Impact of Phase II Trial Design Choice in Oncology

Meredith Anne Martyn

Institute of Clinical Trials and Methodology, UCL

Submitted for the degree of Doctor of Philosophy

Declaration

I, Meredith Anne Martyn, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signature:

Date: 30/12/2023

Abstract

This thesis explores the influence of phase II trial design on the success rates of phase III trials in oncology, given that phase III trial failures have been seen to reach 58%. Historically, single-arm phase II trials were considered standard, but the increasing use of randomised-controlled trials in the last two decades has sparked a debate on optimal design choice.

This thesis considers the implications of each phase II trial design on the drug development pipeline. A narrative synthesis reveals the need for a methodological simulation study that assesses impact of phase II design choice while considering the following key elements: 1) end-of-phase III trial decisions, 2) both alternative and null hypotheses, 3) historical control error, 4) differing phase endpoints, 5) imperfect correspondence of treatment effect between phases.

The first simulation study addresses key elements 1, 2, and 3. However, implementing the remaining key elements; differing phase endpoints and imperfect correspondence; proves challenging. This highlights the struggle seen in practice with using phase II response rates to predict phase III survival outcomes. Therefore, methodology is developed in a proof-of-concept simulation study that considers key elements 1, 4, and 5. Finally, combining the methodology developed throughout, the last study integrates all key elements, drawing parameters from published pairings of phase II-phase III trials. The results compare the effectiveness of each trial design, emphasizing the implications of the chosen design on the overall development plan performance. Additionally, this final study introduces an innovative empirical approach for incorporating imperfect correspondence in simulation studies.

This thesis provides valuable insights for phase II investigators in selecting an optimal phase II design that will benefit a drug development plan as a whole, providing an opportunity to improve phase III cancer trial success rates.

Impact statement

The potential impacts of this thesis are outlined under future methodological research and design of phase II trials.

Future methodological research

This thesis can enhance future research through the work conducted in the narrative synthesis. From this, five key elements were identified as crucial to quantitatively assess the impact of phase II design choice on a development plan as a whole. These can be used as a checklist for future researchers to consider in quantitative studies that compare phase II trial designs.

The final simulation study that was developed throughout this PhD has the highest impact potential from my whole thesis. Not only is it the first quantitative study to consider all key elements, but it was based on five published phase II-phase III trials which allowed the simulations to reflect real clinical trial environments. It assessed how changing phase II design choice, specifically from a randomised-controlled design to a single-arm design, impacted the performance of the development plan. Performance indicators included development plan sample size and likelihood of making correct treatment decisions.

The final study can be extended to include other settings of phase II-phase III trials to assess a wider range of cancer clinical trial environments. Not only this, the simulation study can be extended further to include alternate phase II designs, which can also be in the setting of other disease areas. Moreover, some of the methodology used can be applied to other research areas, more specifically, to investigate the translation between phase II response rates and phase III survival outcomes.

Design of phase II trials

The findings of my simulation study can be used by phase II investigators to inform choice of phase II trial design. For example, the study shows that if a phase II trial is conducted but the hypothesised treatment effect is beyond a reasonable level of uncertainty, then a randomised phase II trial should be chosen. Furthermore, the simulation study demonstrates the ramifications of choosing conservative estimates at the design stage of a single-arm phase II trial. While some investigators may make this choice to reduce phase II sample size compared with a randomised trial, findings suggest the subsequent phase III trials demand a much larger number of participants than otherwise.

Finally, this thesis can be combined with the quantitative methodology papers identified in my narrative synthesis. These studies could be used to form a basis of formal guidance for investigators to choose an optimal phase II design. Better informed decisions regarding phase II trial design will have ripple effects on the success rates of cancer drug development plans, which will ultimately allow cancer patients to have access to novel treatment sooner. Not only this, but as the research has potential to be extended to include alternative phase II designs and other disease areas, it allows improvements to be made to a wider scope of drug development plans. Overall, the impact of the research conducted in this thesis can help improve access to novel treatments on a national level.

Acknowledgements

I used to jest that the source of motivation for this PhD was the unlimited milage that would come with becoming Dr Martyn in Doc Martens, but in reality, it took an awful lot more than that.

There's a very long list of people I'd like to thank. I'd like to thank my supervisors, Hakim, Tim and Max. I am particularly grateful to Hakim and Tim for transforming my project to what it is today, the kindness and patience you've given me. Extra thanks go to Tim for the time and care with the last leg of this thesis.

I would also like to give a huge thank you to my chosen family, whom I'm very lucky to have. First to my academic Ma n' Pa; Dave and Yean. You have been there through every thesis crisis, from picking me outside my office when broken-spirited, to zoom pomodoro-study sessions, to writing retreats, to hauling my bearded dragon's vivarium to hide from my landlord (officially, that is a joke). I want to give wholehearted thanks to my sister and future brother-in-law for their all-too-literal support when I needed it most, particularly this past year. I promise I'll never make you miss Eurovision to take me to A&E again. I'd like to thank The Lads, the most wholesome group of lovely humans who are unrelenting in their encouragement, particularly to time-keepers Alice and Reb. Huge thanks to the Extended Bethnal Green Universe™, particularly Miguel and Lucie and their safe haven of unlimited tea. Thank you to library-buds, Flav and Evey. I'd also like to thank my Mum, my supplier of nostalgic Australian snacks when I needed an extra boost for the all-nighters. I'd also like to thank friends from the MRCCTU who have leant a listening ear when things got tough; Lizzie, Leanne, Ellen, Alex, Hibo, Andy, Henry, Jeremy, Jingyi and Claire. There have been many, many events which has made this process particularly gruelling, without all of your support I wouldn't have got this far. And to those who doubted me, thanks for fuelling my stubbornness.

And finally, I'd like to dedicate this thesis to Ross Martyn. A while ago I showed you a pair of Doc Marten boots and made a silly comment about wearing them at graduation. Six years later, the boots are still here, but you are not. I know you reassured me that I was under no obligation to keep any promises I might have made you, but I hope you're proud I kept this one all the same. Miss you Dad.

Table of Contents

List of tables

List of Figures

1 Introduction

1.1 Clinical Trials and Drug Development Plans

Clinical trials facilitate medical innovation. When a new drug is discovered, it undergoes multiple rounds of testing in controlled environments, to gather information about its pharmacodynamic and pharmacokinetic properties, safety and efficacy (11). Trials are conducted in patients and are designed to collect high-quality data with minimal bias. These trials provide the evidence necessary to regulatory bodies, like the National Institute for health and Care Excellence (NICE) in the UK, to make informed decisions on public funding for these new drugs.

Before granting approval for the wider population, many questions must be addressed. These include questions surrounding safety, optimal dose, side-effects, comparative effectiveness against the current treatment and long-term impacts. It is impossible to answer all objectives in a single study. Furthermore, different research aims require larger sample sizes. Without addressing safety concerns first, many trial participants could be exposed to a potentially dangerous drug. Therefore, a phased testing approach is used. Initial phases begin on a smaller scale and primarily investigate the concerns of potential harm to patients (12). Once safety is established, investigators have the freedom to explore dosing effects and whether the drug works as intended (12). Once understood, larger trials can be conducted to assess if the new treatment is better than the existing standard of care (13). This strategy allows for rigorous investigation of a new drug while minimising risk to participants.

Generally, there are six stages in the development of a new drug: pre-clinical studies, phase 0, phase I, phase II, phase III and phase IV (14).

It is in preclinical studies where drug discovery begins. These studies are conducted in laboratory settings encompassing both in vitro (tissue culture studies) and in vivo (animal) experiments (15). Preclinical studies can be grouped into two categories; exploratory and confirmatory research (16). In the exploratory research, hypotheses are generated to explore ways in which the biological pathways of diseases can be altered to treat a disease, called “targets” (17, 18). Confirmatory research aims to explore the targets and validate their potential. For this, researchers initially explore which drug compounds affect the intended drug targets in cellular models (18, 19). Promising compounds then undergo trials on animals to evaluate dosing and toxicity effects. The dose and compound of the drug is refined until a candidate drug is finalised (18). This candidate drug can then proceed to human trial testing (18).

Phase 0 is the first stage of testing in humans and employs a proof-of-concept approach to assess the new drug in question. The primary aims are to assess whether the new drug modifies the biological target within the body as intended while gathering data on pharmacokinetics and pharmacodynamics for subsequent refinement (20, 21). To minimise exposure to the patients recruited, trials are brief, involve minimal doses and typically enrol groups of 10-15 patients (20, 21). These participants often have drug-resistant conditions or indolent diseases (20). Although individuals in phase 0 trials are not expected to experience any therapeutic benefits from the treatment, the information provided by these trials paves the way for investigations on drug effectiveness in people who might benefit (20, 21).

Historically, phase I trials have focussed on ensuring that an experimental treatment is well tolerated across a diverse range of patients in advanced stages of disease (12). They are typically small, open-label studies with around 10-100 patients. Although the majority of the assessment is centred around the drugs toxicity, initial investigations into dosing, such as maximum tolerated dose, can also be explored (12, 14, 22). More recently, phase I trials have acquired additional dimensions. They can now include early assessments of efficacy and the exploration of new drug combinations, adding more flexible research objectives (12, 23). From these trials it can be determined which drug or regimen is safe to proceed to phase II testing (12).

Phase II trials explore the impact of the new experimental drug within the intended patient population for the first time and determines whether it is worth investigating in a large phase III trial (24). As such, phase II trials can be broadly grouped into two categories: phase IIA and phase IIB. Phase IIA trials generally assess activity of an experimental drug, but can also investigate dose-response relationships, dosing regimen and toxicity in the context of the targeted patient population (14). Phase IIB trials generally have a decision-making aim as to whether the experimental drug should proceed to a phase III confirmatory trial (24, 25). However, both phase IIA and phase IIB trials offer a plethora of design choices due to their wide range of aims, but usually have sample sizes ranging from 30 to a few hundred patients and last approximately 18-36 months (24, 26, 27).

Phase III trials represent the pivotal stage of the drug development process, providing the evidence needed for regulatory bodies to assess whether the drug should be available for the public. The main aim is to confirm treatment efficacy on a large scale, usually in comparison to the standard of care (13). These trials also estimate incidence of side-effects, involving a substantial cohort of 300-3,000 patients (13). At this stage, the gold standard randomised-controlled trial is used to collect evidence about the treatment effect whilst minimising bias (28). The importance of phase III trials lies in their ability to estimate treatment effects that closely reflect reality to be used as a blueprint for the expected impact on the wider population.

Finally, Phase IV trials can be considered an extension of the drug development pipeline. These phases gather long-term observational data on approved drugs to provide insights on the practical implications of the drug on the broader population (29). The sample sizes can reach up to 5000 patients and can last several years (30, 31). Not only this, but they can also offer exploration in patient populations that were not sufficiently investigated in prior trials, and provide updated toxicity information (32). This final phase allows promising drugs to maximise their potential, giving an opportunity for additional hypothesis generation to explore the application of the new treatment in different disease areas.

The entire journey of drug development, starting from identifying targets to obtaining licensing for wider distribution, typically spans 12-15 years and comes with a price tag exceeding £800 million ($1 billion in the USA) (17, 18). If a potentially useful treatment is overlooked, the efforts of the researchers and participants will be in vain and would be denying the benefits to future patients. Alternatively, misjudging a futile drug as promising can cause delays in the pipeline, hindering timely funding for genuinely effective treatments which limits access for those in need.

This thesis will focus on the clinical trial phases of drug development with a focus on phase II and phase III trials. I will now discuss inefficiencies seen in drug development phases and their consequences.

1.2 Inefficiencies in Development Plans

In medical research, opportunity costs are the health benefits that could have been realized had the funding been invested in another promising alternative intervention (33). The cost of misinvestment can be enormous, with reports that UK expenditure on medical research totalled £1.6 billion in 2018/19 (34). Consequently, researchers bear the responsibility of implementing best practices in designing and conducting trials to deliver the most trustworthy research outcomes as quickly as possible. This not only benefits the current participants and future patients but extends to those who might have gained from research in alternative treatments.

In 2016, the Biotechnology Innovation Organization conducted a first-of-its-kind study to analyse clinical trial success rates in novel drugs in the USA between 2006 and 2015 (35). This report inspected 7455 development plans and assessed the success of transitions between phase I trials, phase II trials, phase III trials and gaining approval from the Food and Drugs Administration (FDA). The study revealed that the average success rate from a phase I trial to regulatory FDA filing was 9.6% across 14 major disease categories made up of allergy, autoimmune, cardiovascular, chronic high-prevalence diseases, endocrine, gastroenterology, haematology, infectious disease, metabolic, neurology, oncology, ophthalmology, psychiatry, rare diseases, respiratory and urology (35). Notably, the likelihood of a phase I oncology trial reaching approval exhibited the lowest rate of 5.1% (35).

This result is alarming, especially considering that cancer is the second leading cause of death worldwide following cardiovascular disease (36, 37). Latest estimates indicate that approximately 18 million people are diagnosed with cancer each year, and it is the cause of around 9 million deaths (38). Currently, cancer holds the highest disease burden, as measured by Disability-Adjusted-Life-Years, a measure of impact of a disease on an individual (38). With projections suggesting global number of cancer deaths will surpass those from cardiovascular disease by 2060, there is an increasing urgency to prioritize the improvement of success rates in novel cancer treatment (38).

Given only 5.1% of phase I cancer trials lead to regulatory approval, almost half the overall success rate for all diseases combined, this begs the question: what are the main reasons for this low approval rate in oncology?

1.3 Cancer Clinical Trials

The report by the Biotechnology Innovation Organization uncovers further weaknesses in cancer drug development plans.

Oncology displays the greatest failure after phase III trials, with only 40.1% of new drugs successfully gaining FDA approval following a phase III trial, compared to an overall average of 58.1% (35). This is concerning, particularly as generally, phase III trials in isolation can cost approximately £10 million, representing a heavy opportunity cost (39). In theory, phase III trials are confirmatory in nature, providing investigators with formal estimates of treatment effects expected to be seen in practice (25). As previously mentioned, randomised-controlled trials and large sample sizes are commonplace at this phase to provide unbiased estimates. It is therefore a point of interest to explore the extent by which the methodology of prior phases may be hindering the performance of phase III cancer trials.

A substantial difference in the success of phase II trials was also identified in cancer drug development plans compared to other disease areas. Only 24.6% of oncology phase II trials progressed to phase III trials, compared to the overall average of 30.7% (35). The difference between the transition between phase I and phase II trials was less stark, with the success rate of oncology at 62.8% compared to 63.2% overall (35). It is possible that the diminished success of phase II cancer trials has a domino effect on the next stages of drug development.

The failings between phase II and phase III oncology trials may not come as a surprise to some researchers, as a fierce debate surrounding the correct choice of phase II cancer trial design has persisted for approximately 25 years (40).

Before the mid-1990s, single-arm trials were the favoured design in phase II trials (40). In the simplest form of this design, all participants were recruited in a single-stage and were given the experimental drug and observed for a period of time (41). The outcome of this group was then compared to an external benchmark, such as response rates seen in the control groups from previous trials, named “historical controls”, or a minimum expected response rate (41, 42). However, unreliable historical controls have led to misleading conclusions on treatment, giving rise to the use of randomised-controlled trials in phase II trials in recent years (43, 44). It should be noted that the favoured use of single-arm trials pre-dated the groupings of phase IIA and phase IIB trials, however, both randomised and single-arm trials are still commonplace in both settings today (45-47).

Randomising allocation of patients into two or more groups within a trial minimises selection bias and confounding bias (48). Assuming the simplest design, i.e. single-stage parallel group design, each patient has an equal chance to be recruited to each treatment group. This means the likelihood of systematic differences between the patient groups is reduced. Moreover, randomisation protects against selection bias from an investigator wanting to recruit a patient into a specific group (49-51). This means patient characteristics between each of the treatment groups should be approximately similar, including for known and unknown confounders, meaning any differences found between the patient groups can be assigned to the treatment effect (49, 51, 52).

Each design brings advantages and disadvantages. Although the randomised-controlled design is seen as the gold standard in trial design, some researchers view this as misplaced in a phase II setting due to the increased demand of participants and time compared to a single-arm trial. To follow, the use has also been criticised as redundant, as a potential subsequent phase III trial would conduct the gold-standard randomised-controlled trial to gather unbiased estimates. Additionally, ethics have been brought forward to this debate, particularly if the trial has large effect sizes, as participants in the control group would be denied the better treatment (40).

Single-arm trials are more cost-efficient and less time-consuming than a randomised trial. Due to the single-arm design throughout, these trials can achieve higher levels of power than a randomised trial with the same number of participants. Additionally, as cancer clinical trials have been conducted for decades, there are many well-established sources to choose a historical control. This reduces the risk associated with using an unreliable historical control compared to earlier times. However, the lack of randomisation exposes weaknesses in the trial design, particularly as there may be prominent differences between the treatment group and the historical control chosen that are unknown. This may lead investigators to misattribute differences in patient groups as treatment effect, and potentially overinflate effectiveness of the drug (53, 54).

A summary of the is seen in Table 1. Presents a summary of the advantages and disadvantages of each design as seen in Grayling et al (40).

