.40.1,pch= 16,
col="blue")

points(powalphn2s0.40.15, powalphph3s0.40.15,pch= 16,
col="darkmagenta")

points(powalphn2s0.40.2, powalphph3s0.40.2,pch= 16,
col="green")

# 0.45 plot all alpha

plot(c(0,80), c(800,1400), type="n", xlab=expression(n[2]),
ylab = "Successful Phase III", main = bquote( ~ 1-beta ==
0.45), pch = 1)
```

```
points(powalphn2s0.450.01, powalphph3s0.450.01,pch= 16,
col="black")

points(powalphn2s0.450.05, powalphph3s0.450.05,pch= 16,
col="red")

points(powalphn2s0.450.1, powalphph3s0.450.1,pch= 16,
col="blue")

points(powalphn2s0.450.15, powalphph3s0.450.15,pch= 16,
col="darkmagenta")

points(powalphn2s0.450.2, powalphph3s0.450.2,pch= 16,
col="green")

# 0.5 plot all alpha

plot(c(0,80), c(800,1400), type="n", xlab=expression(n[2]),
ylab = "Successful Phase III", main = bquote( ~ 1-beta ==
0.5), pch = 1)

points(powalphn2s0.50.01, powalphph3s0.50.01,pch= 16,
col="black")

points(powalphn2s0.50.05, powalphph3s0.50.05,pch= 16,
col="red")

points(powalphn2s0.50.1, powalphph3s0.50.1,pch= 16,
col="blue")

points(powalphn2s0.50.15, powalphph3s0.50.15,pch= 16,
col="darkmagenta")

points(powalphn2s0.50.2, powalphph3s0.50.2,pch= 16,
col="green")

# 0.55 plot all alpha

plot(c(0,80), c(800,1400), type="n", xlab=expression(n[2]),
ylab = "Successful Phase III", main = bquote( ~ 1-beta ==
0.55), pch = 1)

points(powalphn2s0.550.01, powalphph3s0.550.01,pch= 16,
col="black")

points(powalphn2s0.550.05, powalphph3s0.550.05,pch= 16,
col="red")

points(powalphn2s0.550.1, powalphph3s0.550.1,pch= 16,
col="blue")
```

```
points(powalphn2s0.550.15, powalphph3s0.550.15,pch= 16,
col="darkmagenta")

points(powalphn2s0.550.2, powalphph3s0.550.2,pch= 16,
col="green")

# 0.6 plot all alpha

plot(c(0,80), c(800,1400), type="n", xlab=expression(n[2]),
ylab = "Successful Phase III", main = bquote( ~ 1-beta ==
0.6), pch = 1)

points(powalphn2s0.60.01, powalphph3s0.60.01,pch= 16,
col="black")

points(powalphn2s0.60.05, powalphph3s0.60.05,pch= 16,
col="red")

points(powalphn2s0.60.1, powalphph3s0.60.1,pch= 16,
col="blue")

points(powalphn2s0.60.15, powalphph3s0.60.15,pch= 16,
col="darkmagenta")

points(powalphn2s0.60.2, powalphph3s0.60.2,pch= 16,
col="green")

# 0.65 plot all alpha

plot(c(0,80), c(800,1400), type="n", xlab=expression(n[2]),
ylab = "Successful Phase III", main = bquote( ~ 1-beta ==
0.65), pch = 1)

points(powalphn2s0.650.01, powalphph3s0.650.01,pch= 16,
col="black")

points(powalphn2s0.650.05, powalphph3s0.650.05,pch= 16,
col="red")

points(powalphn2s0.650.1, powalphph3s0.650.1,pch= 16,
col="blue")

points(powalphn2s0.650.15, powalphph3s0.650.15,pch= 16,
col="darkmagenta")

points(powalphn2s0.650.2, powalphph3s0.650.2,pch= 16,
col="green")

# 0.7 plot all alpha
```

```
plot(c(0,80), c(800,1400), type="n", xlab=expression(n[2]),
ylab = "Successful Phase III", main = bquote( ~ 1-beta ==
0.7), pch = 1)

points(powalphn2s0.70.01, powalphph3s0.70.01,pch= 16,
col="black")

points(powalphn2s0.70.05, powalphph3s0.70.05,pch= 16,
col="red")

points(powalphn2s0.70.1, powalphph3s0.70.1,pch= 16,
col="blue")

points(powalphn2s0.70.15, powalphph3s0.70.15,pch= 16,
col="darkmagenta")

points(powalphn2s0.70.2, powalphph3s0.70.2,pch= 16,
col="green")

# 0.75 plot all alpha

plot(c(0,80), c(800,1400), type="n", xlab=expression(n[2]),
ylab = "Successful Phase III", main = bquote( ~ 1-beta ==
0.75), pch = 1)

points(powalphn2s0.750.01, powalphph3s0.750.01,pch= 16,
col="black")

points(powalphn2s0.750.05, powalphph3s0.750.05,pch= 16,
col="red")

points(powalphn2s0.750.1, powalphph3s0.750.1,pch= 16,
col="blue")

points(powalphn2s0.750.15, powalphph3s0.750.15,pch= 16,
col="darkmagenta")

points(powalphn2s0.750.2, powalphph3s0.750.2,pch= 16,
col="green")

# 0.8 plot all alpha

plot(c(0,80), c(800,1400), type="n", xlab=expression(n[2]),
ylab = "Successful Phase III", main = bquote( ~ 1-beta ==
0.8), pch = 1)

points(powalphn2s0.80.01, powalphph3s0.80.01,pch= 16,
col="black")
```

```
points(powalphn2s0.80.05, powalphph3s0.80.05,pch= 16,
col="red")

points(powalphn2s0.80.1, powalphph3s0.80.1,pch= 16,
col="blue")

points(powalphn2s0.80.15, powalphph3s0.80.15,pch= 16,
col="darkmagenta")

points(powalphn2s0.80.2, powalphph3s0.80.2,pch= 16,
col="green")

# 0.85 plot all alpha

plot(c(0,80), c(800,1400), type="n", xlab=expression(n[2]),
ylab = "Successful Phase III", main = bquote( ~ 1-beta ==
0.85), pch = 1)

points(powalphn2s0.850.01, powalphph3s0.850.01,pch= 16,
col="black")

points(powalphn2s0.850.05, powalphph3s0.850.05,pch= 16,
col="red")

points(powalphn2s0.850.1, powalphph3s0.850.1,pch= 16,
col="blue")

points(powalphn2s0.850.15, powalphph3s0.850.15,pch= 16,
col="darkmagenta")

points(powalphn2s0.850.2, powalphph3s0.850.2,pch= 16,
col="green")

# 0.9 plot all alpha

plot(c(0,80), c(800,1400), type="n", xlab=expression(n[2]),
ylab = "Successful Phase III", main = bquote( ~ 1-beta ==
0.9), pch = 1)

points(powalphn2s0.90.01, powalphph3s0.90.01,pch= 16,
col="black")

points(powalphn2s0.90.05, powalphph3s0.90.05,pch= 16,
col="red")

points(powalphn2s0.90.1, powalphph3s0.90.1,pch= 16,
col="blue")

points(powalphn2s0.90.15, powalphph3s0.90.15,pch= 16,
col="darkmagenta")
```

```
points(powalphn2s0.90.2, powalphph3s0.90.2,pch= 16,
col="green")


############################################################
#success rate of phase III
par(mfrow=c(1,1))
rph3alph3s0.01<-data.sample$prob3_succ[data.sample$alpha ==
0.01]
rph3aln2s0.01<-data.sample$n2[data.sample$alpha == 0.01]


rph3alph3s0.05<-data.sample$prob3_succ[data.sample$alpha ==
0.05]
rph3aln2s0.05<-data.sample$n2[data.sample$alpha == 0.05]


rph3alph3s0.1<-data.sample$prob3_succ[data.sample$alpha ==
0.1]
rph3aln2s0.1<-data.sample$n2[data.sample$alpha == 0.1]


rph3alph3s0.15<-data.sample$prob3_succ[data.sample$alpha ==
0.15]
rph3aln2s0.15<-data.sample$n2[data.sample$alpha == 0.15]


rph3alph3s0.2<-data.sample$prob3_succ[data.sample$alpha ==
0.2]
rph3aln2s0.2<-data.sample$n2[data.sample$alpha == 0.2]
#phase III success rate
plot(c(0,35), c(0.92,1), type="n", xlab=expression(n[2]),
ylab="Rate of Successful Phase III trials", pch=1)
points(rph3aln2s0.01, rph3alph3s0.01, pch= 16)
points(rph3aln2s0.05, rph3alph3s0.05, pch= 16, col="red")
points(rph3aln2s0.1, rph3alph3s0.1, pch= 16, col="blue")
```

```
points(rph3aln2s0.15, rph3alph3s0.15, pch= 16,
col="darkmagenta")

points(rph3aln2s0.2, rph3alph3s0.2, pch= 16, col="green")

lines(lowess(data.sample$n2, data.sample$prob3_succ))

legend("bottomright", legend = c(expression(paste(alpha, "
= ", 0.01)),
expression(paste(alpha, " = ", 0.05)),
expression(paste(alpha, " = ", 0.1)),
expression(paste(alpha, " = ", 0.15)),
expression(paste(alpha, " = ", 0.2))),
col=c("black","red", "blue","darkmagenta","green"),lty =
c(1,1,1,1,1))

############################################################

# success rate of phase II

par(mfrow=c(1,1))

rph2alph3s0.01<-data.sample$prob2_succ[data.sample$alpha ==
0.01]

rph2aln2s0.01<-data.sample$n2[data.sample$alpha == 0.01]


rph2alph3s0.05<-data.sample$prob2_succ[data.sample$alpha ==
0.05]

rph2aln2s0.05<-data.sample$n2[data.sample$alpha == 0.05]


rph2alph3s0.1<-data.sample$prob2_succ[data.sample$alpha ==
0.1]

rph2aln2s0.1<-data.sample$n2[data.sample$alpha == 0.1]


rph2alph3s0.15<-data.sample$prob2_succ[data.sample$alpha ==
0.15]

rph2aln2s0.15<-data.sample$n2[data.sample$alpha == 0.15]


rph2alph3s0.2<-data.sample$prob2_succ[data.sample$alpha ==
0.2]
```

```
rph2aln2s0.2<-data.sample$n2[data.sample$alpha == 0.2]

#phase III success rate

plot(c(0,35), c(0,.4), type="n", xlab=expression(n[2]),
ylab="Rate of Successful Phase II trials", pch=1)

points(rph2aln2s0.01, rph2alph3s0.01, pch= 16)

points(rph2aln2s0.05, rph2alph3s0.05, pch= 16, col="red")

points(rph2aln2s0.1, rph2alph3s0.1, pch= 16, col="blue")

points(rph2aln2s0.15, rph2alph3s0.15, pch= 16,
col="darkmagenta")

points(rph2aln2s0.2, rph2alph3s0.2, pch= 16, col="green")

lines(lowess(data.sample$n2, data.sample$prob2_succ))

legend("topleft", legend = c(expression(paste(alpha, " = ",
0.01)),
expression(paste(alpha, " = ", 0.05)),
expression(paste(alpha, " = ", 0.1)),
expression(paste(alpha, " = ", 0.15)),
expression(paste(alpha, " = ", 0.2))),
col=c("black","red", "blue","darkmagenta","green"),lty =
c(1,1,1,1,1))
```

F.2 Randomised single-stage

```
# loading needed packages

library("plyr")

# fixing a seed to obtain reproducible results

set.seed(2212022)

# randomised phase II trials; sample size effect

# two-sided phase III

#true treatment effects

truedelta<- rnorm(10000, 0, 1)

# histogram of the distribution treatment effect available

hist(truedelta,prob=TRUE, breaks=20, main = "True Treatment
Effect", xlab = expression(Delta))

curve(dnorm(x, mean(truedelta), sd(truedelta)), add=TRUE,
col="darkblue", lwd=2)

p.1<- 0.25 # control arm's probability of success

# corresponding true phase II treatment effects

p.2 <- ((exp(truedelta*pi/sqrt(3))*p.1
))/((p.1*(exp(truedelta*pi/sqrt(3))  - 1)  + 1))  # the
#inverse log of the normally distributed treatment effects


# how the true delta is correlated to p.2

hist(truedelta,prob=TRUE, breaks=20, main = "True Treatment
Effect", xlab = expression(mu[2]))

plot(truedelta, p.2, xlim = c(-4,4), ylim = c(0,1),xlab =
expression(mu[2]), ylab =expression(p[2]) )