Consideration	Randomised-controlled trials	Single-arm trials
Bias	Unbiased treatment effect No confounding when sample size is large enough	If historical control is poorly chosen it can lead to biased treatment effect More susceptible to selection bias
Cost	Expensive, requires more patients and so more time	More cost-effective, requires fewer patients and less time
Patient-acceptability	Randomisation may dissuade patients to enter trial	Guarantee of experimental treatment encourages patients to enter trial
Power	Requires more patients to achieve the same level of power. Alternatively, achieves less power with the same number of patients	Requires less patients to achieve the same level of power. Alternatively, achieves the more power with the same number of patients

Table 1 – Table to show advantages and disadvantages of randomised-controlled trials and single-arm trials in a phase II setting *(40)*

There is no consensus to guide phase II investigators on the choice of optimal phase II trial design, given their specific clinical trial parameters (55). An agreed consensus could improve the success rates of phase II cancer clinical trials and subsequent phase III cancer trials. This thesis aims to identify the circumstances in which each design is preferred to help answer this question. This in turn can streamline the cancer drug development process and allow patients to access life-saving treatment sooner.

1.4 Outline of Thesis

The structure of my thesis is as follows:

- Conduct a narrative synthesis on methodological papers comparing single-arm and randomised controlled trials in a phase II cancer setting with respect to their link with subsequent phase III trials.
- Conduct my own methodological simulation study comparing aspects that have not been investigated.
- Conduct a simulation study using applications from published trials in a phase II-III cancer setting.

2. Narrative Synthesis

2.1 Introduction

For this narrative synthesis, I will first describe one of the first methodological studies published which compared the performance of single-arm trials and RCTs in a phase II setting. I will use it as a baseline to compare with other papers that are identified through this narrative synthesis. I will then identify the research gaps in existing literature, some of which I will address in the rest of my thesis.

2.1.1 Taylor et al.

In 2006, Taylor *et al.* published a seminal paper titled, "Comparing an experimental agent to a standard agent: relative merits of a single-arm or randomized two-arm phase II design" (56). This paper simulated phase II single-arm trials and randomised-controlled trials to assess which design was more likely to conclude in favour of a truly effective treatment. Additionally, Taylor *et al.* calculated the proportion of active phase III trials which would result from these phase II designs. This was one of the first papers to conduct a study to address appropriate phase II cancer trial design choice.

Taylor *et al.* defined P_{0i} as the proportion of true binary response rate in the control treatment arm in institution, i, within a phase II study. It goes on to define P_{1i} as the proportion of true binary response rate in the experimental treatment for institution i. ${P^*}_i$ was then defined as the proportion of binary response rate in a chosen historical control treatment arm for each institution i. Levels of P_{0i} assessed were 10% or 30%. Multiplicative treatment effect from institution i was defined as $\delta_i \geq 1$, such that $P_{1i} = P_{0i}\delta_i$. Average treatment effect, $(1 + \mu_\delta)$, ranged from 1.0-1.8. Levels of sample size assessed for each trial simulated was 30 or 80.

Taylor *et al.* focused on the impact of inter-institution variability on the outcomes from each phase II trial design. To account for this, variability in estimated P_{0i} and ${P^*}_i$ were defined by parameters ω and ϕ which ranged from 0-0.2. Inter-institution variability in δ_i, which is uniformly distributed, was defined by θ which ranged from 0-0.1. Additionally, the authors considered impact on decision thresholds, Δ, specified at either 0 or 0.05.

The authors also explored impact on subsequent phase III trials by simulating five phase II single-arm trials and five phase II randomised-controlled trials with similar parameters described above, but with sample sizes of 40 and with the null hypothesis considered. The proportion of subsequent positive phase III trials was calculated as the proportion of phase II trials that observed experimental response rate to be at least 5% better than the standard.

The results of the simulations provided useful insight: in the presence of high inter-institution variability, single-arm trials were less likely to correctly conclude in favour of treatment than a randomised-controlled trial. On the other hand, when each design had a sample size of 30, randomised-controlled trials were less likely to correctly conclude in favour of the treatment. When considering impact on phase III trials, single-arm trials were less likely to incorrectly recommend a phase III trial in the presence of little variability and under the null hypothesis. Single-arm trials were also more likely to correctly recommend a phase III trial under little variability when treatment effect was small – otherwise, both trial designs performed similarly. The paper ultimately recommended the use of single-arm trials, unless large inter-institution variability was anticipated.

One major criticism of Taylor *et al.* was the limited number of values chosen for the clinical trial parameters. For example, the two sample sizes chosen were restricted to 30 or 80. In practice, clinical trials are not limited to these two choices and instead, sample size can be based on a calculation with pre-specified α and power levels. The values of the parameters explored were also limited, only allowing for two levels of response rates and treatment effects. This meant that the conclusions could not be easily generalised to the typical performance of many phase II trials. Additionally, it is unlikely that different institutions within the same phase II clinical trial would use different historical response rates. Instead, it is common practice to choose one level of response rate to use as a global historical control. Another limitation of Taylor *et al.* is that it is simplistic. For example, it is well reported that there are discrepancies between size of treatment effects found in phase II and subsequent phase III trials. Explanations for this phenomenon include the use of different endpoints between phase II and phase III trials, the differing α and power levels used for sample size calculations and differences in populations (57). However, Taylor *et al.* assumed that all phase II trials that showed at least a 5%-point difference between experimental response rate and control response rate would translate into a positive phase III trial.

Taylor *et al.* conducted the first study to compare the methodology of single-arm trials and randomised-controlled trials in a phase II context. Although the applicability of Taylor *et al.* was limited, it served as a foundation for many medical statisticians to provide further evidence to decide when it is appropriate to use a specific phase II trial design.

To investigate all studies which have conducted research to compare the performance of phase II trial designs, in a similar way to Taylor *et al.,* I conducted a review. Many authors explored different ways to assess performance of phase II trial designs, therefore a more nuanced approach was needed to summarise all the findings of the methodological papers than a literature review. To allow for use of words and text to summarise and explain the findings, a narrative synthesis was chosen using appropriate guidance (58).

2.2 Methods

2.2.1 Search Strategy

To capture the trend of increasing use of alternative designs being used in a phase II setting in the past two decades (59), the narrative synthesis identified methodological studies published from 1st January 2000 to 18th January 2018 through PubMed. The search term included the MeSH terms: ‘randomised controlled trials as topic’ or ‘clinical trial as topic’. Additionally, the following words had to be included in either the title and/or abstract of the paper: uncontrolled, single-arm, one-arm trial, non-randomised, non-controlled, historical control. Spelling variations were accounted for.

To identify papers for the final narrative synthesis, a screening process to assess relevance was completed which involved reading titles, then abstracts, then full texts.

Inclusion criteria are listed below:

- Methodological papers

- Comparison of performance between traditional single-arm trial designs and traditional randomised-controlled trial designs in phase II settings
- Full text papers
- In English language
- Accessible by UCL library
- No duplicates
- Studies which generated novel quantitative data

Titles deemed irrelevant were excluded from the Narrative Synthesis. Title screening was chosen as the search criteria generated 1955 results, many of which were trial reports which were easily identifiable through the titles. Those with uncertain or clear relevance proceeded to abstract screening. This was repeated for the full text review. When relevance was uncertain, full text articles were verified with an independent reviewer. Novel data was determined as a study which generated quantitative outcomes, as opposed to summarising quantitative outcomes from previous studies.

2.2.2 Data extraction

Two types of data extraction were performed on the final papers chosen for review: descriptive data and quantitative study data.

The type of descriptive data collected from papers included year published, title, first author, objectives, methods, main results, strengths, weaknesses, and next steps/gaps.

For quantitative analysis, the different methodological papers assessed the two designs in different ways which made it difficult to directly compare the results. For example, some focussed on the impact of various degrees of historical control error in a single-arm trial, which was compared to a randomised controlled trial (56, 60-64). Some focussed on how rigorous each design faired against type I and type II error by assessing the proportion of correct conclusions made under the alternate and null hypothesis (56, 60-65). Other papers looked at the impact of the phase II study designs on the subsequent phase III trials (54, 56, 60, 62, 66). While considering this, two papers even accounted for the correspondence of treatment effect between phase II and phase III trials (54, 60). Another considered different endpoints between each phase (60).

Therefore, I made note of five key elements which were considered across the studies when comparing the two phase II designs. I then assessed how many of these elements each methodological paper considered.

Therefore, quantitative study data consisted of five yes/no questions describing the study parameters considered, which included:

- Did the study consider impact on subsequent phase III trial?
- Did the study consider both null and alternative hypothesis?
- Did the study consider historical control error?
- Did the study consider different phase II and phase III endpoints?
- Did the study consider correspondence between phase II and phase III treatment effect?

These elements would ascertain the robustness of each papers results. For example, the performance of the phase II trial in isolation from a subsequent phase III trial cannot provide information on the consequences of the new experimental drug on the development plan as a whole. Furthermore, it is important to reduce both type I and type II errors in a trial, therefore, considering the quality of performance under both the null and alternative hypothesis is necessary. Consideration of historical control error was deemed crucial as this is one of the most prominent criticisms of single-arm trial design. Finally, consideration of differing endpoints between phases and correspondence of treatment effect between phases was considered a necessary component to reflect real-practice of phase II-phase III development plans.

Results of data were compiled into tables. Details of the methodologies and data results of each paper are described and assessed.

2.3 Results

Figure 1 presents the screening process for final papers selected for narrative synthesis. It should be noted that some papers were excluded for more than one reason. In this figure, single-arm trial is abbreviated to SAT, and randomised controlled trial is abbreviated to RCT.

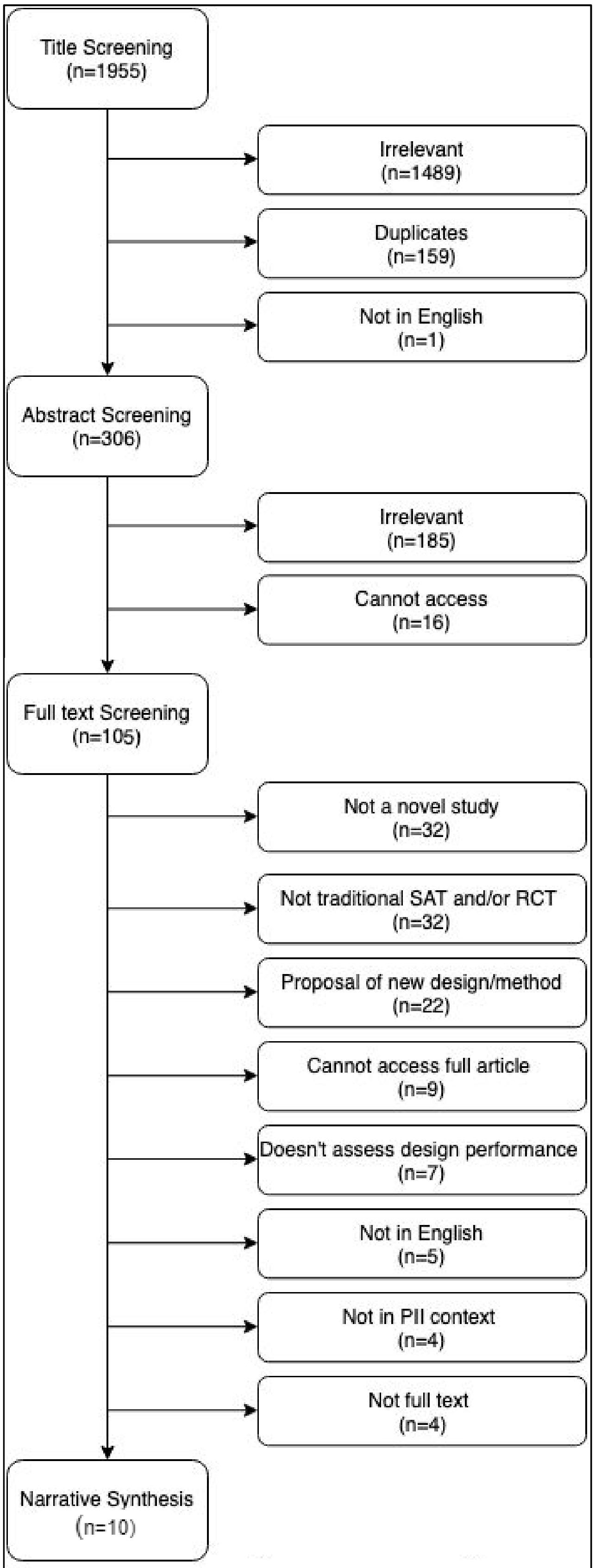


Figure 1 – Flow diagram showing selection of final papers from narrative synthesis

2.3.1 Paper Selection

The PubMed search identified 1955 papers published from 1st January 2000 to 18th January 2018. Search results were extracted into an excel spreadsheet and duplicates were identified through entries that had identical titles and authors. After title screening all papers, 306 proceeded to abstract screening, from which 105 proceeded to full text screening. 10 were then chosen to be included for the narrative synthesis, including the original paper by Taylor *et al.*

Papers that were labelled “irrelevant” from the title or abstract screening were rejected from the narrative synthesis as they were either stand-alone trial reports, used “non-randomised trials” as a definition for observational studies, were literature reviews or only investigated the performance of one trial design.

Of the 105 articles reviewed for full text screening, 32 were eliminated as they were not novel studies. These included review articles, opinion pieces, or papers which did not produce new quantitative data.

Another 32 were eliminated as they did not assess traditional single-arm trials or randomised-controlled trials. These included papers which assessed the performance of a specific type of single-arm trial or randomised-controlled trial design, or grouped results of single-arm trials with other observational studies to compare against randomised-controlled trials.

A further 22 papers were eliminated due to the paper proposing a new trial design or statistical method. These included new methods using Bayesian statistics to strengthen the validity of historical controls used in practice, introducing an adaptive element to the trial, or a new meta-analysis method to combine the results of both single-arm trials and randomised-controlled trials.

Seven papers did not strictly assess performance of single-arm trials against randomised-controlled trials. These included papers which gave descriptive statistics on the type of trials conducted in a phase II setting, or even the difference in the quality of evidence reporting between the two designs.

For nine papers I was not granted access to read the full article, and four that were not full texts were abstracts only. It should be noted that the one paper that was eliminated through title-screening for not being in English was checked at a later date to confirm that the rest of the text was also not available in English.

Ten methodological papers were chosen for the final narrative synthesis review. *Figure 1* illustrates the elimination process. Each of the nine papers (excluding the Taylor *et al.*) are described in the following Detailed Description section. They are described in alphabetical order from the first author.

2.3.2 Descriptive Data

Results of the descriptive data from the 10 chosen papers are presented in Table 2. A summary of the Taylor paper is given first and highlighted in yellow, and other papers selected for the narrative synthesis are in alphabetical order of the first author.

2.3.3 Quantitative Study Data

Results of the quantitative study data collected from the 10 chosen papers are presented in *Table 3*. Results of the Taylor paper is first and highlighted in yellow, and the remaining nine papers selected for the narrative synthesis are in alphabetical order of the first author. It should be noted that RCT used throughout this table stands for randomised-controlled trial.

It is apparent that very few strategies considered all the five main yes/no questions. The only paper that considered all five elements was Herberger *et al.*

Strikingly, most papers considered the null hypothesis when assessing performance of single-arm trials and randomised-controlled trials. However, four out of 10 papers did not consider historical control error. Specifically these papers are Grayling et al., Maitland et al., Monzon et al and Sharma et al. (54, 65-67). This is surprising, as one of the most prominent criticisms for single-arm trials is the level of bias that may be present in estimated treatment effect due to using a non-concurrent control arm (63).

As seen from the data, five out of 10 papers did not consider impact on subsequent phase III trials. This may be due to a large oversight, as performance of trials solely within a phase II setting does not necessarily translate to more effective treatments becoming available to the public. Of the five papers that did consider impact on phase III trials, only two considered the correspondence between phase II and phase III treatment effect, and only one considered different endpoints used. From this, it is clear that the consequences of phase II design choice on phase III trials needs further evaluation.

ID #	Year	Author	Title	Objectives	Methods	Main results	Strengths	Issues not addressed	Next Steps/ Gaps
1	*2006*	*Taylor et al.*	*Comparing an experimental agent to a standard agent: relative merits of a one-arm or randomized two-arm phase II design (56)*	*To investigate impact of variability on conclusions of a treatment in single-arm trials and RCTs*	*Simulations of RCT and single-arm trials to explore conclusions on treatment*	*-Single-arm trials are more likely to lead to wrong conclusions with large variability* *-RCTs are more likely to lead to wrong conclusions with smaller sample size* *-Any difference between RCT and single-arm trials are negligible*	*-Authors considered multiple types of trial variabilities such as between-institution,* P_0 *and treatment effect variabilities* *-Authors considered decision thresholds*	*-The paper only considered a small range of parameters, i.e. sample sizes of 30 or 80. This limits applicability of the results* *-Impact on phase III trials did not consider the conduct of the phase III trial itself.*	*-Extend range of trial parameters simulated to reflect real-life practice* *-Extend scope of impact on phase III trials*
2	2015	Grayling *et al.*	Do single-arm trials have a role in drug development plans incorporating randomised trials (67)	To compare various phase II development plans to assess most appropriate phase II design: single-arm trials and/or RCTs	-Authors simulated results of six different phase II development plans. -Authors assessed impact of clinician opinion by using a prior distribution	-Development plans that involve a group sequential design uses the smallest sample size to achieve the same level of power	- Authors compared trials as part of a phase II development plan -Accounted for stopping rules within a trial design -Considered more than one type of trial design -Considered four types of optimality criteria	-The authors did not consider historical control error - Assessed a stand-alone RCT, but not a stand-alone single-arm trial -Only considered one set of anecdotal parameters of P_0 and P_1	-Could account for historical control error -Could explore wider range of P_0 and P_1 - Could assess results from a stand-alone single-arm trial

3	2009	Hunsberger *et al.*	A comparison in Phase II Study Strategies (60)	To compare different phase II study strategies to determine most efficient drug development path to make a conclusion on drug efficacy	-Simulations of four phase II-phase III development plans were completed -Main outcomes were total number of patients recruited, total time taken across a development plan, and probability of concluding correctly about the treatment	-In a single-arm trial, overestimating historical control can reduce the probability of correctly concluding in favour of the treatment from anticipated 81% to 9% -Integrated phase II/phase III development plan performed the most consistently, without sacrificing probability of concluding correctly and can do so with minimal patients and time.	-Authors considered multiple development plans - Considered amount of time it would take to complete a trial -Considered impact on phase III trial, correlation between phase II and phase III treatment effects and different phase II-phase III endpoints	-Phase II endpoint used was progression-free survival. It would be more realistic if it considered a binary endpoint - Only one set of anecdotal values considered for P_0, P_1 and treatment effect -Did not consider a stand-alone single-arm trial -Correlation of treatment effects between phase II and phase III only considered for three out of the four development plans simulated	-Could consider binary endpoints for phase II -Could consider a wider range for P_0, P_1, treatment effect and historical control error -Could include correlation of phase II and phase III treatment effect for the remaining development plan
4	2010	Maitland *et al.*	Analysis of the yield of Phase II Combination Therapy Trials in Medical Oncology (66)	To investigate the hypothesis that phase II single-arm trials lead to a low proportion of practice-changing phase III trials	-A database search was conducted to identify phase II “combination chemotherapy” trials and associated phase III trials -Authors investigated associations between phase II trial characteristics and practice-changing phase III trials	-phase II RCTs were more likely to draw negative conclusions than single-arm trials, to a statistically significant degree	-Authors used data from real phase II and phase III trials -Authors prospectively identified phase III trials from phase II trials	-results of 22 phase II RCTs compared against the results of 341 phase II single-arm trials -Authors only looked at phase III trials which “changed clinical practice”, so the breakdown of negative phase III trials is unknown -Little else is investigated comparing phase II single-arm and RCT designs.	-Could have searched for phase II trials beyond combination chemotherapy and beyond trials published between 2001-2002 -Could have investigated designs of phase II trials that led to negative phase III trials -Could have investigated further phase II characteristics such as sample size, treatment effect size, P_0 and P_1

5	2015	Monzon *et al.*	Correlation of single arm versus randomised phase 2 oncology trial characteristics with phase 3 outcome (54)	To investigate the hypothesis that phase II single-arm trials are worse at predicting phase III outcomes than phase II RCTs	-Database search in the top journals identified phase III oncology trials and associated phase II trials -Authors compared associations of phase III outcomes with phase II trial design	-There was no significant difference in the association between positive phase III outcome and phase II design. -The only significant finding was that positive phase II trials were more likely to lead to positive phase III trials	-Authors used data from real trials -Authors identified phase II trials and their associated phase III trials to assess phase II impact on phase III trials	-The authors only looked at top 16 oncology journals, which could introduce selection bias -80% of phase II trials found were single-arm trials -Definition of "positive phase II trial" includes 95% CI which is not appropriate for all phase II settings -Did not look at designs of negative phase II trials -Did not adjust for multiple testing	-Could conduct systematic search to identify phase III trials, and could further identify negative phase II trials -Could adjust definition of "positive phase II" trial to be more appropriate -Could stratify by characteristics of phase II design i.e. sample size
6	2014	Moroz *et al.*	Comparison of anticipated and actual control group outcomes in randomised trials in paediatric oncology provides evidence that historically controlled studies are biased in the favour of novel treatment (61)	To recreate Paediatric Oncology RCTs as single-arm trials using P_0 used in sample size calculation, and compare outcomes	-Database search to find eligible RCTs that have sufficient reporting of sample size calculation	-Single-arm trials overestimated treatment effect by an average of 3.8% -There was a significant difference in treatment effects estimated between RCTs and single-arm trials	-Authors used real data from 47 RCTs -Authors conducted a sensitivity analysis on types of trials i.e. superiority, equivalence, and non-inferiority	-The authors did not distinguish between phase II and phase III RCTs -RCTs chosen were conducted in Paediatric Oncology, therefore has limited application -RCT was only selected if they had sufficient reporting of sample size calculation - if it included trials with poorer reporting the results might have been different - Authors only used published trials, therefore might have had publication bias in RCTs chosen	-Explore RCTs for all cancers -Consider unpublished trials/ null trials i.e. trials registered on AllTrials.net that have not published data -Separate trials into phase II or phase III, and by sample size -Could contact authors for underreported trials for sample size calculation

7	2011	Pond *et al.*	Quantitative evaluation of single-arm versus randomized phase II cancer clinical trials (62)	To improve on the Taylor *et al.* paper with values of parameters that are more reflective of true practice	-Simulations of RCT and single-arm trials completed at phase II level that Taylor *et al.* used with adjusted parameters -Accounted for historical control error and proportion of active phase III trials	-In the presence of historical control error and between-institution variability, single-arm trials had a lower proportion of subsequent active phase III trials than phase II RCTs	-Authors used more realistic parameters than the Taylor *et al***.** paper -Authors accounted for historical control error and subsequent proportion of active phase III trials	-Phase II trial designs did not have matching error rates -Only explored limited range of historical control error i.e. ± 10%, ± 5% or 0% -Assessed phase III single-arm trials following phase II single-arm trials which is not reflective of real practice -Limited range of P_1 explored i.e. only considered P_1=P_0, P_1=P_0+0.15 or P_1=P_0+0.2	-Could match error rates of phase II trials -Could consider only RCT design at phase III -Could explore wider range of historical control error and P_1
8	2015	Sambucini *et al.*	Comparison of single-arm vs. randomised phase two clinical trial: Bayesian approach (63)	To compare ability of RCTs and single-arm trials in making the correct decision about treatment	-Simulations using Bayesian statistics using investigators opinion as a prior	-With historical control error present, single-arm trials were more likely to make correct decisions about treatment when the treatment effect was large -The more sceptical an investigator was about the historical control used, the more an RCT is preferred	-Authors assessed performance of trial designs by looking at likelihood of concluding correctly about treatment, a good performance measure -It considered the opinion of an investigator -It compared trials with the same sample size, which is often a limiting factor in practice	-The authors only looked at scenarios where P_1 was a maximum of P_0+0.15 -The authors only considered the case where historical control was overestimated	-Could consider further flexibility of P_1>P_0+0.15 -Could consider underestimation of historical control -Could consider additional trial design aspects i.e. decision threshold, allocation ratio -Could consider impact on phase III -Could consider typical values of α and power levels in phase II and phase III trials

9	2012	Sharma *et al.*	Resampling phase III data to assess phase II trial designs and endpoints (65)	To resample patient data from existing phase III trials to recreate and compare phase II RCT and single-arm trials	-Authors identified a negative and positive phase III trial (TARGET and the AE 941). They resampled patient data to recreate phase II RCT and single-arm trials, and compared true negative and true positive rates	-Single-arm trials had true positive rates of 55%, vs RCTs which had 96.5% -Single-arm trials had false positive rates of 0.9%, vs RCTs which had 25%	-Authors used real data -Authors used both positive and negative phase III trials with similar parameters from which to resample their data	-The authors only looked at two trials -Only one type of single-arm trial was assessed compared to multiple types of RCTs. -Authors reported that the example of the negative phase III trial was perhaps not truly negative	-Could use additional pairs of phase III trials to resample from -Could include different single-arm trial designs -Could simulate a true null situation -Could vary α and power levels -Could compare results of trial designs that use the same sample size
10	2010	Tang *et al.*	Comparison of Error Rates in Single-Arm Versus Randomized Phase II Cancer Clinical Trials (64)	Improvement on Taylor *et al.* paper using more realistic parameters	-Recreate a single-arm trial by resampling patients from the positive N9741 trial for the experimental arm, and use historical control arm from Saltz trial -Simulate RCT and single-arm trials and compare the prevalence of error rates	-If N9741 trial was run as a single-arm trial, it would have rejected the effective treatment -As variability increases within trials, single-arm trials have higher than anticipated error rates, while RCT error rates remain the same	-Authors used data from real trials -Authors adapted simulation from Taylor *et al.* paper so it was more reflective of real-life practice	-The authors only recreated one version of a single-arm trial. They could have explored multiple other trials that use historical controls as a comparison. -Comparison of trials with the same sample size was not made, which is usually a limiting factor in real practice. -Null condition not explicitly discussed.	-Further exploration of recreating N9741 single-arm trial -Consider RCT and single-arm trials of the same size -Consider impact on phase III -Consider typical values of α and power levels in phase II and phase III trials

Table 2 – Summary of final papers selected by narrative synthesis

ID #	Paper	Author	Impact on phase III considered? Y/N	Null considered? Y/N	Account for historical control error? Y/N	Different phase II /phase III endpoints considered? Y/N	Correspondence between phase II and phase III treatment effect considered? Y/N
1	*Comparing an experimental agent to a standard agent: relative merits of a one-arm or randomized two-arm PII design*	*Taylor et al.*	*Y*	*Y*	*Y*	*N*	*N*
2	Do single-arm trials have a role in drug development plans incorporating randomised trials	Grayling *et al.*	N	N	N	N	N
3	A comparison in Phase II study strategies	Hunsberger *et al*	Y	Y	Y	Y	Y
4	Analysis of the yield of phase II combination therapy trials in medical oncology	Maitland *et al.*	Y	Y	N	N	N
5	Correlation of single arm versus randomised phase 2 oncology trial characteristics with phase 3 outcome	Monzon *et al.*	Y	Y	N	N	Y

6	Comparison of anticipated and actual control group outcomes in randomised trials in paediatric oncology provides evidence that historically controlled studies are biased in the favour of novel treatment	Moroz *et al.*	N	N	Y	N	N
7	Quantitative evaluation of single-arm versus randomized phase II cancer clinical trials	Pond *et al.*	Y	Y	Y	N	N
8	Comparison of single-arm vs. randomised PII clinical trial: Bayesian approach	Sambucini *et al.*	N	Y	Y	N	N
9	Resampling phase III data to assess phase II trial designs and endpoints	Sharma *et al.*	N	Y	N	N	N
10	Comparison of error rates in single-arm versus randomized phase II cancer clinical trials	Tang *et al.*	N	Y	Y	N	N
		Total	**5**	**8**	**6**	**1**	**6**

Table 3 – Summary of quantitative information extracted from final papers selected by narrative synthesis

2.3.4 Detailed Description of Each Study Selected by Narrative Synthesis

In this section I provide detailed descriptions of the nine additional papers to the Taylor *et al.* paper that were selected for the narrative synthesis. They are described in alphabetical order of the first author.

2.3.4.1 Grayling et al. - Do single-arm trials have a role in drug development plans incorporating randomised trials?

In 2015, Grayling *et al.* extended the scope of existing research by including development plans which make up more than one trial in a phase II setting (67). The development plans were assessed by four optimality criteria: smallest expected sample size, smallest expected sample size given a pre-specified threshold for power, maximum power-per-patient, and maximum power-per-patient given a pre-specified threshold for power.

The paper performed two analyses: 1) development plan performance and 2) impact on performance given clinicians' opinions.

To assess development plan performance, the authors simulated six development plans:

- A Simon's two-stage single-arm trial followed by a randomised-controlled trial.
- A Simon's two-stage single-arm trial that allows for early stopping, followed by a randomised-controlled trial.
- A single-stage randomised-controlled trial.
- A 3-stage group-sequential randomised-controlled trial with early go/no-go stopping criteria.
- A group-sequential randomised-controlled trial with early go/no-go decisions based on sample size of each stage in Simon's two-stage single-arm trial.
- A group-sequential randomised-controlled trial with early go/no-go decisions based on total sample size in Simon's two-stage single-arm trial.

The authors used data observed in a phase II trial conducted by Digumarti *et al.* as an example of values seen in phase II clinical trials (68). From this, the development plans were simulated using parameters P_0=0.15, P_1=0.3, α=0.0025 and power=0.74.

To assess the impact on performance of phase II development plans given the opinions of the clinicians, probabilities were computed such that each of the development plans were optimal after placing reference priors on P_1. Beta distributions were used as reference priors to represent either the opinion of a sceptical or enthusiastic clinician to reflect beliefs in the likely values of P_1. They defined a sceptic as one who believed there was a 10% chance that $P_1 > P_0$, and an enthusiast as one who believed was a 10% chance that $P_1 < P_0$.

The authors found that generally, group-sequential style of randomised-controlled trials were preferred in a phase II setting. Under the first optimality criteria, single-arm trials were preferred by both sceptics and enthusiasts. Under all other optimality criteria, randomised-controlled trials were more likely to be preferred by an enthusiast. Unlike other methodological studies, Grayling *et al.* allows more than one trial to be completed in a single phase before continuing to the next. Additionally, the four different optimality criteria provide thorough critiques of each development plan in the context of sample size and power.

Limitations of Grayling *et al.* are the small range of values explored for P_1 and P_0, they only considered situations where $P_1 > P_0$ and there was lack of consideration for historical control error. Without this, it is difficult to truly assess the performance of the development plan under all circumstances, or indeed, how well a development plan can also identify a non-effective treatment.

2.3.4.2 Hunsberger et al. – A Comparison of Phase II Study Strategies

In 2009, Hunsberger *et al.* investigated the most efficient phase II-phase III development plan to determine drug efficacy through simulations (60).

Firstly, the authors described four phase II-phase III development plans that they considered:

1. A phase II single-arm trial followed by a phase III randomised-controlled trial.
2. A phase II randomised-controlled trial followed by a phase III randomised-controlled trial.
3. An integrated phase II-phase III randomised-controlled trial with a futility interim analysis with an intermediate endpoint.

4. A stand-alone phase III randomised-controlled trial with a futility interim analysis.

Development plans 1,2 and 3 used Progression Free Survival (PFS) as phase II endpoints, and all development plans used Overall Survival (OS) as phase III endpoints. Parameters used average estimates from pancreatic cancer trials to simulate results.

For development plans 1 and 2, only promising phase II trials would lead to phase III trials. Phase II trials used sample size calculations with α=0.1 and power=0.9. Additional parameters for phase II sample size consisted of experimental PFS (PFS_1)=2-4 months, and control PFS (PFS_0)=3 months and expected Hazard Ratio (HR)=1.5. For subsequent phase III trials, sample size was powered using α=0.05 and power=0.9. Additional parameters for phase III sample size consisted of control overall survival (OS_0)= 6 months, experimental overall survival (OS_1)=7.8 months, expected HR=1.3. Trials were assumed to recruit 15 patients per month, with a minimum of six months follow up.

For development plans 3 and 4, development plans would only continue past the interim analysis if there was sufficient evidence that the treatment was effective above a pre-specified threshold. Development plan 3 was planned to maintain 81% power throughout. While development plan 3 used PFS and OS as endpoints, development plan 4 only considered OS throughout.

All development plans were assessed under the null and alternative hypotheses. For development plans 2, 3 and 4, correspondence between phase II and phase III treatment effect was assessed through correlation of phase II PFS outcomes, and phase III OS outcomes. Distribution for the phase III control overall survival, OS_0, was assumed to be exponential, with median months m_0. Phase III treatment effect was defined as Δ_0 such that phase III experimental overall survival, OS_1, had an exponential distribution with $m_0\Delta_0$. Provisional phase II outcomes, such as experimental progression-free survival, PFS_1, used exponential distributions with median values of $m\rho m_0$. Further, control progression-free survival in phase II, PFS_0, used $\Delta_0 m\rho m_0$. PFS times therefore did not have exponential distributions, but when

OS and PFS were different, the correlation would be small and PFS would approximate the exponential distribution.

The main performance measures of the development plans were development plan sample size and time taken to complete the development plan.

The authors found that when historical control was overestimated in a phase II single-arm trial, the probability of correctly concluding in favour of the treatment reduced from an anticipated 81% to 51%, or in some situations even 9%. It was also found that integrated phase II-phase III trials produced the most consistent results, and most often used the least number of patients and time throughout the development plan to reach similar conclusions.

It should be noted that Hunsberger *et al.* is the only one identified in this narrative synthesis that considered all five elements highlighted in the introduction. It completed very thorough research on phase II design and impact on phase III, by not only considering traditional designs, but using a holistic approach that considers full development plans.

However, the limitations of Hunsberger *et al.* include only exploring one set of P_1, P_0, treatment effect and historical control error. Additionally, the main outcome Hunsberger *et al.* investigated was the total number of patients and time taken throughout a development plan, where a more thorough investigation on the probability of making correct decisions on a drug is arguably more useful. Additionally, it only considered correlation between phase II and phase III treatment effects for development plans 2, 3 and 4. Finally, it considered the use of PFS endpoints at phase II, when binary outcomes are more often used. By not considering phase II binary response rates, it ignored another major challenge that exists in phase II-phase III cancer clinical trials, which is to what degree phase II response rates can predict phase III survival (69-71).