Deltatrue<- data.frame(p.2,truedelta)

alpha=0.05 # type I error for phase II and III trials

power=0.8 # 1-type II error for phase II and III trials

delta1=0.3 # delta1 is the clinically significant
#difference we wish to detect in phase III

delta2=0.2# delta2 is the csd we wish to detect in phase II
```

```
sigma<-1 # sd for underlying treatment effect


# designs used in evaluation of randomised designs' sample
#size
data.sample <-
read.table("C:/Users/enada/OneDrive/Desktop/Nada Elbeltagi-
PhD 17.01.22/Write up/samplesize.csv", header=TRUE,
sep=",")
data.sample$power<-1-data.sample$beta
data.sample$n2<- (round((p.1*(1 - p.1) + (p.1+delta2)*(1 -
(p.1+delta2)))*((qnorm(1-
data.sample$alpha)+qnorm(data.sample$power))^2)/
(delta2^2)/2)*2)*2 # formula to calculate sample sizes in
#randomised phase II


data.sample<-data.sample[!(data.sample$n2>300),] #getting
rid of any combinations where n2 is larger than 300
data.sample<-data.sample[!(data.sample$n2<=0),] #getting
rid of any combinations where n2 is larger than 300
data.sample$halfn2<-data.sample$n2/2
l<- length(data.sample$alpha)


get.sample.sizeph3 <- function(sigma,alpha,power,delta1)
  # function to do standard sample size calculation
assuming sigma = 1
  # N.B. alpha is one-sided error rate and sample size is
total for two arms
{
 round((2 * 2 *sigma*sigma*(qnorm(1-
(alpha/2))+qnorm(power))^2/(delta1^2))/2)*2
}
n3<-get.sample.sizeph3(sigma, alpha, power, delta1)
```

```
# sample size of the phase III trial depends on alpha,
#power, sigma
# and delta1 which is the treatment effect we wish to
#detect
mu1<-0 # control arm mean
sigma1<- sigma2 <-1
control3<- rnorm(10000,mu1,sigma1) # patients available for
#control arm in phase III
ph3delta<-c()
mu2<-list()
experimental3<-list()
ph3samplecontrol <- list()
ph3sampleexp <- list()
ph3test<-list()
pv3<-list()
control2<-rbinom(10000,1,p.1) # patients available for
control arm in phase II
ph2delta<-c()
experimental2<-list()
ph2samplecontrol<-list()
ph2sampleexp<-list()
ph2vecs<-vector()
ph2grp<-vector()
total<- list()
ph2test<-list()
pv2<-c()
ph2tabexp<-list()
ph2tabcont<-list()
x<-c()
n<-c()
phase_III_ready <- FALSE
```

```
phase_II_ready<- TRUE

r2<- lapply(1:l, matrix, data=NA,nrow=1, ncol=1)
r3<- lapply(1:l, matrix, data=NA,nrow=1, ncol=1)
result<-lapply(1:l, matrix, data=NA,nrow=1, ncol=1)
td2<- lapply(1:l, matrix, data=NA,nrow=1, ncol=1)
td3<- lapply(1:l, matrix, data=NA,nrow=1, ncol=1)

for (i in 1:l){
  pop <- 500000
# population of patients
  while( pop > (n3)){
    # discontinue loop if pop less than phase 3 sample size
    if(phase_III_ready & phase_II_ready){
      #run phase III trial
      mu2<- (sqrt(3)*log(((p.1*ph2delta)-
ph2delta)/(p.1*(ph2delta-1))))/pi
      # mu2 is the treatment effect tested in phase 3 and
#corresponds to p.2
      experimental3<-rnorm(10000,mu2,sigma2)
      # patients available to be entered in the
#experimental arm
      ph3samplecontrol<- sample(control3, n3/2,
replace=FALSE) # sample n2/2 patients for the phase III
      #half patients sampled in control arm
      ph3sampleexp<- sample(experimental3, n3/2,
replace=FALSE) # sample the remaining patients for the
#phase III
      #half patients sampled in experimental arm
```

```
      ph3test<-t.test(ph3samplecontrol,ph3sampleexp,
var.equal= TRUE, mu= 0, alternative = "t") #running the
two-sample t-test:
      pv3 <- ph3test$p.value # pvalue extracted
      if (pv3 <= alpha & mu2>0){
        # ph3 successful if pvalue<=alpha and the treatment
#effect>0
        td3[[i]]<- c(td3[[i]], mu2)
        r3[[i]]<- c(r3[[i]],1)
        pop <- pop - n3# take out used patients in trial
      }
      else {
        # ph3 unsuccessful
        td3[[i]]<- c(td3[[i]], mu2)
        r3[[i]]<-c(r3[[i]],0)
        pop<- pop - n3# take out used patients in trial
}
     phase_III_ready <- FALSE
    } else {
# continue loop if pop > phase II and 3 sample size
      if(pop> n3 + data.sample$n2[i]) {
        phase_II_ready<- TRUE
        #phase II simulations
        ph2delta<- sample(p.2,1,replace = FALSE)
        experimental2<-rbinom(10000,1,ph2delta)
        ph2samplecontrol <- sample(control2,
data.sample$halfn2[i], replace=FALSE) # sampled 1/2 n in
control
        ph2sampleexp <-
sample(experimental2,data.sample$halfn2[i], replace=FALSE)
# sampled n/2 in experimental
```

```
        ph2samplecontrol<- factor(ph2samplecontrol,c(0,1),
labels = c('fail','success'))
        ph2sampleexp<- factor(ph2sampleexp,c(0,1), labels =
c('fail','success'))
        ph2tabexp<-table(ph2sampleexp)
        ph2tabcont<-table(ph2samplecontrol)
        x<-c(ph2tabcont[2],ph2tabexp[2])
        n<-c(ph2tabcont[2]+ph2tabcont[1],
ph2tabexp[2]+ph2tabexp[1])
        ph2test<-prop.test(x, n, alternative = c("l"),
conf.level = (1- alpha), correct = FALSE) # run ph2 trial
        pv2 <- ph2test$p.value # extract the p-value
        if (pv2 <= data.sample$alpha[i] & !is.na(pv2)){
          #ph2 success
          td2[[i]]<- c(td2[[i]], ph2delta)
          r2[[i]] <- c(r2[[i]],1)
          phase_III_ready <- TRUE
          pop <- pop - data.sample$n2[i] # take out used
#patients in trial
        }
        else {
          # phase II failure
          td2[[i]]<- c(td2[[i]], ph2delta)
          td3[[i]]<- c(td3[[i]], NA)
          r2[[i]] <- c(r2[[i]],0)
          phase_III_ready <- FALSE
          r3[[i]] <- c(r3[[i]], 3)
          pop <- pop - data.sample$n2[i] # take out used
#patients in trial
        }
      } else {break
```

```
}
    }
  }
}
tot.ph2<-list()
tot.ph3succ<- list()
tot.ph3fail<- list()
tot.ALL<- list()
tot.ph3<-list()
tot.ph2succ<-list()
tot.ph2fail<- list()
b <-list()
r3.f<-list()
r2.f<-list()
a<-list()
#collate results
for (i in 1:l){
  result[[i]]<- cbind(r2[[i]],r3[[i]], td2[[i]], td3[[i]])
  result[[i]]<-data.frame(result[[i]])
  result[[i]]<-result[[i]][-1,]
  names(result[[i]])<- c("ph2out", "ph3out", "p2", "m2")
  tot.ph2[[i]]<-nrow(result[[i]])
  r3.f[[i]] <- factor(result[[i]][,2], levels = c(1,0,3),
labels = c("success", "fail", "not run"))
  r2.f[[i]] <- factor(result[[i]][,1], levels =
c(1,0),labels = c("success", "fail"))

  b[[i]]<- data.frame(table(r3.f[[i]]))
  a[[i]]<- data.frame(table(r2.f[[i]]))
```

```
  tot.ph3fail[[i]]<- b[[i]][2,2]
  tot.ph3succ[[i]]<-b[[i]][1,2]

  tot.ph2fail[[i]]<- a[[i]][2,2]
  tot.ph2succ[[i]]<-a[[i]][1,2]

  tot.ph3[[i]]<- b[[i]][1,2] + b[[i]][2,2]

  tot.ALL[[i]]<- (data.sample$n2[i]*tot.ph2[[i]])+
(n3*tot.ph3[[i]])
  assign(paste0("res",i),as.data.frame(result[[i]]))
}
ph2<-unlist(tot.ph2)
ph3<-unlist(tot.ph3)
ph2fail<-unlist(tot.ph2fail)
ph3fail<-unlist(tot.ph3fail)
ph2succ<- unlist(tot.ph2succ)
ph3succ<- unlist(tot.ph3succ)

data.sample<-cbind(data.sample, ph2)
data.sample<-cbind(data.sample, ph3)
data.sample<-cbind(data.sample, ph2fail)
data.sample<-cbind(data.sample, ph3fail)
data.sample<-cbind(data.sample, ph2succ)
data.sample<-cbind(data.sample, ph3succ)
# calculate success rates
data.sample$prob3_succ<-ph3succ/ph3
data.sample$prob2_succ<-ph2succ/ph2
```

```
########################
# understanding results#
par(mar=c(5.1, 4.1, 4.1, 2.1))
#alpha
alph3s0.01<-data.sample$ph3succ[data.sample$alpha == 0.01]
aln2s0.01<-data.sample$n2[data.sample$alpha == 0.01]

alph3s0.05<-data.sample$ph3succ[data.sample$alpha == 0.05]
aln2s0.05<-data.sample$n2[data.sample$alpha == 0.05]

alph3s0.1<-data.sample$ph3succ[data.sample$alpha == 0.1]
aln2s0.1<-data.sample$n2[data.sample$alpha == 0.1]

alph3s0.15<-data.sample$ph3succ[data.sample$alpha == 0.15]
aln2s0.15<-data.sample$n2[data.sample$alpha == 0.15]

alph3s0.2<-data.sample$ph3succ[data.sample$alpha == 0.2]
aln2s0.2<-data.sample$n2[data.sample$alpha == 0.2]
# plot: sample size against randomised design sample sizes
plot(c(0,300), c(350,1500), type="n",
xlab=expression(n[2]), ylab="Successful Phase III trials",
pch=1)
points(aln2s0.01, alph3s0.01, pch= 16)
points(aln2s0.05, alph3s0.05, pch= 16, col="red")
points(aln2s0.1, alph3s0.1, pch= 16, col="blue")
points(aln2s0.15, alph3s0.15, pch= 16, col="darkmagenta")
points(aln2s0.2, alph3s0.2, pch= 16, col="green")
lines(lowess(data.sample$n2, data.sample$ph3succ))
legend("topright",  legend = c(expression(paste(alpha, " =
", 0.01)),
```

```
expression(paste(alpha, " = ", 0.05)),
expression(paste(alpha, " = ", 0.1)),
expression(paste(alpha, " = ", 0.15)),
expression(paste(alpha, " = ", 0.2))),
col=c("black","red", "blue","darkmagenta","green"),lty =
c(1,1,1,1,1), cex=.5)

############################################################

# power and alpha of the phase II trials and their effect

#extract the number of ph3 successes with a power of 0.4

#AND ALL ALPHAS

#pow=0.4

powalphph3s0.40.01<-data.sample$ph3succ[data.sample$power
== 0.4 & data.sample$alpha == 0.01]

powalphph3s0.40.05<-data.sample$ph3succ[data.sample$power
== 0.4 & data.sample$alpha == 0.05]

powalphph3s0.40.1<-data.sample$ph3succ[data.sample$power ==
0.4 & data.sample$alpha == 0.1]

powalphph3s0.40.15<-data.sample$ph3succ[data.sample$power
== 0.4 & data.sample$alpha == 0.15]

powalphph3s0.40.2<-data.sample$ph3succ[data.sample$power ==
0.4 & data.sample$alpha == 0.2]


powalphn2s0.40.01<-data.sample$n2[data.sample$power == 0.4
& data.sample$alpha == 0.01]

powalphn2s0.40.05<-data.sample$n2[data.sample$power == 0.4
& data.sample$alpha == 0.05]

powalphn2s0.40.1<-data.sample$n2[data.sample$power == 0.4 &
data.sample$alpha == 0.1]

powalphn2s0.40.15<-data.sample$n2[data.sample$power == 0.4
& data.sample$alpha == 0.15]

powalphn2s0.40.2<-data.sample$n2[data.sample$power == 0.4 &
data.sample$alpha == 0.2]