Hunsberger *et al.* is significant and introduces many nuanced themes in assessing phase II design such as differing phase II and phase III endpoints and correspondence between phase II and phase III treatment effects. However, the realism of the endpoints chosen limits the application. Additionally, as it only explored parameters using one anecdotal trial, it is difficult to generalise results. However, ideas introduced by the authors can be used as a strong foundation in combination with the Taylor *et al.* paper for further research.

2.3.4.3 Maitland et al. – Analysis of the Yield of Phase II Combination Therapy Trials in Medical Oncology

In 2010 Maitland *et al.* explored the hypothesis that single-arm trials lead to a low proportion of positive phase III trials (66). As such, real phase II-phase III trials were identified through a database search to explore trends between phase II and associated phase III trials.

Initially, Maitland *et al.* identified phase II “combination chemotherapy” trials through MedLine that were published between 2001 and 2002. Authors identified 363 phase II trials, from which they identified 10 subsequent positive phase III randomised-controlled trials. The paper broke down phase II trial characteristics associated with positive phase III randomised-controlled trials, which included phase II design, disease area, conclusive outcome, and quality of methodological reporting.

In terms of phase II design, randomised-controlled trials were more likely to draw negative conclusions than single-arm trials to a statistically significant degree (p=0.004). This could suggest that randomised-controlled trials are more efficient at identifying futile treatment than single-arm trials, as their concurrent control arms are less affected by biases. However, this may not be a fair statement, as only 22 phase II randomised-controlled trials were identified compared to 341 phase II single-arm trials. With such a large discrepancy in the number of randomised-controlled trials vs single-arm trials found, it is difficult to draw any conclusions.

The strength of Maitland *et al.* is that it avoids publication bias: through its database search it first identifies phase II trials, and then their associated phase III trials. This allows both negative and positive phase II trials to be identified, differing to the method Monzon *et al.* used to find published trials.

Although the authors seemed to identify phase II and phase III trials through a robust database search, very few findings were found between phase II design and phase III outcomes. Furthermore, the trials identified were limited to “combination chemotherapy” trials published from 2001-2002. The paper could have contributed more to this field by looking at other trial characteristics such as sample size, treatment effects and by extending their initial database scope. It would have also been useful to know the trends between phase II design and negative phase III trials.

While Maitland *et al.* is promising with its robust database search to identify phase II and phase III trials, too little information is analysed on phase II design to contribute significantly to this field.

2.3.4.4 Monzon et al. - Correlation of single arm versus randomised phase 2 oncology trial characteristics with phase 3 outcome

In 2015, Monzon *et al.* investigated published phase II and phase III trials to analyse associations between phase II design and phase III success (54).

The paper completed a database search in high impact factor journals to identify 189 phase III trials. From these, 336 associated phase II trials were identified. Trial characteristics from both phase II and phase III trials were collected, and correlations between phase II study features and phase III outcomes were examined using chi-square or Fisher’s exact tests.

Monzon *et al.* found that there was no difference in association between positive phase III outcome and phase II design. In fact, the only statistically significant finding was association between positive phase III outcome and positive phase II outcome, independent of phase II design. These findings seemingly contradict all papers identified for the narrative synthesis thus far. The authors also seem to agree with Taylor *et al.*: that phase II design choice has negligible, if any, impact on phase III outcome.

Monzon *et al.* is the only paper identified in this narrative synthesis that was solely based on quantitative data from real-life clinical trials. With data extracted from multiple phase II and phase III trials, it is the only paper identified which can attempt to summarise typical information seen.

Despite the results, the methodology can be improved. For example, identifying phase III trials only through high-impact factor journals leads to publication bias in trials chosen for analysis. Additionally, only phase II trials which led to phase III trials were considered for analysis when it would be equally important to know the typical characteristics and designs used in negative phase II trials. Lastly, the authors did not adjust for multiple testing, so the significant finding between positive phase III outcome and positive phase II outcome should be treated with caution. Knowing more design characteristics, such as sample size or α and power level would have also been useful. With a few adjustments to the methodology, it could contribute further useful information to the field.

2.3.4.5 Moroz et al. - Comparison of anticipated and actual control group outcomes in randomised trials in paediatric oncology provides evidence that historically controlled studies are biased in the favour of novel treatment

In 2015, Moroz *et al.* similarly explored impact of design choice by using published information from existing clinical trials (61).

A total of 48 randomised-controlled trials were identified in the database search in the field of paediatric oncology. Authors recreated the identified randomised-controlled trials as single-arm trials by using the anticipated proportion of responders in the control arm, P_0, used in the sample size calculations as a historical control. The conclusions of the randomised-controlled trials and recreated single-arms were compared.

Moroz *et al.* found that if the randomised-controlled trials had run as single-arms, treatment effect would have been overestimated by an average of 3.8%-points, with some exceeding estimates by over 10%-points. Estimates made in randomised-controlled trials and single-arms for the same outcomes were significantly different.

Moroz *et al.* improves on the research Monzon *et al.* conducted in terms of methodology and rigour. Randomised-controlled trials chosen were not limited by impact factor from journals, therefore publication bias would have had a lesser impact. Additionally, choosing the same P_0 for sample size calculation and historical control is common practice, so it would be reasonable to conclude that the recreated single-arms are comparable to the randomised-controlled trials.

However, the results are not without limitations. It does not differentiate between phase II or phase III randomised-controlled trials used for analysis. Phase III trials tend to be more thoroughly researched, therefore the P_0 used as a sample size calculation is more likely to be reflective of true P_0, than a P_0 chosen for a phase II sample size calculation. Additionally, it is rare for a single-arm trial to be conducted in a phase III setting. The paper would be vastly improved if the authors distinguished between the results of randomised-controlled trials chosen from phase II or phase III trials. Additionally, all randomised-controlled trials selected were in the context of paediatric oncology, for which the disease is more heterogeneous among children (61). Therefore, the results may generate more variable results than oncology in general, limiting the generalisability of the results from this study.

Though Moroz *et al.* presents vital research, without distinguishing between the phase II and phase III randomised-controlled trials included for analysis, application is limited.

2.3.4.6 Pond et al. – Quantitative evaluation of single-arm versus randomized phase two cancer clinical trials

Pond *et al.* used Taylor *et al.* as an example study and aimed to improve on the methodology (56, 62). In particular, Pond *et al.* used ranges of values that were more reflective of common practice and considered the null hypothesis.

Similar to Taylor *et al.*, Pond *et al.* assessed performance on the ability of each design to correctly conclude either in favour of effective treatment, or against ineffective treatment. Values explored for the proportion of responders in the control arm, P_0, ranged from 0.05-0.75 in increments of 0.05. This was compared against three levels of proportion of responders in the experimental arm, P_1, being equal to P_0, P_0+0.15 and P_0+0.2. Pond *et al.* extended on Taylor *et al.*'s work by allowing historical control error to be equal to ±10%-points from true P_0. Methods were also adapted such that inter-institution variability for P_0, historical control, and treatment effect were sampled from binomial distributions instead of uniform distributions. Values of variability simulated for the binomial distributions were 20, 50, 100 and 200, where 20 and 200 represented high and low variation respectively. Lastly, Pond *et al.* updated the decision-thresholds to conclude in favour of new treatment so that they were based on commonly used single-arm and randomised-controlled trial designs defined in Simon *et al.* (72) and Jung *et al.* (73). Sample sizes were based on three different combinations of α and power levels commonly used in Simon and

Jung phase II trials (72, 73). If a positive phase II trial was simulated, it was considered that a subsequent phase III would be conducted. The primary outcome was proportion of phase III trials that would have been conducted testing truly effective treatment.

The paper found that in the presence of uncertainty and variability such as inter-institution variability and historical control error, single-arm phase II trials were less likely to lead to subsequent phase III trials with effective treatment compared to phase II randomised-controlled trials. This contrasts the conclusions made by Taylor *et al.* and suggests that design choice at phase II clearly had an impact on the success of phase III trials.

Accounting for historical control error within a range of standard agent response rates, considering the null hypothesis and using a more expansive range of values for P_0 are the main advantages to Pond *et al.* compared to Taylor *et al.* It highlights that the choice of phase II design impacts phase III trials. However, the assessment on impact on phase III trials was quite simplistic, as it assumed that an active drug studied in phase II that concluded in favour of the treatment would translate into a positive phase III trial. This could have been improved by considering correspondence between phase II and phase III treatment effect, and even differing phase II-phase III endpoints.

Although the authors did include a wider range of parameters typically seen in real life, P_1 was still limited to a maximum of +20%-points of the control response rate, and historical control error was limited to ±5%-points or ±10%-points of the control response rate. Further exploration of these parameters would have provided a more comprehensive view of impact on phase II designs on phase III trials.

2.3.4.7 Sambucini et al. - Comparison of single-arm vs. randomised phase two clinical trial: Bayesian approach

In 2015, Sambucini *et al.* used Bayesian analysis to compare the performance of randomised-controlled trials and single-arm trials while using a prior to represent the opinions of clinicians (63).

The authors assessed performance by comparing the ability of each design to make correct conclusions about treatment. This was done by comparing results of probability density functions for proportions of control and experimental response rate (P_0, P_1), specifically when $P_0 > P_1$ and $P_0 < P_1$ for each design. Sample sizes simulated for each design were 30, 60, 90, 100 and 150. Historical control error values explored were P_0+5%-points, P_0+10%-points or P_0+15%-points. Clinician's opinions were represented by reference priors placed on P_1.

Overall, the results showed that when there are large treatment effects, single-arm trials were preferred even when historical control error was present. The results also found that the more sceptical a clinician was regarding the accuracy of the historical control, the stronger the preference for a randomised-controlled trial.

One way Sambucini *et al.* contributes new information to the field is by thoroughly exploring impact of a clinician's opinion in the presence of historical control error. A clinician's opinion is likely to be known prior to a trial being conducted, and the chance of historical control error is always present. Therefore, knowing the full impact of these two parameters could help future investigators choose a design which is most likely to lead to the correct conclusion.

Although the consideration of historical control error is a strength in Sambucini *et al.*, the authors only considered historical control error in one direction, when historical control was larger than true P_0. It is also likely a historical control could be underestimated, therefore the full extent of possible impact from historical control error is incomplete. Additionally, the maximum level of P_1 explored was P_0+0.15. Although in practice it is rare to see large treatment effects, especially in cancer settings, biomarker subgroups can display large treatment effect sizes. Biomarker subgroup analysis has increasing interest as it leads to larger differences in subpopulations. If these subgroups did have larger treatment effects, the results from this paper could not be applied in selecting the best phase II trial design. Therefore, improvements could be made to the Sambucini *et al.* paper by considering a wider range of values, and by extending the analysis to consider impact to phase III success rates.

2.3.4.8 Sharma et al. – Resampling phase III data to assess phase II trial designs and endpoints.

In 2012, Sharma *et al.* investigated the ability of phase II designs to detect positive or negative phase III results (65). In the same vein as Tang *et al.*, phase II trials were simulated by sampling patients from existing phase III trials. In this case, patients were resampled from a positive and negative phase III trial in metastatic colon cancer.

Sharma *et al.* achieved this by using 770 patients from the positive phase III TARGET trial, and 259 patients from the negative AE941 trial as a sample pool (74, 75). All patients in the pool were resampled 5000 times to simulate a phase II single-arm two-stage trial and a phase II randomised-controlled trial. Both phase II trial designs were simulated using an α=0.1. Single-arm trials were simulated with 37 patients, and randomised-controlled trials were simulated with 40-70 patients. The only single-arm trial endpoint considered was response rate, and four types were considered for randomised-controlled trials; response rate, mean tumour sizes, progression-free survival (PFS) at 90 days and overall PFS .

The authors observed a higher true positive response rate in randomised-controlled trials than single-arm trials (96.5% vs 55%). Interestingly, the authors also reported lower false-positive rates in single-arm trials compared to randomised-controlled trials (0.9% vs 25%). These results lead to an interesting question: if the choice of design leads to a higher rate of false positives or false negatives, which one is preferred?

These results provide heightened levels of external validity in assessing both phase II trial designs by using real trial data. Through this, problems in the sensitivity for single-arms and the specificity in randomised-controlled trials to determine if there was a treatment effect could be identified. These results may provide useful guidance to investigators. Depending on the research question of a phase II trial, the difference in the sensitivity/specificity levels may be a factor in choosing a design. However, the results of the simulations are limited to the two phase III trials chosen, so may not reflect typical sensitivity and specificity levels seen in all phase II randomised-controlled trial and single-arm trials. The authors also acknowledged that the results may be skewed as the AE941 trial may not be truly negative, as marginal drug activity was found. A more comprehensive study should repeat simulations for more pairs of positive and negative phase III trials.

2.3.4.9 Tang et al. – Comparison of Error Rates in Single-Arm Versus Randomized Phase II Cancer Clinical Trials

In 2010 Tang *et al.* published a paper as an extension of Taylor *et al.* (64). The authors analysed performance of single-arm trials and randomised-controlled trials in two ways; through real data simulations and simulations based on statistical models.

The real data simulations were based on patient data from two phase III trials to assess type I and II error rates in single-arm phase II designs. The phase III N9741 trial demonstrated improved survival for patients who received experimental treatment, FOLOFOX, compared with a control arm (76). A total of 672 patients were used from the experimental arm for the real data simulations. To obtain a historical control for the simulations, the phase III trial conducted by Saltz *et al.* was used which had the same control treatment as the N9741 trial (77). 1000 trials were simulated with 50 patients in each, sampled from the experimental arm from the N9741 trial.

The statistical model simulations assessed the ability of each design to conclude correctly in the presence of historical control, patient temporal drift and patient selection effect variabilities. Historical control success rates were based on a beta distribution with predefined parameters: control success rate, θ_0 (either 20% or 50%), and a 90% confidence interval, ω. The number of studies used to specify the historical control was either four or eight. It was assumed patient temporal drift and selection effects followed a normal distribution with a mean of 0.05, and variance of 0.01. α and power levels of these simulated trials were 0.1 and 0.8 respectively. Levels of treatment effect, δ, explored were 0%, 5%, 10%, 15% and 20%. Two-sided one-sample z-tests were used to assess simulated single-arm trials, and two-sample z-tests were used to assess simulated randomised-controlled trials.

One of the interesting findings was that if the phase III N9741 trial had been run as a phase II single-arm trial, it would not have concluded in favour of the effective treatment. Through statistical simulations, the authors also found that as variability increases in the phase II trials, single-arm trials have higher error rates than randomised-controlled trials. These corroborate with findings in Pond *et al.* that choice of phase II design can impact phase III conclusions.

The authors further highlight ways in which phase II trial design choice can impact conclusions made on new treatments. By using real clinical trial data to mimic a phase II single-arm trial, they demonstrate that choice of phase II design can impact phase III conclusions, thereby stifling progress of a new drug through the pharmaceutical pipeline. For their statistical model simulations, the consideration of patient temporal drift and patient selection effect affirms the necessity of choosing a historical control arm carefully to minimise error. Similar to Taylor *et al.* and Pond *et al.*, the applicability of the statistical model simulations are restricted as few values of sample size, control success rate and treatment effect were explored. The paper could be improved by considering a wider range of these parameter values, and further investigating subsequent impact on phase III trials.

2.4 Discussion

The initial findings in Taylor *et al.* suggested that phase II design choice had a negligible impact on phase III design. Through this narrative synthesis I have identified nine additional papers which have since made further assessment on the performance of phase II trial designs. A variety of methods were used, including statistical simulations, Bayesian methods, retrospectively assessing published clinical trial data and using existing clinical trial data as sources to recreate trials.

Five elements were identified that quantitative studies should consider to robustly assess the impact of phase II trial design on subsequent phase III trials. These considerations were: impact on subsequent phase III trials, both alternative and null hypotheses, historical control error, differing phase II and phase III endpoints and correspondence between phase II and phase III treatment effects.

Of the 10 papers selected for the narrative synthesis, five considered impact on phase III trials, eight papers considered the null hypothesis, six papers accounted for historical control error, one considered different phase II-phase III endpoints and two considered correspondence between phase II and phase III treatment effect. Only one paper, Hunsberger *et al.*, considered all five elements.

Despite the initial findings of Taylor *et al.*, eight of the nine other papers concluded that phase II design choice influenced conclusions made about treatment. Many papers suggested that when there was uncertainty surrounding the historical control error, a randomised-controlled trial was preferred. It should be noted this was not always the case, as one paper found that when historical control was overestimated and treatment effect was large, a single-arm trial was preferred. One paper also found that when sample sizes were small, single-arm trials out-performed randomised-controlled trials.

Even though uncertainty in historical control can negatively impact the performance of a phase II single-arm trial, quantifying the value of uncertainty before it makes a negative impact is difficult. This is partly due to the widely varying analysis all papers used, and that some papers failed to investigate the full ranges of P_0, P_1 and historical control error itself. Therefore, further work is required to establish this. Several questions about subsequent impact on phase III trials also remain unanswered. Although five papers did consider the impact on phase III trials, many of them only considered whether a phase III trial was likely to be conducted, not the result of the subsequent phase III trial itself. Moreover, only two considered the correspondence between phase II and phase III treatment effects, and only one considered different endpoints that would be used at each phase. However, the differing endpoints that Hunsberger *et al.* considered were not reflective of real-life practice. With these research gaps still present, it is difficult to quantify in which clinical trial scenarios each design performs optimally.

To date, there is no consensus on how to choose the most appropriate phase II design to maximise the performance of phase III trials. In four of the papers, performance was defined as the ability to correctly detect effective treatment, and correctly detect ineffective treatment (56, 62-64). Addressing these issues could provide robust research to contribute to formal recommendations on phase II design choice in the future, and improve the success rates of cancer development plans.

My review of the literature between 2000-2018, has highlighted many issues in the area of phase II design choice, and had identified clear gaps that need to be addressed. Therefore, my future work aims to complete a study that considers the following five key elements, updated given the results from the narrative synthesis:

1. Impact on phase III trial conclusions
2. Both alternative and null hypothesis
3. Historical control error
4. Phase II binary response rate and phase III time-to-event survival endpoints
5. Imperfect correspondence between phase II and phase III treatment effects.

The work presented throughout the rest of the thesis aims to carry out research which considers all five of these points simultaneously to provide well-rounded evidence. This research can support investigators in choosing the most appropriate phase II trial design to maximise development plan success rates.

3. Simulation Studies with Identical Binary Endpoints at Phase II and III

3.1 Introduction

The narrative synthesis identified five key elements which should be considered in a quantitative methodological study to robustly assess the impact of phase II design choice on a drug development plan. These key elements include the consideration of:

1. Phase III trial conclusions
2. Both alternative and null hypotheses
3. Historical control error
4. Phase II binary response rate and phase III time-to-event survival endpoints
5. Imperfect correspondence between phase II and phase III treatment effects

In this chapter, I will conduct four simulation studies which account for the first three key elements. This will be done by assessing performance of two different phase II-phase III development plans; those that have a single-arm phase II trial, and those that have a randomised phase II trial. For simplicity, I will only simulate each design with a single-stage and one experimental arm. Additionally, I will not distinguish between simulating a phase IIA or a phase IIB trial. Therefore, the phase II trials simulated can be considered to represent the overall phase II process. However, the decision to continue to a simulated phase III trial can be seen to resemble the decision at the end of a phase IIB trial.

It should be noted that I am simulating all the clinical trial data, with all observations available and no censoring, thereby assuming no missing data.

These four simulation studies will use binary endpoints at phase II and III.

Study 1 will address the first two key elements by looking at the performance of the whole phase II-phase III development plan under both the alternative and null hypothesis. There will be no historical control error, which means hypothesised values for P_0 (true proportion of responders in the control arm) and P_1 (true proportion of responders in the experimental arm) used in phase II sample size calculations are reflective of the truth.

Study 2 will build on Study 1 by introducing the third key element, historical control error. Here, hypothesised values of P_0 used for phase II sample size estimation will not reflect the truth. For phase III trials that follow a randomised phase II trial, it will be assumed that phase II estimates for P_0 and P_1 are reflective of the truth, and therefore phase III sample size will be appropriately powered given the true values of P_0 and P_1. However, for phase III trials that follow a single-arm phase II trial, the historical control error will be carried over into the phase III sample size calculation.

Study 3 will build on Study 2 by allowing stopping rules at the end of phase II and will allow phase III trials to base sample size on observed phase II estimates of P_0 and P_1.

Study 4 will expand on Study 3 by extending the range of historical control error and will assess different levels of P_0 and P_1.

Details of Study 1, Study 2, Study 3 and Study 4 can be seen in Table 4.

Dimensions of a study	Study 1	Study 2	Study 3	Study 4
Key development	1) Introduces phase II impact on phase III trials 2) Assesses alternative and null hypothesis	Introduces historical control error	Allows phase III sample size to be based on phase II estimates	1) Extends range of historical control error 2) Assesses different values of P_0 and P_1.
Phase II stopping rules	No	No	Yes	Yes
Values used for phase II sample size calculation	Hypothesised P_0=0.1, Hypothesised P_1=0.2, 0.3, 0.4	Hypothesised P_0=0.1, Hypothesised P_1=0.2, 0.3, 0.4	Hypothesised P_0=0.1, Hypothesised P_1=0.2, 0.3, 0.4	Hypothesised P_0=0.4, Hypothesised P_1=0.5, 0.6, 0.7
Historical control error	No	P_0=$(HP_0$±1%-point) P_0=$(HP_0$±2%-points) P_0=$(HP_0$±5%-points)	P_0=$(HP_0$ ±1%-point) P_0=$(HP_0$±2%-points) and P_0=$(HP_0$±5%-points)	P_0=$(HP_0$ ±1%-point), P_0=$(HP_0$±2%-points) P_0=$(HP_0$±5%-points of true P_0) P_0=$(HP_0$±15%-points)

Values used for phase III sample size calculation	After single-arm phase II: hypothesised P_0 and true P_1 After randomised phase II: true P_0 and true P_1	After single-arm phase II: hypothesised P_0 and true P_1 After randomised phase II: true P_0 and true P_1	Phase II $\widehat{P_0}$ and phase II $\widehat{P}_1$	Phase II $\widehat{P_0}$ and phase II $\widehat{P}_1$

Table 4 - Elements to be incorporated in each of the four studies conducted

3.2 Methods

Two development plans were created for the simulation studies:

- A single-arm phase II trial followed by a phase III randomised-controlled trial.
- A randomised phase II trial followed by a phase III randomised-controlled trial.

Randomised phase III trials were chosen to reflect standard practice (13).

The two main measures of performance were:

- Proportion of times a development plan correctly concludes in favour of effective treatment out of 10000 repetitions.
- Proportion of times a development plan does not reject the null hypothesis for ineffective treatment out of 10000 repetitions.

Both these measures assessed each development plan's ability to make correct decisions on treatment, a performance measure commonly used in many papers identified in the narrative synthesis (56, 60, 62, 63, 65). Performance measures also helped address the second key element; impact of phase II design choice under both the alternative and null hypothesis.

The randomised and single-arm trial designs tested different hypotheses. Randomised trial designs compared proportions of response rate between two groups. However, single-arm phase II trials compared observed proportion of responders in the experimental arm with a hypothetical proportion of responders in the control arm i.e., the historical control. Error surrounding this historical control addressed the third key element, impact of phase II design choice in the presence of historical control error. This will mimic situations where chosen historical control does not reflect the true population control response rate.

There were two broad approaches that assessed sample size; calculating sample size based on fixed α and $1-\beta$ (i.e. power) or using a fixed sample size. In ideal circumstances, sample size would always be calculated based on fixed α and $1-\beta$ which would vary depending on the trial phase. However, in practice, phase II investigators can have limited resources of patients in a phase II setting, with papers reporting a median sample size in phase II trials as 100 participants (78). To reflect this, sample size in the simulations were assessed using:

1. Ideal circumstances where investigators had access to unlimited phase II participants, for which sample size was calculated using fixed α=0.15 and $1 - \beta$=0.8.
2. Restricted circumstances where investigators were limited to fixed sample sizes of 50, 74 and 100.

In the first approach, a different sample size calculation was conducted for each phase II design: the A'hern method for single-arm trials and likelihood ratio-test for randomised trials. There will also be differences in phase III sample size calculations between the two development plans. In Study 1 and Study 2, the hypothesised values of P_0 and P_1 will remain the same across phase II and phase III trials. In study 3 and 4, phase III sample size calculations are based on observed estimates from the phase II trials.

For the second approach, sample sizes of 50, 74 and 100 were chosen to represent the instances when investigators had limitations when recruiting patients. Therefore, a fixed sample size of 50 represented a very low access to participants, 74 represented low access to patients, and 100 represented the median size of phase II trials.

The rest of the methods section provides a list of definitions and notation that will be used throughout the simulations. To follow, the simulation plan is described using the ADEMP(S) structure by Morris *et al.* (79).

3.2.1 Definitions

Table 5 describes the simulation parameters. They are divided into three categories: 1) unknown truth, 2) pre-phase II trial, which is the information investigators use to design the phase II trial and 3) post-phase II trial which is the information estimated from the trial results.

	Terminology	Description
unknown truth	P_0	True proportion of responders in the control arm
	P_1	True proportion of responders in the experimental arm
	Treatment effect	$(P_1 - P_0)$ i.e., Difference between P_0 and P_1
	Negative historical control error	$HP_0 < P_0$ i.e., when hypothesised control response proportion is less than the truth for sample size calculation
	Positive historical control error	$HP_0 > P_0$ i.e., when hypothesised control response proportion is more than the truth for sample size calculation
Pre-phase II trial	HP_0	Hypothesised control response proportion investigators expect to see in the trial (used in sample size calculations).
	HP_1	Hypothesised experimental response proportion investigators expect to see in the trial (used in sample size calculations).
	α	Designed probability threshold of accepting the alternative hypothesis when there is truly no effect between treatments
	$1 - \beta$	Designed probability threshold of rejecting the null hypothesis when there is truly an effect between treatments (if applicable)
post-phase II trial	$\widehat{P_0}$	Observed control arm response proportion estimate from the trial (only available when randomised-controlled trial is used at phase II)
	$\widehat{P_1}$	Observed experimental response proportion estimate from the trial

Table 5 – Simulation definitions

3.2.2 Simulation Plan for Study 1,2,3 and 4 – ADEMP(S)

The following section will be split up to describe Aims, Data generating mechanisms, Estimands, Methods and Performance measures. An additional sub-section will describe S for Simulation sample size & Monte Carlo error.

3.2.2.1 Aims

The aim of these simulation studies is to investigate impact of phase II trial design performance on a development plan under three of the five key elements identified in the narrative synthesis. These included the evaluation of:

1. Phase III trial conclusions
2. Both alternative and null hypotheses
3. Historical control error

Simulations with consideration of these three key elements will be investigated in four studies.

3.2.2.2 Data Generating Mechanisms

A *data generating mechanism* refers to the set of values that were used to generate a dataset. For each of the four studies, parameters were adjusted to create new datasets. Table 6 describes all the parameters and their ranges of values which were considered in the four studies. The parameters were divided into two categories: states of nature i.e., parameters that cannot be controlled for when designing a phase II trial, and parameters under the investigators' control.

It should be noted that hypothesised values of P_0 (HP_0) were anchored throughout the simulations, with historical error represented by differing the value of true P_0 around the anchored HP_0. This was chosen to reflect real-life practice where an investigator chooses an HP_0 before knowing the truth. Before completing the trial, the P_0 is unknown and could theoretically be any value. Therefore, the simulations represent the various scenarios from an investigators point of view where the unknown P_0 could be different to their chosen HP_0, to varying degrees.

	Parameters used in simulation	Values simulated			
		Study 1	Study 2	Study 3	Study 4
States of nature i.e. parameters that cannot be controlled for	P_0	$P_0 = HP_0$			P_0 = (HP_0 −15%-points)
					P_0 = (HP_0 −10%-points)
			P_0 = (HP_0 −5%-points)	P_0 = (HP_0 −5%-points)	P_0 = (HP_0 −5%-points)
			P_0 = (HP_0 −2%-points)	P_0 = (HP_0 −2%-points)	P_0 = (HP_0 −2%-points)
			P_0 = (HP_0 −1%-points)	P_0 = (HP_0 −1%-points)	P_0 = (HP_0 −1%-points)
			$P_0 = HP_0$	$P_0 = HP_0$	$P_0 = HP_0$
			P_0 = (HP_0 +1%-points)	P_0 = (HP_0 +1%-points)	P_0 = (HP_0 +1%-points)
			P_0 = (HP_0 +2%-points)	P_0 = (HP_0 +2%-points)	P_0 = (HP_0 +2%-points)
			P_0 = (HP_0 +5%-points)	P_0 = (HP_0 +5%-points)	P_0 = (HP_0 +5%-points)
					P_0 = (HP_0 +10%-points)
					P_0 = (HP_0 +15%-points)

	P_1	$P_1 = HP_1$	$P_1 = HP_1$	$P_1 = HP_1$	$P_1 = HP_1$
Parameters under the investigators' control	HP_0	0.1	0.1	0.1	0.4
	HP_1	0.2, 0.3, 0.4	0.2, 0.3, 0.4	0.2, 0.3, 0.4	0.5, 0.6, 0.7
	α (one-sided)	phase II = 0.15	phase II = 0.15	phase II = 0.15	phase II = 0.15
		phase III = 0.025	phase III = 0.025	phase III = 0.025	phase III = 0.025
	$1 - \beta$	phase II = 0.8	phase II = 0.8	phase II = 0.8	phase II = 0.8
		phase III = 0.9	phase III = 0.9	phase III = 0.9	phase III = 0.9

Table 6 – *Data generating mechanisms* used for preliminary simulations

There are many moving parts to the simulations, therefore, I have broken up the section for *data generating mechanisms* into three parts. First, I will describe true and hypothesised values for P_0 and P_1 for each of the four studies. Second, I will describe the sample size calculation assumptions for each of the four studies. Lastly, I will describe how trial arms will be simulated for each of the four studies.

True and hypothesised values of P_0 and P_1

Study 1

For Study 1, hypothesised P_0 (HP_0) and hypothesised P_1 (HP_1) used for sample size calculation were reflective of the truth i.e., HP_0=P_0 and HP_1= P_1. HP_0 was anchored at 0.1 throughout, and HP_1 ranged from 0.2, 0.3 and 0.4. A level of HP_0=0.1 was considered as low response rates of between 10%-15% are commonly found in early-phase cancer clinical trials (80). The three levels of HP_1 were chosen to demonstrate varying levels of treatment effect, similar to those explored in papers identified in the narrative synthesis (62, 64, 67). No historical control error was considered in this study for the sake of simplicity.

Phase II sample size was assessed in four ways; fixed at either 50, 74, 100 patients or calculated with a one-sided α=0.15 and $1-\beta$=0.8. As previously mentioned, the fixed sample sizes were chosen to reflect situations where phase II investigators have limited access to participants, with median phase II sample size found to be 100 (78). Many α and $1-\beta$ levels are used in practice to calculate phase II sample size, however, a one-sided 0.15 α threshold with a 0.8 $1-\beta$ power threshold was chosen. This was to represent the most extreme difference in phase II sample size calculations to contrast the phase III sample size calculations, but within the range that is still seen in common practice (81). As phase III trials are designed to be larger, they have less restrictions on sample size. Therefore, all phase III sample sizes were calculated with one-sided α =0.025 and $1-\beta$=0.9, chosen to reflect standard practice (82). For all phase II and phase III sample size calculations, true P_0 and true P_1 values were used.

For development plans with a phase II single-arm, a *data generating mechanism* existed for each value of HP_1= P_1 i.e., 0.2, 0.3 and 0.4, and was repeated under each of the four variations of phase II sample size. Simulations were repeated under the alternative hypothesis and null hypothesis, formally written as $H_0: P_1 \leq P_0$ and $H_1: P_1 > P_0$. This was repeated for development plans with a randomised phase II trial. This means Study 1 had three values of HP_1 and four levels of phase II sample size, which were simulated under the alternative and null hypothesis for both development plans. This gave a total of 48 (3 x 4 x 2 x 2) *data generating mechanisms* for Study 1.

Study 2

Study 2 used the same 48 data generating used in Study 1, but additionally included historical control error where hypothesised proportion of control responders was not equal to the truth, i.e., hypothesised P_0≠ true P_0. Seven levels of historical control error were assessed: where true P_0 = hypothesised P_0−5%-points, true P_0 = hypothesised P_0−2%-points, true P_0 = hypothesised P_0−1%-point, no error, true P_0 = hypothesised P_0+1%-point, true P_0 = hypothesised P_0+2%-points and true P_0 = hypothesised P_0+5%-points. A maximum of 5%-points of historical control error was initially chosen to examine the impact on trial performance when only a small amount of error was present. In development plans with a single-arm phase II trial, hypothesised P_0 and true P_1 were used to calculate the sample size in the subsequent phase III trial. In development plans with a randomised phase II trial, true P_0 and true P_1 was used to calculate subsequent phase III sample size. With seven levels of historical control error, there were 336 [48 x 7] *data generating mechanisms* for Study 2.

Study 3

The only difference between Study 2 and Study 3 was how phase III sample size was calculated. Study 2 used hypothesised control response rate (HP_0), and true experimental response rate (P_1) for phase III trials that followed a single-arm phase II trial. Conversely, true response rates for both the control and experimental arms (P_0, P_1) were used for phase III trials that follow a randomised phase II trial. In Study 3, both development plans used phase II estimates of control and experimental response rates ($\widehat{P_0}$, $\widehat{P_1}$) to calculate phase III sample size. This was the only change from study 2 to study 3, therefore 336 *data generating mechanisms* were present for Study 3 also.