#0.45
```

```
powalphph3s0.450.01<-data.sample$ph3succ[data.sample$beta
== 0.55 & data.sample$alpha == 0.01]

powalphph3s0.450.05<-data.sample$ph3succ[data.sample$beta
== 0.55 & data.sample$alpha == 0.05]

powalphph3s0.450.1<-data.sample$ph3succ[data.sample$beta ==
0.55 & data.sample$alpha == 0.1]

powalphph3s0.450.15<-data.sample$ph3succ[data.sample$beta
== 0.55 & data.sample$alpha == 0.15]

powalphph3s0.450.2<-data.sample$ph3succ[data.sample$beta ==
0.55 & data.sample$alpha == 0.2]


powalphn2s0.450.01<-data.sample$n2[data.sample$beta == 0.55
& data.sample$alpha == 0.01]

powalphn2s0.450.05<-data.sample$n2[data.sample$beta == 0.55
& data.sample$alpha == 0.05]

powalphn2s0.450.1<-data.sample$n2[data.sample$beta == 0.55
& data.sample$alpha == 0.1]

powalphn2s0.450.15<-data.sample$n2[data.sample$beta == 0.55
& data.sample$alpha == 0.15]

powalphn2s0.450.2<-data.sample$n2[data.sample$beta == 0.55
& data.sample$alpha == 0.2]


# 0.5

powalphph3s0.50.01<-data.sample$ph3succ[data.sample$power
== 0.5 & data.sample$alpha == 0.01]

powalphph3s0.50.05<-data.sample$ph3succ[data.sample$power
== 0.5 & data.sample$alpha == 0.05]

powalphph3s0.50.1<-data.sample$ph3succ[data.sample$power ==
0.5 & data.sample$alpha == 0.1]

powalphph3s0.50.15<-data.sample$ph3succ[data.sample$power
== 0.5 & data.sample$alpha == 0.15]

powalphph3s0.50.2<-data.sample$ph3succ[data.sample$power ==
0.5 & data.sample$alpha == 0.2]
```

```
powalphn2s0.50.01<-data.sample$n2[data.sample$power == 0.5
& data.sample$alpha == 0.01]

powalphn2s0.50.05<-data.sample$n2[data.sample$power == 0.5
& data.sample$alpha == 0.05]

powalphn2s0.50.1<-data.sample$n2[data.sample$power == 0.5 &
data.sample$alpha == 0.1]

powalphn2s0.50.15<-data.sample$n2[data.sample$power == 0.5
& data.sample$alpha == 0.15]

powalphn2s0.50.2<-data.sample$n2[data.sample$power == 0.5 &
data.sample$alpha == 0.2]


# 0.55

powalphph3s0.550.01<-data.sample$ph3succ[data.sample$power
== 0.55 & data.sample$alpha == 0.01]

powalphph3s0.550.05<-data.sample$ph3succ[data.sample$power
== 0.55 & data.sample$alpha == 0.05]

powalphph3s0.550.1<-data.sample$ph3succ[data.sample$power
== 0.55 & data.sample$alpha == 0.1]

powalphph3s0.550.15<-data.sample$ph3succ[data.sample$power
== 0.55 & data.sample$alpha == 0.15]

powalphph3s0.550.2<-data.sample$ph3succ[data.sample$power
== 0.55 & data.sample$alpha == 0.2]


powalphn2s0.550.01<-data.sample$n2[data.sample$power ==
0.55 & data.sample$alpha == 0.01]

powalphn2s0.550.05<-data.sample$n2[data.sample$power ==
0.55 & data.sample$alpha == 0.05]

powalphn2s0.550.1<-data.sample$n2[data.sample$power == 0.55
& data.sample$alpha == 0.1]

powalphn2s0.550.15<-data.sample$n2[data.sample$power ==
0.55 & data.sample$alpha == 0.15]

powalphn2s0.550.2<-data.sample$n2[data.sample$power == 0.55
& data.sample$alpha == 0.2]
```

```
# 0.6

powalphph3s0.60.01<-data.sample$ph3succ[data.sample$power
== 0.6 & data.sample$alpha == 0.01]

powalphph3s0.60.05<-data.sample$ph3succ[data.sample$power
== 0.6 & data.sample$alpha == 0.05]

powalphph3s0.60.1<-data.sample$ph3succ[data.sample$power ==
0.6 & data.sample$alpha == 0.1]

powalphph3s0.60.15<-data.sample$ph3succ[data.sample$power
== 0.6 & data.sample$alpha == 0.15]

powalphph3s0.60.2<-data.sample$ph3succ[data.sample$power ==
0.6 & data.sample$alpha == 0.2]


powalphn2s0.60.01<-data.sample$n2[data.sample$power == 0.6
& data.sample$alpha == 0.01]

powalphn2s0.60.05<-data.sample$n2[data.sample$power == 0.6
& data.sample$alpha == 0.05]

powalphn2s0.60.1<-data.sample$n2[data.sample$power == 0.6 &
data.sample$alpha == 0.1]

powalphn2s0.60.15<-data.sample$n2[data.sample$power == 0.6
& data.sample$alpha == 0.15]

powalphn2s0.60.2<-data.sample$n2[data.sample$power == 0.6 &
data.sample$alpha == 0.2]


# 0.65

powalphph3s0.650.01<-data.sample$ph3succ[data.sample$power
== 0.65 & data.sample$alpha == 0.01]

powalphph3s0.650.05<-data.sample$ph3succ[data.sample$power
== 0.65 & data.sample$alpha == 0.05]

powalphph3s0.650.1<-data.sample$ph3succ[data.sample$power
== 0.65 & data.sample$alpha == 0.1]

powalphph3s0.650.15<-data.sample$ph3succ[data.sample$power
== 0.65 & data.sample$alpha == 0.15]

powalphph3s0.650.2<-data.sample$ph3succ[data.sample$power
== 0.65 & data.sample$alpha == 0.2]
```

```
powalphn2s0.650.01<-data.sample$n2[data.sample$power ==
0.65 & data.sample$alpha == 0.01]

powalphn2s0.650.05<-data.sample$n2[data.sample$power ==
0.65 & data.sample$alpha == 0.05]

powalphn2s0.650.1<-data.sample$n2[data.sample$power == 0.65
& data.sample$alpha == 0.1]

powalphn2s0.650.15<-data.sample$n2[data.sample$power ==
0.65 & data.sample$alpha == 0.15]

powalphn2s0.650.2<-data.sample$n2[data.sample$power == 0.65
& data.sample$alpha == 0.2]


# 0.7

powalphph3s0.70.01<-data.sample$ph3succ[data.sample$power
== 0.7 & data.sample$alpha == 0.01]

powalphph3s0.70.05<-data.sample$ph3succ[data.sample$power
== 0.7 & data.sample$alpha == 0.05]

powalphph3s0.70.1<-data.sample$ph3succ[data.sample$power ==
0.7 & data.sample$alpha == 0.1]

powalphph3s0.70.15<-data.sample$ph3succ[data.sample$power
== 0.7 & data.sample$alpha == 0.15]

powalphph3s0.70.2<-data.sample$ph3succ[data.sample$power ==
0.7 & data.sample$alpha == 0.2]


powalphn2s0.70.01<-data.sample$n2[data.sample$power == 0.7
& data.sample$alpha == 0.01]

powalphn2s0.70.05<-data.sample$n2[data.sample$power == 0.7
& data.sample$alpha == 0.05]

powalphn2s0.70.1<-data.sample$n2[data.sample$power == 0.7 &
data.sample$alpha == 0.1]

powalphn2s0.70.15<-data.sample$n2[data.sample$power == 0.7
& data.sample$alpha == 0.15]

powalphn2s0.70.2<-data.sample$n2[data.sample$power == 0.7 &
data.sample$alpha == 0.2]
```

```
# 0.75

powalphph3s0.750.01<-data.sample$ph3succ[data.sample$power
== 0.75 & data.sample$alpha == 0.01]

powalphph3s0.750.05<-data.sample$ph3succ[data.sample$power
== 0.75 & data.sample$alpha == 0.05]

powalphph3s0.750.1<-data.sample$ph3succ[data.sample$power
== 0.75 & data.sample$alpha == 0.1]

powalphph3s0.750.15<-data.sample$ph3succ[data.sample$power
== 0.75 & data.sample$alpha == 0.15]

powalphph3s0.750.2<-data.sample$ph3succ[data.sample$power
== 0.75 & data.sample$alpha == 0.2]


powalphn2s0.750.01<-data.sample$n2[data.sample$power ==
0.75 & data.sample$alpha == 0.01]

powalphn2s0.750.05<-data.sample$n2[data.sample$power ==
0.75 & data.sample$alpha == 0.05]

powalphn2s0.750.1<-data.sample$n2[data.sample$power == 0.75
& data.sample$alpha == 0.1]

powalphn2s0.750.15<-data.sample$n2[data.sample$power ==
0.75 & data.sample$alpha == 0.15]

powalphn2s0.750.2<-data.sample$n2[data.sample$power == 0.75
& data.sample$alpha == 0.2]


# 0.8

powalphph3s0.80.01<-data.sample$ph3succ[data.sample$power
== 0.8 & data.sample$alpha == 0.01]

powalphph3s0.80.05<-data.sample$ph3succ[data.sample$power
== 0.8 & data.sample$alpha == 0.05]

powalphph3s0.80.1<-data.sample$ph3succ[data.sample$power ==
0.8 & data.sample$alpha == 0.1]

powalphph3s0.80.15<-data.sample$ph3succ[data.sample$power
== 0.8 & data.sample$alpha == 0.15]
```

```
powalphph3s0.80.2<-data.sample$ph3succ[data.sample$power ==
0.8 & data.sample$alpha == 0.2]


powalphn2s0.80.01<-data.sample$n2[data.sample$power == 0.8
& data.sample$alpha == 0.01]

powalphn2s0.80.05<-data.sample$n2[data.sample$power == 0.8
& data.sample$alpha == 0.05]

powalphn2s0.80.1<-data.sample$n2[data.sample$power == 0.8 &
data.sample$alpha == 0.1]

powalphn2s0.80.15<-data.sample$n2[data.sample$power == 0.8
& data.sample$alpha == 0.15]

powalphn2s0.80.2<-data.sample$n2[data.sample$power == 0.8 &
data.sample$alpha == 0.2]


# 0.85

powalphph3s0.850.01<-data.sample$ph3succ[data.sample$power
== 0.85 & data.sample$alpha == 0.01]

powalphph3s0.850.05<-data.sample$ph3succ[data.sample$power
== 0.85 & data.sample$alpha == 0.05]

powalphph3s0.850.1<-data.sample$ph3succ[data.sample$power
== 0.85 & data.sample$alpha == 0.1]

powalphph3s0.850.15<-data.sample$ph3succ[data.sample$power
== 0.85 & data.sample$alpha == 0.15]

powalphph3s0.850.2<-data.sample$ph3succ[data.sample$power
== 0.85 & data.sample$alpha == 0.2]


powalphn2s0.850.01<-data.sample$n2[data.sample$power ==
0.85 & data.sample$alpha == 0.01]

powalphn2s0.850.05<-data.sample$n2[data.sample$power ==
0.85 & data.sample$alpha == 0.05]

powalphn2s0.850.1<-data.sample$n2[data.sample$power == 0.85
& data.sample$alpha == 0.1]

powalphn2s0.850.15<-data.sample$n2[data.sample$power ==
0.85 & data.sample$alpha == 0.15]
```

```
powalphn2s0.850.2<-data.sample$n2[data.sample$power == 0.85
& data.sample$alpha == 0.2]


# 0.9

powalphph3s0.90.01<-data.sample$ph3succ[data.sample$power
== 0.9 & data.sample$alpha == 0.01]

powalphph3s0.90.05<-data.sample$ph3succ[data.sample$power
== 0.9 & data.sample$alpha == 0.05]

powalphph3s0.90.1<-data.sample$ph3succ[data.sample$power ==
0.9 & data.sample$alpha == 0.1]

powalphph3s0.90.15<-data.sample$ph3succ[data.sample$power
== 0.9 & data.sample$alpha == 0.15]

powalphph3s0.90.2<-data.sample$ph3succ[data.sample$power ==
0.9 & data.sample$alpha == 0.2]


powalphn2s0.90.01<-data.sample$n2[data.sample$power == 0.9
& data.sample$alpha == 0.01]

powalphn2s0.90.05<-data.sample$n2[data.sample$power == 0.9
& data.sample$alpha == 0.05]

powalphn2s0.90.1<-data.sample$n2[data.sample$power == 0.9 &
data.sample$alpha == 0.1]

powalphn2s0.90.15<-data.sample$n2[data.sample$power == 0.9
& data.sample$alpha == 0.15]

powalphn2s0.90.2<-data.sample$n2[data.sample$power == 0.9 &
data.sample$alpha == 0.2]