Study 4

Study 4 used a different value of HP_0 than other studies, HP_0=0.4. It also used different values for HP_1: 0.5, 0.6, 0.7. Different levels of HP_0 and HP_1 were chosen to expand upon the range used in Studies 1-3, similar to Pond *et al.* from the narrative synthesis (62). This provided the additional benefit of being able to assess if relative treatment effect size could impact development plans in a different way to Studies 1-3, while absolute treatment effect size remains the same. Historical control error range was extended where true P_0 = hypothesised P_0−15%-points, true P_0 = hypothesised P_0−10%-points, true P_0 = hypothesised P_0+10%-points and P_0 = hypothesised P_0+15%-points, similar to the historical control errors investigated by Sambucini *et al.* seen in the narrative synthesis (63). Therefore, 11 levels of P_0 were assessed under four levels of phase II sample size, three levels of HP_1, under both the alternative and null hypothesis and across both development plans. This gave a total of 528 [11 x 4 x 3 x 2 x 2] *data generating mechanisms* in Study 4.

Sample size calculation assumptions

All four studies compared the two phase II-phase III development plans, those with a single-arm phase II trial, and those with a randomised phase II trial. Each of these phase II trial designs used different assumptions to calculate sample size.

A'hern was the method chosen to calculate single-arm phase II trials, as this method is commonly used to calculate single-stage single-arm trials in practice (83). A'hern aims to test basic level of efficacy in a single-arm trial, when efficacy is a proportion such as proportion of responders to treatment. Hypothesised P_1 is considered minimum proportion of efficacy needed to conclude in favour of treatment, and hypothesised P_0 represents the level at which the treatment is ineffective (84).

The A'hern method provides investigators with a cutoff for minimum number of responders to conclude in favour of treatment, while ensuring that P_1> P_0, for a given sample size for prespecified levels of P_0, P_1,α and β.

The A'hern method is used as an improvement on Fleming's single-stage sample size procedure which uses the normal approximation to the binomial distribution which is incorrect for small sample sizes. These inconsistencies can lead to the possibility of rejecting the null hypothesis when the value of P_0 is within the confidence interval for the estimate of P_1. Therefore, A'hern uses the exact binomial distribution.

For example, with a P_0=0.05 and P_1=0.2, and one-sided α =0.05 and (1- β)=0.9 the Fleming sample size procedure determines that the sample size should be 34, with a minimum number of responders as 5. However, the observed α for this sample size is actually 3%, and observed (1-β) is only 0.84. A'hern sample size gives 38 for the sample design criteria, with a minimum number of responders as 5. These values reach the desired threshold of observed one-sided α =0.05 and power of 90% (85). Details of the Fleming sample size calculation is given in the appendix.

Therefore, A'hern uses the Fleming sample size procedure, but tests various cutoff points for the observed α and β levels with the exact binomial distribution for ranges that fall from 0.8-4 times the Fleming sample size.

Cut-off points are determined such that:

$$\text{cut-off} = N_{PII}\; x \left\{HP_0 + \left[\left(\frac{z_\alpha}{(z_\alpha + z_{1-\beta})}\right) x\; (HP_1 - HP_0)\right]\right\}$$

Where z_α and $z_{1-\beta}$ are the standardized normal deviates of the targeted α and $(1-\beta)$ levels.

The only exception to this is when the Fleming sample size is <30, for which the A'hern method tests every possible cutoff point.

The smallest sample size tested that satisfies the original design criteria, with observed one-sided α =0.05 and power=0.9, is determined to be the A'hern sample size and corresponding minimum number of responders. These cutoffs make it clear that the confidence interval of P_1 will exclude P_0 by at least a small amount.

For development plans with a randomised phase II trial, the phase II sample size was calculated with a one-sided likelihood ratio test with one-sided α=0.15 and $1-\beta$=0.8 (81). A likelihood ratio test was chosen as this is a common method to assess proportions in randomised clinical trials (86). All sample sizes for phase III trials were calculated a one-sided likelihood ratio test with one-sided α=0.025 and $1-\beta$=0.9 (82).

Simulating trial arms

In all four studies, binary endpoints were used in all phase II and phase III trials. In cancer clinical trials it is common practice for binary outcomes to represent a patient's response to treatment; for partial or complete response to be an outcome of interest, and this is sometimes the primary interest at phase II (87). A binary response rate is commonly used in phase II cancer trial settings, where definition of response is defined prior to the trial, such as shrinkage of tumour of >10mm (1, 2). Therefore, in the simulation studies, Bernoulli random draws were used to simulate patients' responses for each treatment group in a trial. True response proportion defined "success rate", where success was when a patient responded to treatment (88).

For single-arm phase II trials, one experimental arm was simulated. For a given *data generating mechanism,* the associated P_1 represented probability of success, i.e. true response proportion of the experimental treatment. The number of Bernoulli random draws was equal to the required sample size obtained using the A'hern method.

For randomised phase II trials, both experimental and control arms were simulated. For a given *data generating mechanism*, Bernouilli random draws were used in both arms, where sample size was calculated using a one-sided likelihood ratio test for associated hypothesised P_1 and hypothesised P_0.

Phase III randomised-controlled trials were simulated in different ways depending on the study. For Study 1 and Study 2, phase III randomised-controlled trials mimicked the way randomised phase II studies were generated with different α and $1-\beta$ levels for sample size calculation. For simplicity, phase III trials were generated regardless of phase II outcome, but individual phase III trials were retrospectively dismissed if the prior phase II trial did not conclude in favour of treatment. For development plans with a single-arm phase II trial, subsequent phase III randomised-controlled trials carried forward any historical control error which then impacted phase III sample size calculation used to generate trial arms. For development plans with a randomised phase II trial, subsequent phase III randomised-controlled trials used values of that reflected the truth for phase III sample size calculation used to generate trial arms, i.e. true P_0 and true P_1.

For Study 3 and Study 4, phase III samples sizes that followed a randomised phase II trial were calculated using phase II estimates of $\widehat{P_0}$ and $\widehat{P_1}$ to simulate trial arms. In development plans with a single-arm phase II trial, subsequent phase III sample size was calculated using the hypothesised P_0 (with any historical control error) and the phase II estimate of $\widehat{P_1}$ to simulate trial arms.

3.2.2.3 Estimands/ target

The target in my simulations was hypothesis rejection at phase II and phase III, to help evaluate the performance of development plans, specifically, the proportion of development plans that either correctly concluded in favour of treatment, or proportion that correctly failed to reject the null hypothesis.

The specific null hypothesis was that the response rate in the experimental arm is less than or equal to the response rate in the control arm across phase II and phase III. As previously stated, the specific null hypothesis is $H_0: P_1 \leq P_0$ across a phase II-phase III development plan, and the specific alternative hypothesis is $H_1: P_1 > P_0$.

3.2.2.4 Methods of analysis

One method of analysis was performed for each development plan which is presented in Table 7.

Development plan	Phase II analysis	Phase III analysis
Phase II single-arm x Phase III RCT	Comparison of the number of responders in the simulated trial with minimum number of responders needed to conclude in favour of treatment according to the A'hern method. This calculation determined the minimum number of responders needed where the 85% confidence interval did not include hypothesised P_0.	Comparison of the two observed proportions using a likelihood ratio test that assumes $Y_i \sim Binomial$ with one-sided p-value < 0.025.
Phase II RCT x Phase III RCT	Comparison of the two observed proportions using a likelihood ratio test that assumes $Y_i \sim Binomial$ with one-sided p-value < 0.15.	

Table 7 – Method of analysis for the two development plans

3.2.2.5 Performance measures

There were two performance measures, one under the alternative hypothesis and one under the null hypothesis. Under the alternative hypothesis ($P_1 > P_0$), the performance of the development plans was measured by the observed proportion of simulations that correctly reject the null hypothesis. This was labelled "observed true positive". Under the null hypothesis ($P_1 \leq P_0$), the performance of the development plans was measured by the observed proportion of times they did not reject the null hypothesis. This was labelled "observed true negative".

In Study 1 and Study 2, a negative result from a development plan was recorded when a phase II trial failed to reject the null hypothesis, even if the subsequent phase III trial concluded in favour of treatment. If a phase II trial within a development plan concluded in favour of the experimental treatment, but the subsequent phase III trial failed to reject the null hypothesis, it will also be recorded as a negative result. A positive result of a development plan was only recorded when both phase II and phase III concluded in favour of the treatment.

In Study 3 and Study 4 decision rules existed, and phase III trials only followed from phase II trials when they rejected the null hypothesis. Therefore, a negative result from a development plan was collected when a development plan stopped after phase II or failed to reject the null hypothesis at the end of phase III. A positive result of a development plan was recorded when both the phase II and phase III trials concluded in favour of treatment.

3.2.2.6 Simulation Sample Size & Monte Carlo error

A simulation sample size was calculated to find the minimum number of repetitions needed to reduce the Monte Carlo error. It was calculated as 10000 repetitions for each study.

This was calculated using the formula provided by Morris *et al.* (79). It was determined that the Monte Carlo standard error for the chosen performance measures would be highest when proportion of development plans that conclude correctly is 50%. Therefore, to ensure Monte Carlo standard error remained below the desired level of 0.5%, the minimum number of simulations was calculated as:

$$n_{sim} = \left\{ \frac{E(\text{proportion of development plans concluding correctly}) \; x \; [1 - E(\text{proportion of development plans concluding correctly})]}{\left(\text{Monte Carlo SE}_{\text{req}}\right)^2} \right\}$$

$$n_{sim} = \left\{ \frac{0.5 \; x \; (1 - 0.5)}{0.005^2} \right\} = 10000$$

Where $Monte\ Carlo\ SE_{req}$ is equal to 0.5%

n_{sim} is the number of simulation repetitions

A Monte Carlo standard error of 0.5% was chosen as an acceptable and sufficiently small error rate. It is important to note that the formula assumed a binomial distribution for end of development plan conclusions after phase III.

Example code of Study 3 can be found in the appendix. In this instance, a development plan with a randomised phase II trial is simulated under the alternate hypothesis, with a HP_0=0.1, and HP_1=0.2 with 1%-point of historical control error, i.e. P_0 =0.11. All analysis was conducted using Stata version 16.0, except A'hern sample size calculations which were conducted using PASS v.20.0.2.

3.3 Results

The results are divided into four main sections for each study. Within each study section, there are two subsections of results. The first subsection compares the performance of the two development plans when the phase II sample size is fixed at 50, 74 and 100 participants. The second subsection compares the performance of the two development plans when the phase II trials used a sample size calculation with fixed α=0.15 and $1-\beta$=0.8.

3.3.1 Study 1

3.3.1.1 fixed sample size

Figure 2 describes Study 1 when phase II sample sizes are fixed at 50, 74 and 100. In this study, hypothesised P_0=true P_0 and hypothesised P_1=true P_1. Future references to hypothesised P_0 and hypothesised P_1 will be written as HP_0 and HP_1 respectively.

Figure 2a reports the results under the alternative hypothesis, i.e. $P_1 > P_0$. Furthermore, HP_0 and HP_1 reflected the truth, such that HP_0=P_0 and HP_1=P_1. The x-axis represents values of HP_1 ranging from 0.2 to 0.4. As a reminder, HP_0 remained anchored at 0.1. The y-axis represents true positive proportion i.e., proportion of 10000 repetitions that a development plan correctly rejected the null hypothesis. The grey solid line represents the "analytical level" of true positive proportion that development plans should have achieved, with the power that would be expected if the tests were perfectly calibrated at both stages, i.e. 80% and 90% respectively

(80% x 90% = 72%). The main outcome measure was assessed on whether development plans were able to achieve this target level of true positive proportion.

All orange lines represent results of development plans with single-arm phase II trials with varying fixed phase II sample sizes. The solid line represents a phase II trial with a fixed sample size of 50, the dashed line represents a phase II trial with a fixed sample size of 74 and the dotted line represents a phase II trial with a fixed sample size of 100.

All blue lines represent results of development plans with randomised phase II trials with varying fixed phase II sample sizes. The solid, dashed, and dotted lines represent results with fixed phase II sample sizes of 50, 74 and 100 respectively.

Figure 2a shows that development plans with single-arm phase II trials had higher true positive proportions than development plans with randomised phase II trials for the cases I considered. This is particularly true when HP_1=0.2, resulting in a small treatment effect. Using fixed phase II sample size of 50 as an example, when HP_1=0.2 the true positive proportion the development plan with a randomised phase II trial achieved was ~45.5%. For those with a single-arm trial, true positive proportion was 72.3%. However, as HP_1 and sample size increased, the advantage of using a single-arm phase II trial in a development plan lessened. This can be seen when HP_1=0.4 and phase II sample size is fixed at 100. In these circumstances, both development plans achieved true positive proportions of 90%.

Figure 2b reflects the results under the null hypothesis, i.e. $P_1 \leq P_0$. Here, HP_1 was not reflective of the truth, with HP_0=P_0=P_1. The x-axis represents values of HP_1 ranging from 0.2 to 0.4. Again, HP_0=0.1. The y-axis represents true negative proportion i.e., proportion of 10000 repetitions when a development plan did not reject the null hypothesis. The grey solid line represents the “analytical level” of true positive proportion that development plans should have achieved, assuming that the power at phase II and phase III were 1-(15% x 2.5%) = 99.7%. The main outcome measure was whether development plans achieved this target level of true negative proportion.

Figure 2b shows that development plans with single-arm phase II trials performed similarly to development plans with randomised phase II trials. Both development plans performed at a high level regardless of HP_1 value where true negative proportions remained stable, ranging from 99.4-99.8%.

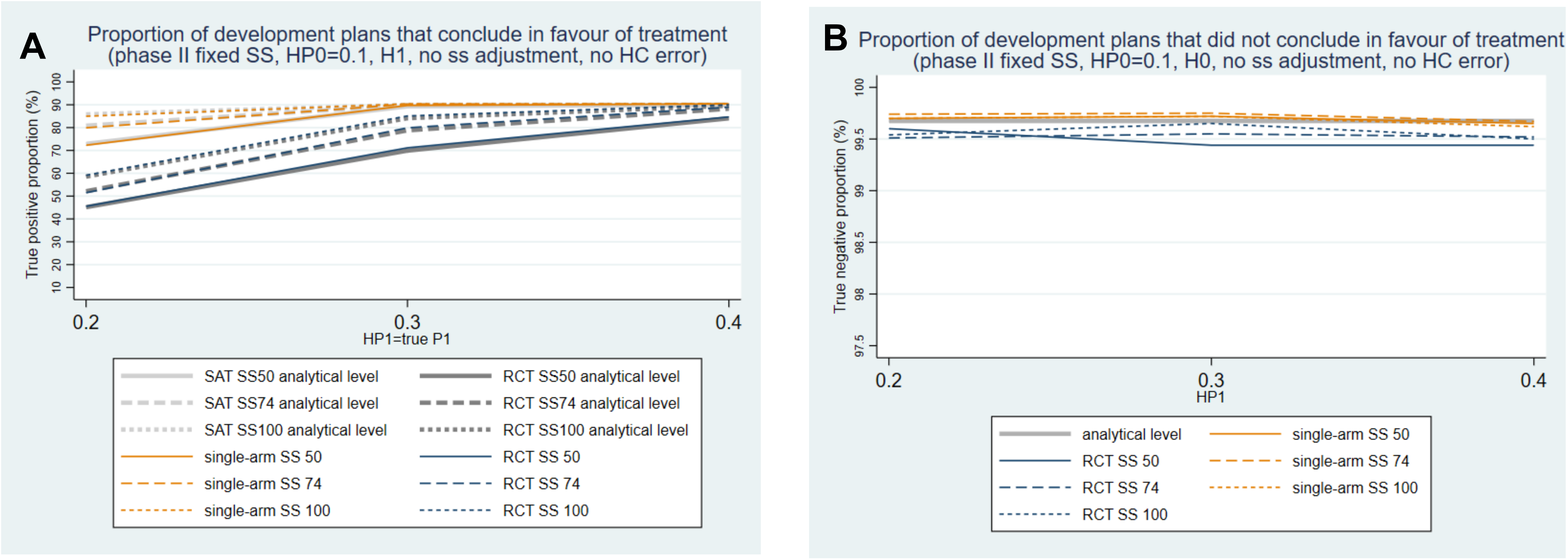


Figure 2 – Study 1 true positive proportion and true negative proportion with fixed phase II sample size and HP_1=0.2, 0.3, 0.4

Figure 2a and Figure 2b display results under the alternative and null hypothesis respectively.

3.3.1.2 fixed alpha and beta

Figure 3 presents the results of Study 1, when the phase II sample size was calculated with α=0.15 and $1-\beta$=0.8.

Similar to Figure 2, blue lines represent results for development plans with single-arm phase II trials and orange lines represent results for development plans with randomised phase II trials. The x-axis represents values of HP_1 ranging from 0.2 to 0.4. It should be noted that in all instances HP_0=0.1.

Figure 3a displays results under the alternative hypothesis, i.e., $P_1 > P_0$. Furthermore, HP_0 and HP_1 reflected the truth, such that HP_0=P_0 and HP_1=P_1. True positive proportion is displayed along the y-axis. The grey solid line represents the "analytical level" of true positive proportion that development plans aimed to achieve if tests were perfectly calibrated at both stages; 72%. Figure 3a shows that under the alternative hypothesis, there was little difference between the true positive proportion of the two development plans. When HP_1=0.2, both development plans achieved a true positive proportion of 71.7-71.5%. As HP_1 became larger, development plans with single-arm phase II trials had an increased true positive proportion compared to development plans with randomised phase II trials; 76.1% compared to 73.3%.