# panel plots of the effect of the power

par(mar=c(5.1,1,1,1))

par(oma=c(1,1,1,1))

par(mfrow=c(3,4))

# 0.4 plot all alpha

plot(c(0,300), c(350,1400), type="n",
xlab=expression(n[2]), ylab = "Successful Phase III", main
= bquote( ~ 1-beta == 0.4), pch = 1)
```

```
points(powalphn2s0.40.01, powalphph3s0.40.01,pch= 16,
col="black")

points(powalphn2s0.40.05, powalphph3s0.40.05,pch= 16,
col="red")

points(powalphn2s0.40.1, powalphph3s0.40.1,pch= 16,
col="blue")

points(powalphn2s0.40.15, powalphph3s0.40.15,pch= 16,
col="darkmagenta")

points(powalphn2s0.40.2, powalphph3s0.40.2,pch= 16,
col="green")

# 0.45 plot all alpha

plot(c(0,300), c(350,1400), type="n",
xlab=expression(n[2]), ylab = "Successful Phase III", main
= bquote( ~ 1-beta == 0.45), pch = 1)

points(powalphn2s0.450.01, powalphph3s0.450.01,pch= 16,
col="black")

points(powalphn2s0.450.05, powalphph3s0.450.05,pch= 16,
col="red")

points(powalphn2s0.450.1, powalphph3s0.450.1,pch= 16,
col="blue")

points(powalphn2s0.450.15, powalphph3s0.450.15,pch= 16,
col="darkmagenta")

points(powalphn2s0.450.2, powalphph3s0.450.2,pch= 16,
col="green")

# 0.5 plot all alpha

plot(c(0,300), c(350,1400), type="n",
xlab=expression(n[2]), ylab = "Successful Phase III", main
= bquote( ~ 1-beta == 0.5), pch = 1)

points(powalphn2s0.50.01, powalphph3s0.50.01,pch= 16,
col="black")

points(powalphn2s0.50.05, powalphph3s0.50.05,pch= 16,
col="red")

points(powalphn2s0.50.1, powalphph3s0.50.1,pch= 16,
col="blue")
```

```
points(powalphn2s0.50.15, powalphph3s0.50.15,pch= 16,
col="darkmagenta")

points(powalphn2s0.50.2, powalphph3s0.50.2,pch= 16,
col="green")

# 0.55 plot all alpha

plot(c(0,300), c(350,1400), type="n",
xlab=expression(n[2]), ylab = "Successful Phase III", main
= bquote( ~ 1-beta == 0.55), pch = 1)

points(powalphn2s0.550.01, powalphph3s0.550.01,pch= 16,
col="black")

points(powalphn2s0.550.05, powalphph3s0.550.05,pch= 16,
col="red")

points(powalphn2s0.550.1, powalphph3s0.550.1,pch= 16,
col="blue")

points(powalphn2s0.550.15, powalphph3s0.550.15,pch= 16,
col="darkmagenta")

points(powalphn2s0.550.2, powalphph3s0.550.2,pch= 16,
col="green")

# 0.6 plot all alpha

plot(c(0,300), c(350,1400), type="n",
xlab=expression(n[2]), ylab = "Successful Phase III", main
= bquote( ~ 1-beta == 0.6), pch = 1)

points(powalphn2s0.60.01, powalphph3s0.60.01,pch= 16,
col="black")

points(powalphn2s0.60.05, powalphph3s0.60.05,pch= 16,
col="red")

points(powalphn2s0.60.1, powalphph3s0.60.1,pch= 16,
col="blue")

points(powalphn2s0.60.15, powalphph3s0.60.15,pch= 16,
col="darkmagenta")

points(powalphn2s0.60.2, powalphph3s0.60.2,pch= 16,
col="green")

# 0.65 plot all alpha
```

```
plot(c(0,300), c(350,1400), type="n",
xlab=expression(n[2]), ylab = "Successful Phase III", main
= bquote( ~ 1-beta == 0.65), pch = 1)

points(powalphn2s0.650.01, powalphph3s0.650.01,pch= 16,
col="black")

points(powalphn2s0.650.05, powalphph3s0.650.05,pch= 16,
col="red")

points(powalphn2s0.650.1, powalphph3s0.650.1,pch= 16,
col="blue")

points(powalphn2s0.650.15, powalphph3s0.650.15,pch= 16,
col="darkmagenta")

points(powalphn2s0.650.2, powalphph3s0.650.2,pch= 16,
col="green")

# 0.7 plot all alpha

plot(c(0,300), c(350,1400), type="n",
xlab=expression(n[2]), ylab = "Successful Phase III", main
= bquote( ~ 1-beta == 0.7), pch = 1)

points(powalphn2s0.70.01, powalphph3s0.70.01,pch= 16,
col="black")

points(powalphn2s0.70.05, powalphph3s0.70.05,pch= 16,
col="red")

points(powalphn2s0.70.1, powalphph3s0.70.1,pch= 16,
col="blue")

points(powalphn2s0.70.15, powalphph3s0.70.15,pch= 16,
col="darkmagenta")

points(powalphn2s0.70.2, powalphph3s0.70.2,pch= 16,
col="green")

# 0.75 plot all alpha

plot(c(0,300), c(350,1400), type="n",
xlab=expression(n[2]), ylab = "Successful Phase III", main
= bquote( ~ 1-beta == 0.75), pch = 1)

points(powalphn2s0.750.01, powalphph3s0.750.01,pch= 16,
col="black")
```

```
points(powalphn2s0.750.05, powalphph3s0.750.05,pch= 16,
col="red")

points(powalphn2s0.750.1, powalphph3s0.750.1,pch= 16,
col="blue")

points(powalphn2s0.750.15, powalphph3s0.750.15,pch= 16,
col="darkmagenta")

points(powalphn2s0.750.2, powalphph3s0.750.2,pch= 16,
col="green")

# 0.8 plot all alpha

plot(c(0,300), c(350,1400), type="n",
xlab=expression(n[2]), ylab = "Successful Phase III", main
= bquote( ~ 1-beta == 0.8), pch = 1)

points(powalphn2s0.80.01, powalphph3s0.80.01,pch= 16,
col="black")

points(powalphn2s0.80.05, powalphph3s0.80.05,pch= 16,
col="red")

points(powalphn2s0.80.1, powalphph3s0.80.1,pch= 16,
col="blue")

points(powalphn2s0.80.15, powalphph3s0.80.15,pch= 16,
col="darkmagenta")

points(powalphn2s0.80.2, powalphph3s0.80.2,pch= 16,
col="green")

# 0.85 plot all alpha

plot(c(0,300), c(350,1400), type="n",
xlab=expression(n[2]), ylab = "Successful Phase III", main
= bquote( ~ 1-beta == 0.85), pch = 1)

points(powalphn2s0.850.01, powalphph3s0.850.01,pch= 16,
col="black")

points(powalphn2s0.850.05, powalphph3s0.850.05,pch= 16,
col="red")

points(powalphn2s0.850.1, powalphph3s0.850.1,pch= 16,
col="blue")

points(powalphn2s0.850.15, powalphph3s0.850.15,pch= 16,
col="darkmagenta")
```

```
points(powalphn2s0.850.2, powalphph3s0.850.2,pch= 16,
col="green")

# 0.9 plot all alpha

plot(c(0,300), c(350,1400), type="n",
xlab=expression(n[2]), ylab = "Successful Phase III", main
= bquote( ~ 1-beta == 0.9), pch = 1)

points(powalphn2s0.90.01, powalphph3s0.90.01,pch= 16,
col="black")

points(powalphn2s0.90.05, powalphph3s0.90.05,pch= 16,
col="red")

points(powalphn2s0.90.1, powalphph3s0.90.1,pch= 16,
col="blue")

points(powalphn2s0.90.15, powalphph3s0.90.15,pch= 16,
col="darkmagenta")

points(powalphn2s0.90.2, powalphph3s0.90.2,pch= 16,
col="green")


############################################################

# success rate of phase III

par(mfrow=c(1,1))

rph3alph3s0.01<-data.sample$prob3_succ[data.sample$alpha ==
0.01]

rph3aln2s0.01<-data.sample$n2[data.sample$alpha == 0.01]


rph3alph3s0.05<-data.sample$prob3_succ[data.sample$alpha ==
0.05]

rph3aln2s0.05<-data.sample$n2[data.sample$alpha == 0.05]


rph3alph3s0.1<-data.sample$prob3_succ[data.sample$alpha ==
0.1]

rph3aln2s0.1<-data.sample$n2[data.sample$alpha == 0.1]
```

```
rph3alph3s0.15<-data.sample$prob3_succ[data.sample$alpha ==
0.15]

rph3aln2s0.15<-data.sample$n2[data.sample$alpha == 0.15]


rph3alph3s0.2<-data.sample$prob3_succ[data.sample$alpha ==
0.2]

rph3aln2s0.2<-data.sample$n2[data.sample$alpha == 0.2]

# plot rate of phase III success

plot(c(0,300), c(0.8,1), type="n", xlab=expression(n[2]),
ylab="Rate of Successful Phase III trials", pch=1)

points(rph3aln2s0.01, rph3alph3s0.01, pch= 16)

points(rph3aln2s0.05, rph3alph3s0.05, pch= 16, col="red")

points(rph3aln2s0.1, rph3alph3s0.1, pch= 16, col="blue")

points(rph3aln2s0.15, rph3alph3s0.15, pch= 16,
col="darkmagenta")

points(rph3aln2s0.2, rph3alph3s0.2, pch= 16, col="green")

lines(lowess(data.sample$n2, data.sample$prob3_succ))

legend("bottomright", legend = c(expression(paste(alpha, "
= ", 0.01)),
expression(paste(alpha, " = ", 0.05)),
expression(paste(alpha, " = ", 0.1)),
expression(paste(alpha, " = ", 0.15)),
expression(paste(alpha, " = ", 0.2))),
col=c("black","red", "blue","darkmagenta","green"),lty =
c(1,1,1,1,1))

############################################################
# success rate of phase II

par(mfrow=c(1,1))

rph2alph3s0.01<-data.sample$prob2_succ[data.sample$alpha ==
0.01]

rph2aln2s0.01<-data.sample$n2[data.sample$alpha == 0.01]
```

```
rph2alph3s0.05<-data.sample$prob2_succ[data.sample$alpha ==
0.05]

rph2aln2s0.05<-data.sample$n2[data.sample$alpha == 0.05]


rph2alph3s0.1<-data.sample$prob2_succ[data.sample$alpha ==
0.1]

rph2aln2s0.1<-data.sample$n2[data.sample$alpha == 0.1]


rph2alph3s0.15<-data.sample$prob2_succ[data.sample$alpha ==
0.15]

rph2aln2s0.15<-data.sample$n2[data.sample$alpha == 0.15]


rph2alph3s0.2<-data.sample$prob2_succ[data.sample$alpha ==
0.2]

rph2aln2s0.2<-data.sample$n2[data.sample$alpha == 0.2]

# plot of phase II success rate

plot(c(0,300), c(0,.6), type="n", xlab=expression(n[2]),
ylab="Rate of Successful Phase II trials", pch=1)

points(rph2aln2s0.01, rph2alph3s0.01, pch= 16)

points(rph2aln2s0.05, rph2alph3s0.05, pch= 16, col="red")

points(rph2aln2s0.1, rph2alph3s0.1, pch= 16, col="blue")

points(rph2aln2s0.15, rph2alph3s0.15, pch= 16,
col="darkmagenta")

points(rph2aln2s0.2, rph2alph3s0.2, pch= 16, col="green")

lines(lowess(data.sample$n2, data.sample$prob2_succ))

legend("topleft", legend = c(expression(paste(alpha, " = ",
0.01)), expression(paste(alpha, " = ", 0.05)),
expression(paste(alpha, " = ", 0.1)),
expression(paste(alpha, " = ", 0.15)),
expression(paste(alpha, " = ", 0.2))),
col=c("black","red", "blue","darkmagenta","green"),lty =
c(1,1,1,1,1))
```