Figure 3b displays results under the null hypothesis i.e. $P_1 \leq P_0$. Here, HP_1 was not reflective of the truth, with HP_0=P_0=P_1. True negative proportion is along the y-axis and HP_1 is along the x-axis. The grey solid line represents the "analytical level" of true negative proportion that development plans aimed to achieve if both tests were perfectly calibrated at both stages; 99.7%.

Figure 3b shows that under the null hypothesis, true negative proportion remained stable under all simulated values of HP_1. Development plans with single-arm phase II trials achieved higher levels of true negative proportion than those with randomised phase II trials. This is because in this instance, historical control happened to be perfectly accurate in single-arm phase II trials with HP_0=true P_0=0.1. Therefore, the estimate of the active experimental arm where true P_1=0.1, was always compared against a value that reflected the truth in the control arm. However, in randomised phase II trials, estimates from two active treatment arms were compared. Although both true P_0=0.1 and true P_1=0.1, due to random chance, estimates of P_0 and P_1 would vary. In small sample sizes, it was more difficult for randomised trials to obtain reliable estimates for both P_0 and P_1, leading to increased chances that trials would erroneously detect a treatment effect. As larger anticipated treatment effects lead to smaller required sample sizes, the performance of development plans with randomised phase II trials diminished as the difference between HP_0 and HP_1 increased. For example, when HP_1 was anticipated to be 0.4, randomised phase II trials only recruited 13 participants in each arm. However, differences were marginal, as all true negative proportions were between 99.1-99.6%.

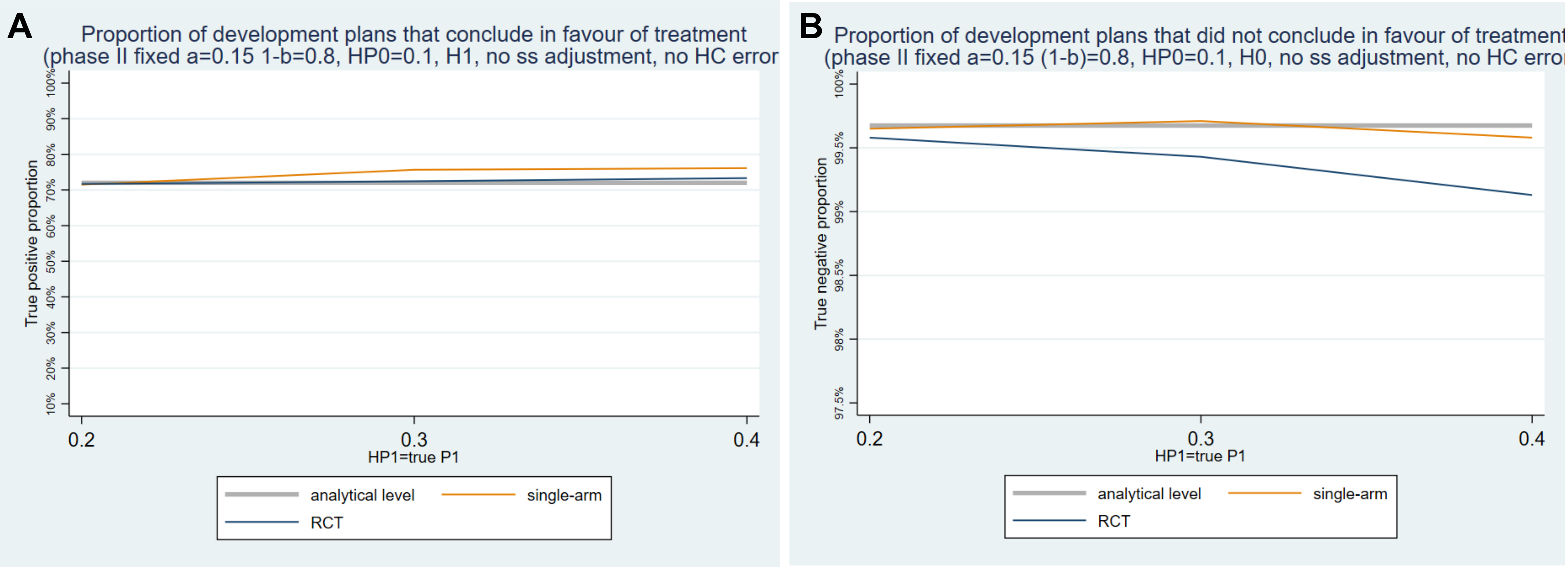


Figure 3 – Study 1 true positive proportion and true negative proportion with phase II sample size calculations using fixed α=0.15 and 1-β=0.8 with HP_1=0.2, 0.3, 0.4

Figure 3a and Figure 3b display results under the alternative and null hypothesis respectively.

3.3.2 Study 2

3.3.2.1 fixed sample size

Figure 4 depicts the results of Study 2 under the alternative hypothesis when phase II sample sizes were fixed at 50, 74 and 100.

The y-axis represents true positive proportion. The x-axis represents the degree of historical control error. The central point indicates no historical control error i.e., HP_0 = true P_0, and phase II sample size was powered appropriately. To the left of the central point, there is up to −5%-points of negative historical control error (HP_0 < true P_0). Here, HP_0 was less than the truth, and phase II sample size was smaller than required. To the right of the central point, there is up to +5%-points of positive historical control error (HP_0 > true P_0). Here, HP_0 was more than the truth, and phase II sample size was larger than required.

In all instances HP_0=0.1. Figure 4a represents results for when HP_1=0.2, Figure 4b represents results for when HP_1=0.3 and Figure 4c represents results for when HP_1=0.4. The grey solid line represents the “analytical level” of true positive proportion that development plans aimed to achieve if the tests were perfectly calibrated at both stages; 72%.

In all cases considered, development plans with single-arm phase II trials had a greater true positive proportion than those with randomised phase II trials when they used the same fixed sample size. However, the advantage of using phase II single-arm trials was reduced in the presence of positive historical control error (HP_0 > true P_0). An example can be seen in Figure 4a when phase II fixed sample size was 50. When there was no historical control error (HP_0 = true P_0), development plans with single-arm phase II trials achieved a true positive proportion of 72.3%, and those with randomised phase II trials achieved 45.5%. However, in the presence of +5%-points of positive historical control error (HP_0 > true P_0), development plans with single-arm phase II trials achieved a true positive proportion of 80.8%, and those with randomised phase II trials achieved 77.1%.

Figure 5 depicts the results of Study 2 under the null hypothesis when phase II sample sizes were fixed at 50, 74 and 100.

The y-axis represents true negative proportion, and the x-axis represents range of historical control error. In all instances HP_0=0.1, and Figure 5a, Figure 5b and Figure 5c represent results for when HP_1=0.2, 0.3 and 0.4 respectively. The grey solid line represents the “analytical level” of true negative proportion that development plans aimed to achieve if the tests were perfectly calibrated at both stages; 99.7%.

Figure 5a shows that development plans with randomised phase II trials had stable true negative proportion throughout all levels of historical control error, ranging from 98.9%- 99.7%. Development plans with single-arm phase II trials had reduced true negative proportion in the presence of negative historical control error (HP_0 < true P_0) of 98% compared to true negative proportion in the presence of positive historical control error (HP_0 > true P_0), 100%. This did not change as treatment effect increased as Figure 5b and Figure 5c show similar results.

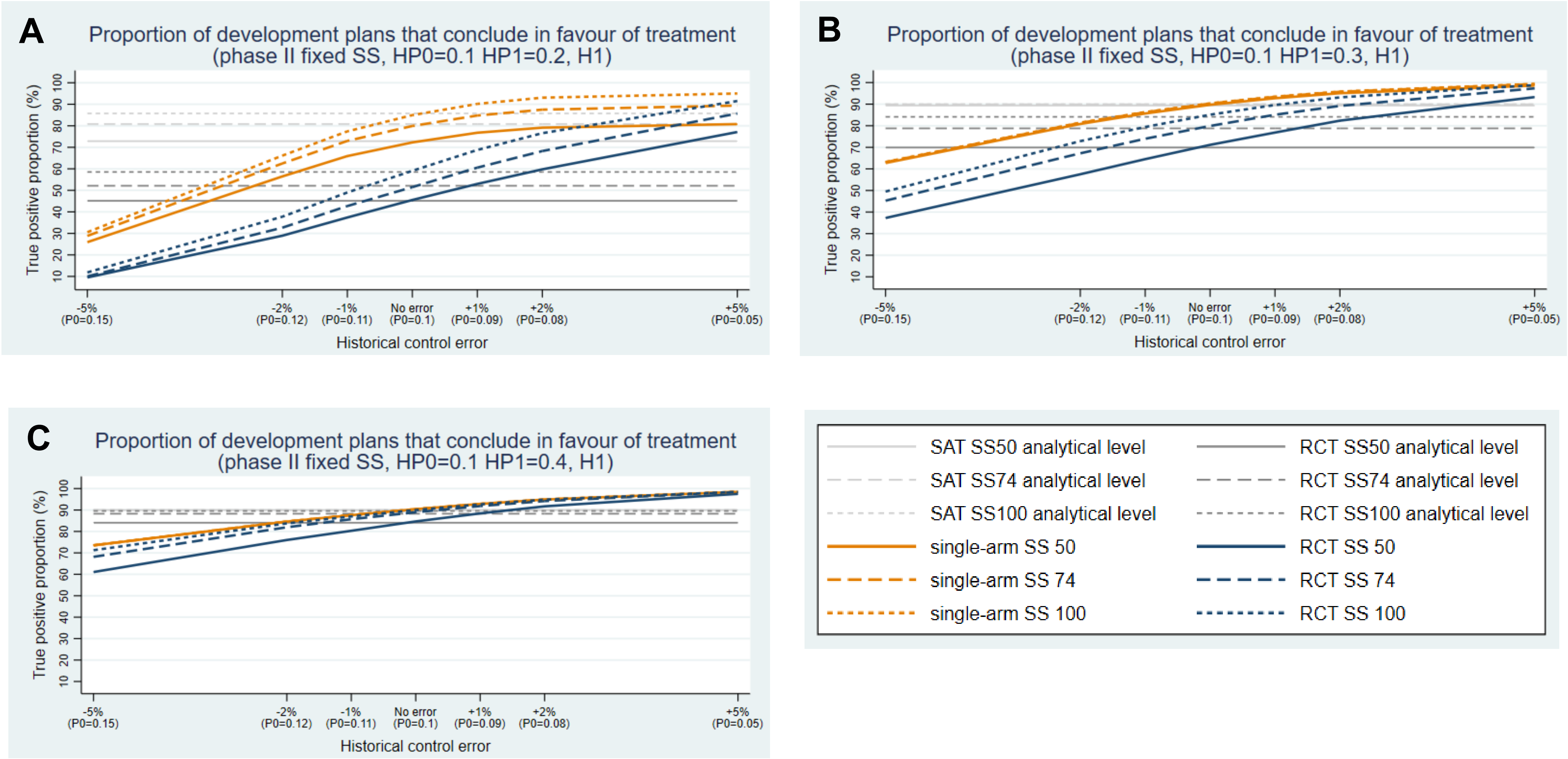


Figure 4 – Study 2 true positive proportion with phase II fixed sample size. Fig 4a, 4b and 4c display results for when HP_1=0.2, 0.3, 0.4 respectively

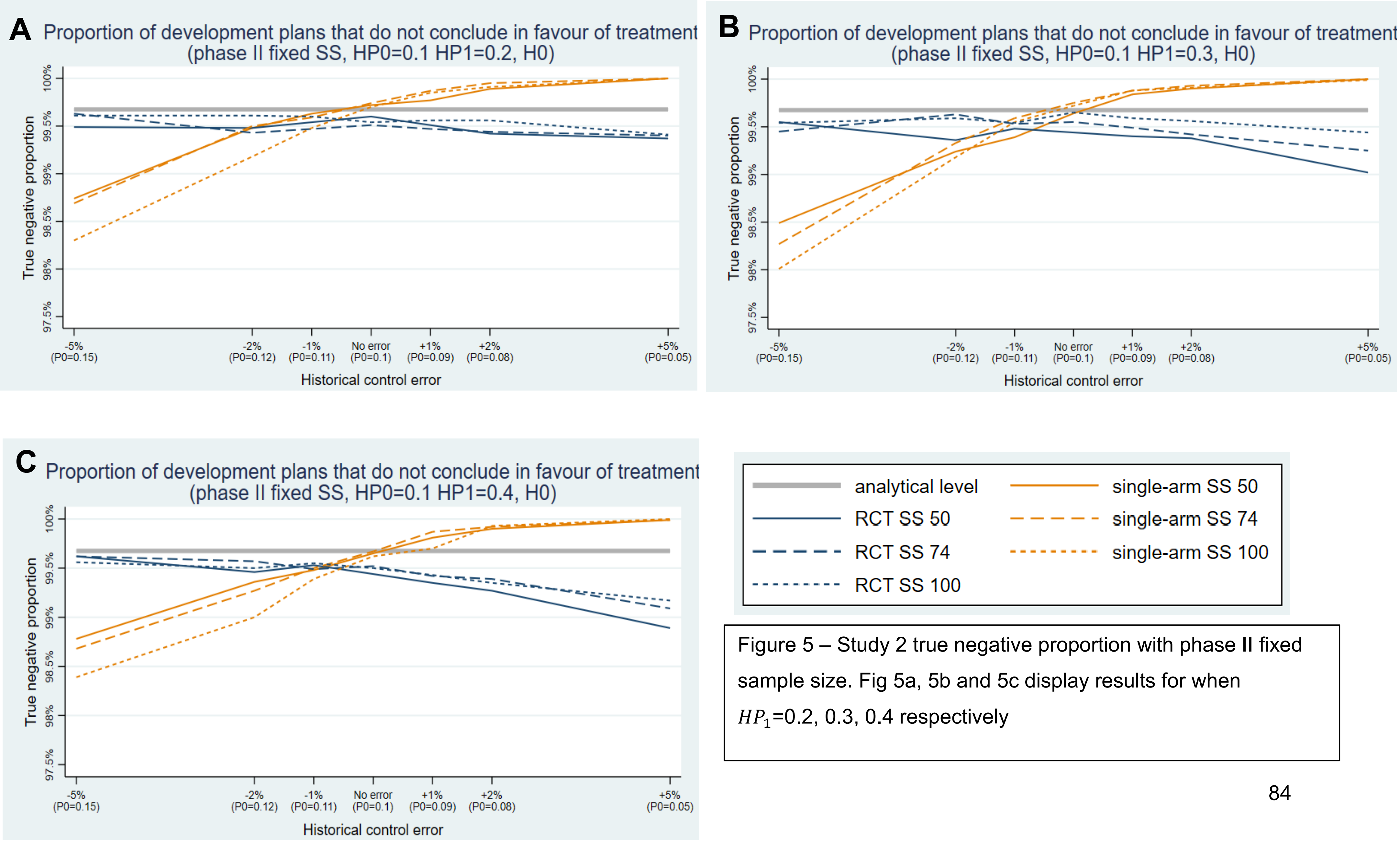


Figure 5 – Study 2 true negative proportion with phase II fixed sample size. Fig 5a, 5b and 5c display results for when HP_1=0.2, 0.3, 0.4 respectively

3.3.2.2 fixed alpha and beta

Figure 6 depicts the results of Study 2 when phase II sample size calculation used fixed α=0.15 and $1-\beta$=0.8.

In all instances, HP_0=0.1 and HP_1 varied from 0.2, 0.3 and 0.4.

Similar to Figure 5, the x-axes represent the range of historical control error. However, the y-axis represents true positive proportions under the alternative hypothesis in Figure 6a, and under the null hypothesis in Figure 6b.

Figure 6a shows that in the presence of negative historical control error (HP_0 < true P_0), development plans with single-arm phase II trials had a higher true positive proportion compared with equivalent development plans with randomised phase II trials that used the same HP_1. When HP_1=0.2, both development plans performed poorly in negative historical control error (HP_0 < true P_0) but improved as HP_1 increased. For example, when HP_1=0.2 and there was −5%-points of negative historical control error (HP_0 < true P_0), development plans with single-arm and randomised phase II trials had true positive proportions of 25.7% and 14% respectively. When HP_1=0.4 and there was −5%-points of negative historical control error (HP_0 < true P_0), development plans with single-arm phase II trials and randomised phase II trials had true positive proportions of 84.4% and 68.1% respectively.

In the presence of +2%-points or more of positive historical control error (HP_0 > true P_0), development plans with randomised phase II trials had a higher true positive proportion than development plans with single-arm phase II trials. This can be seen when HP_1=0.2. In +2%-points of positive historical control error, development plans with single-arm phase II trials and randomised phase II trials had true positive proportions of 78.3% and 88.5% respectively.

Figure 6b shows that both development plans performed in a similar way under the null hypothesis regardless of historical control error or HP_1. The proportion of times a development plan did not conclude in favour of treatment ranged from 98.6% to 100% in all cases considered.

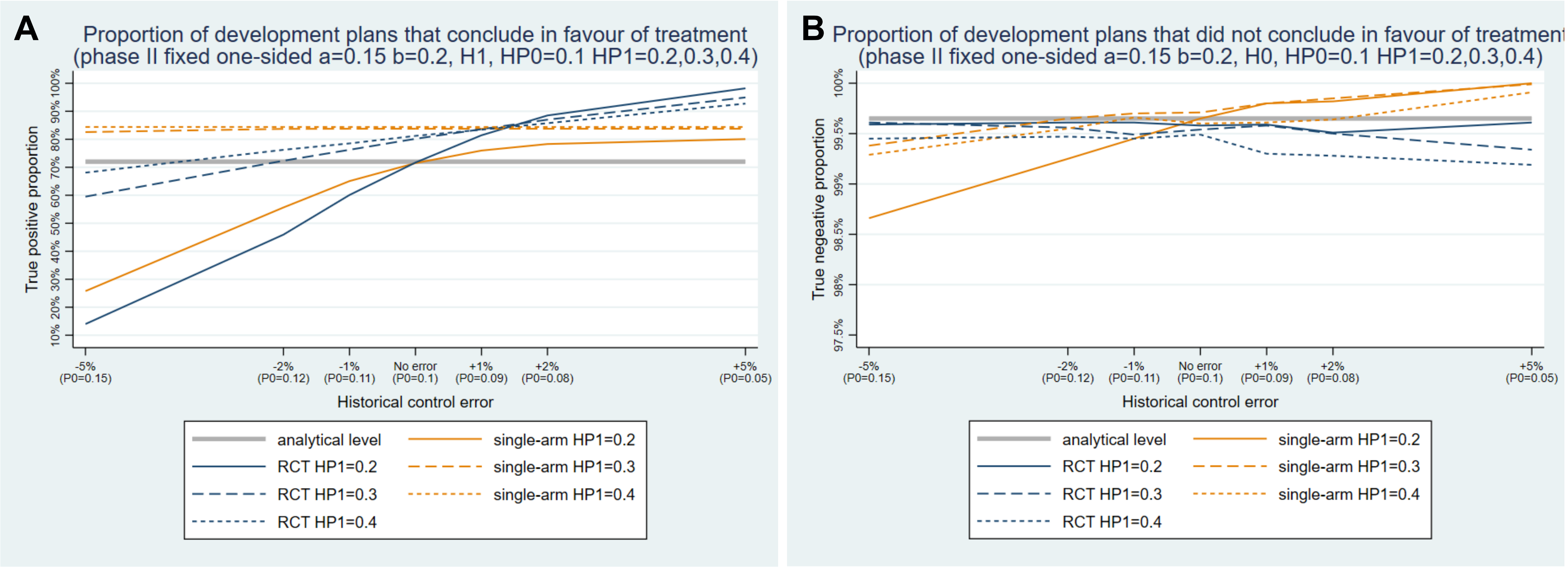


Figure 6 – Study 2 true negative proportion and true positive proportions when phase II sample size calculations use fixed α=0.15 and $1-\beta$=0.8 with HP_1=0.2, 0.3, 0.4

3.3.3 Study 3

3.3.3.1 fixed sample size

Figure 7 depicts the results of Study 3 when phase II sample sizes were fixed at 50, 74 and 100.

Figure 7a, Figure 7b and Figure 7c display results under the alternative hypothesis for when HP_0=0.1 and HP_1 are 0.2, 0.3 and 0.4 respectively. The y-axes represent true positive proportion and x-axis represents range of historical control error.

Figure 7a and Figure 7b show that when HP_1=0.2 or 0.3, development plans with single-arm phase II trials outperformed development plans with randomised phase II trials, even when randomised trials used more participants. This is evident in Figure 7a when there was no historical control error (HP_0 = true P_0) and development plans with single-arm phase II trials with a fixed sample size of 50 achieved a true positive proportion of 64.5%, while development plans with randomised phase II trials and a fixed sample size of 100 achieved a true positive proportion of 44.8%.

Figure 7c shows that development plans with randomised phase II trials outperformed those with single-arm phase II trials when HP_1=0.4 in the presence of negative historical control error (HP_0 < true P_0). For example, when phase II sample size was fixed at 100 and there was −5%-points of negative historical control error (HP_0 < true P_0), development plans with single-arm phase II trials achieved a true positive proportion of 76%, while those using randomised phase II trials achieved a true positive proportion of 82.6%. However, when there was no historical control error or positive historical control error (HP_0 > true P_0), all development plans with single-arm trials performed better than ones with randomised phase II trials. This can be seen when there was no historical control error (HP_0 = true P_0) and all development plans with single-arm phase II trials had true positive proportions that ranged from 90-90.9% and those with randomised phase II trials had true positive proportions that ranged from 79.1-87.9%.

Figure 8 represents results for Study 3 under the null hypothesis. Results of HP_0=0.1 with HP_1=0.2, 0.3 and 0.4 were identical, and therefore displayed in one graph. They are identical because the data generating mechanisms remained the same as they all used the same fixed phase II sample sizes, and under the null hypothesis the truth simulated was P_0=P_1=0.1.

The y-axis represents true negative proportion and x-axis represents range of historical control error.

Under the null hypothesis, development plans with randomised phase II trials remained stable throughout negative and positive historical control error, varying between 99.4-99.7% regardless of phase II sample size. However, development plans with single-arm phase II trials performed worse in the presence of negative historical control error (HP_0 < true P_0) and increased in positive historical control error (HP_0 > true P_0). It should be noted that absolute difference in performance between the two development plans was marginal, as all values ranged from 98.3% to 100%.

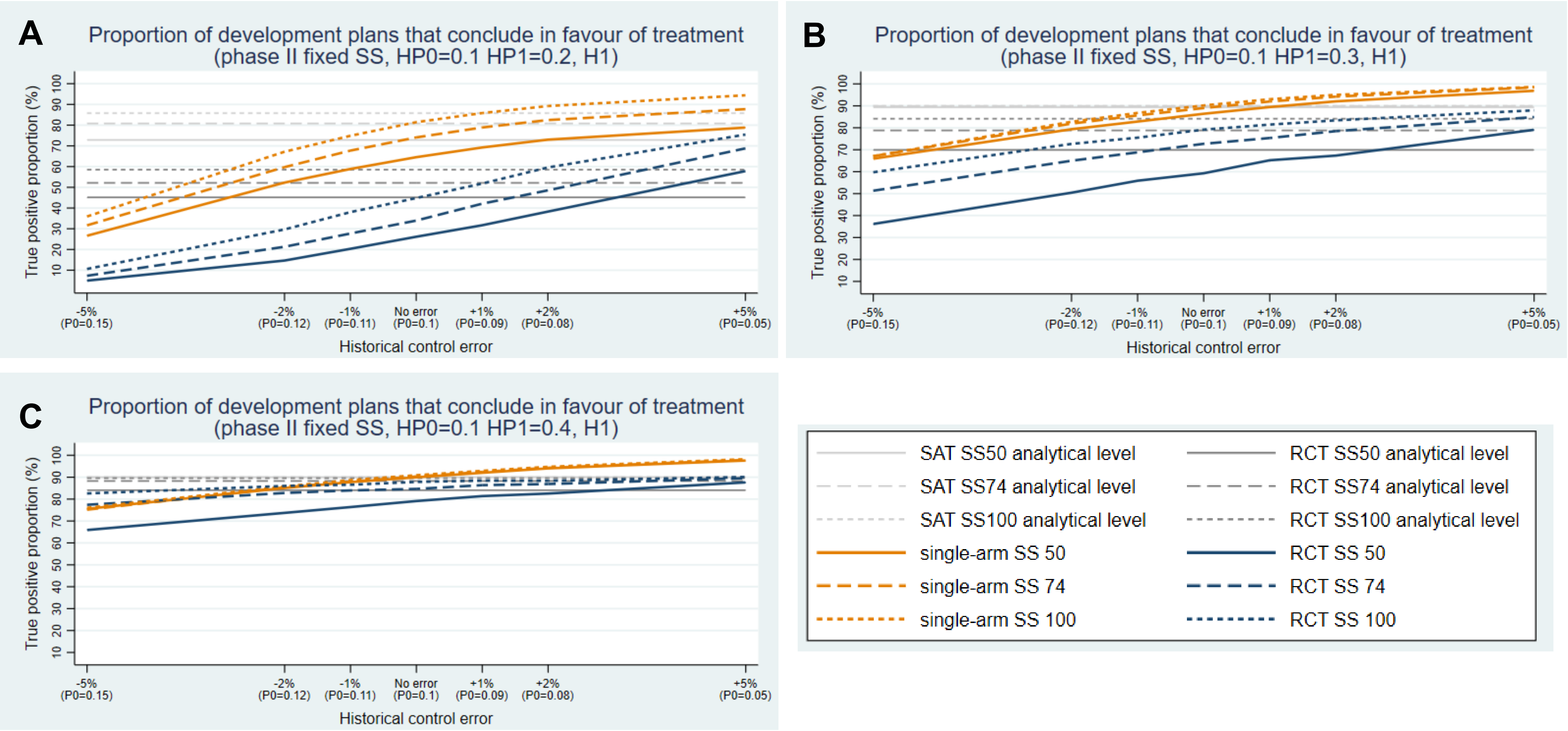


Figure 7 – Study 3 true positive proportions using phase II fixed sample size of 50, 74 and 100.

Figure 7a, Figure 7b and Figure 7c display results under the alternative hypothesis when HP_1=0.2, 0.3 and 0.4 respectively.

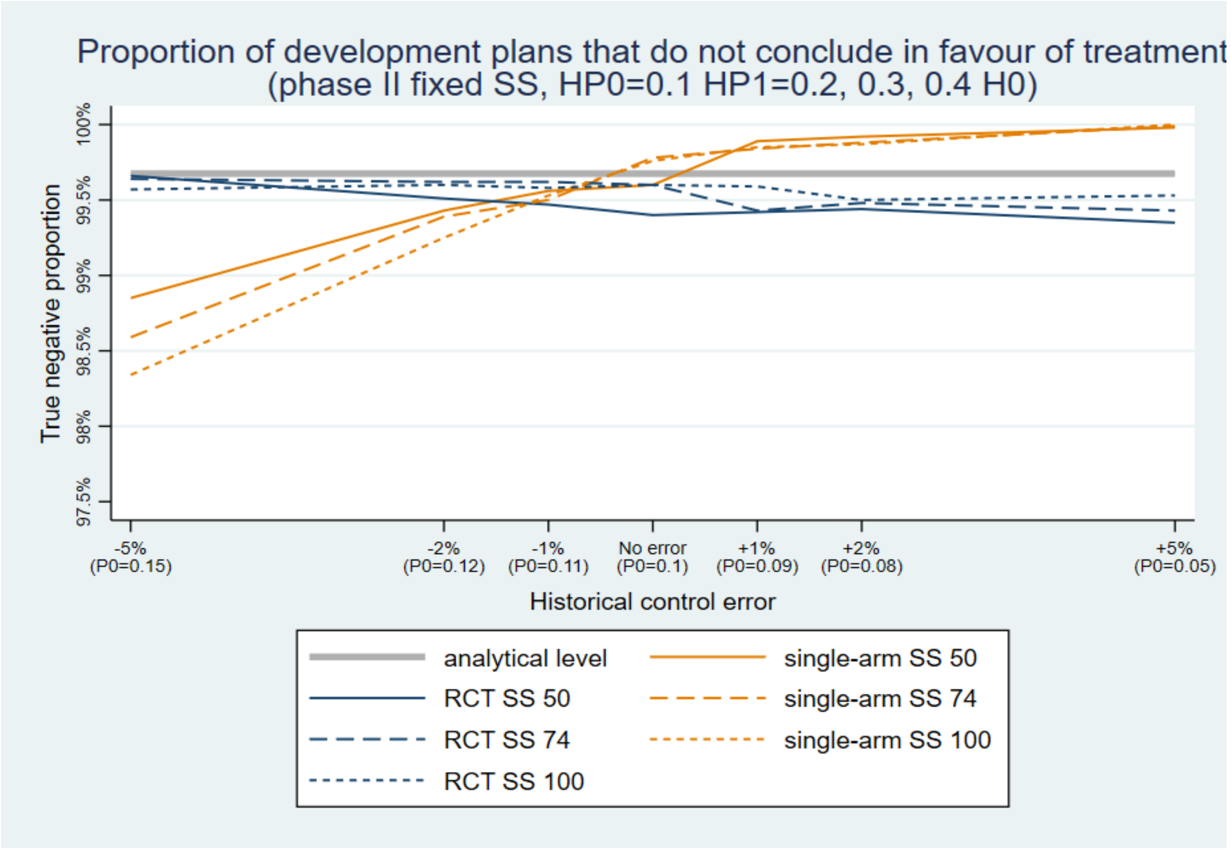


Figure 8 – Study 3 true negative proportions using phase II fixed sample size of 50, 74 and 100 for HP_1=0.2, 0.3 and 0.4 under the null hypothesis.

3.3.3.2 fixed alpha and beta

Figure 9 depicts the results of Study 3 when phase II sample size calculation used fixed α=0.15 and $1-\beta$=0.8.

Figure 9a displays results under the alternative hypothesis where the y-axis represents true positive proportion and x-axis displays range of historical control error.

Figure 9a shows that in the presence of negative historical control error (HP_0 < true P_0), development plans with single-arm phase II trials performed better than equivalent development plans with randomised phase II trials with the same value of HP_1. An example of this can be seen when HP_1=0.2. Development plans with single-arm phase II trials had a true positive proportion that was 8.9%-points more than those with randomised phase II trials (26.6% versus 17.7%). Similarly, when HP_1 =0.3 and 0.4, development plans with single-arm phase II trials had true positive proportions that were between 7-11.8%-points more than development plans with randomised phase II trials.

Figure 9b represents results under the null hypothesis where the y-axis represents true negative proportion.

Figure 9b shows more variable results than previous studies, no matter the level of HP_1 or historical control error. However, in all scenarios both development plans performed well, as true negative proportions remained above 98.7%.

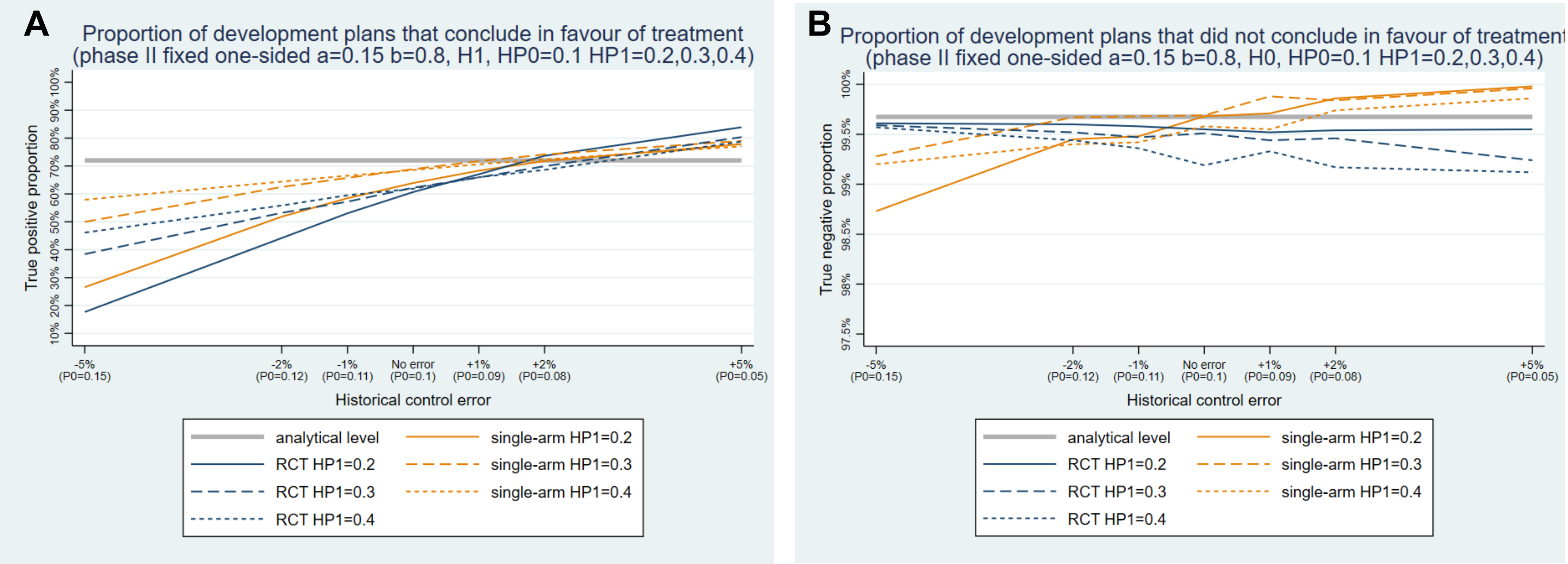


Figure 9 – Study 3 true negative proportion using phase II sample size calculations with fixed α=0.15 and $1-\beta$=0.8 when HP_0=0.1 and HP_1=0.2, 0.3, 0.4

Figure 9a and Figure 9b display results under the alternative and null hypothesis respectively.

3.3.4 Study 4

3.3.4.1 fixed sample size

Figure 10 and Figure 11 depicts the results of Study 4 when phase II sample sizes were fixed at 50, 74 and 100.

Figure 10a, Figure 10b and Figure 10c display results under the alternative hypothesis where HP_0=0.4 and HP_1 were 0.5, 0.6 and 0.7 respectively. The y-axes represent true positive proportion; and x-axes represent range of historical control error.

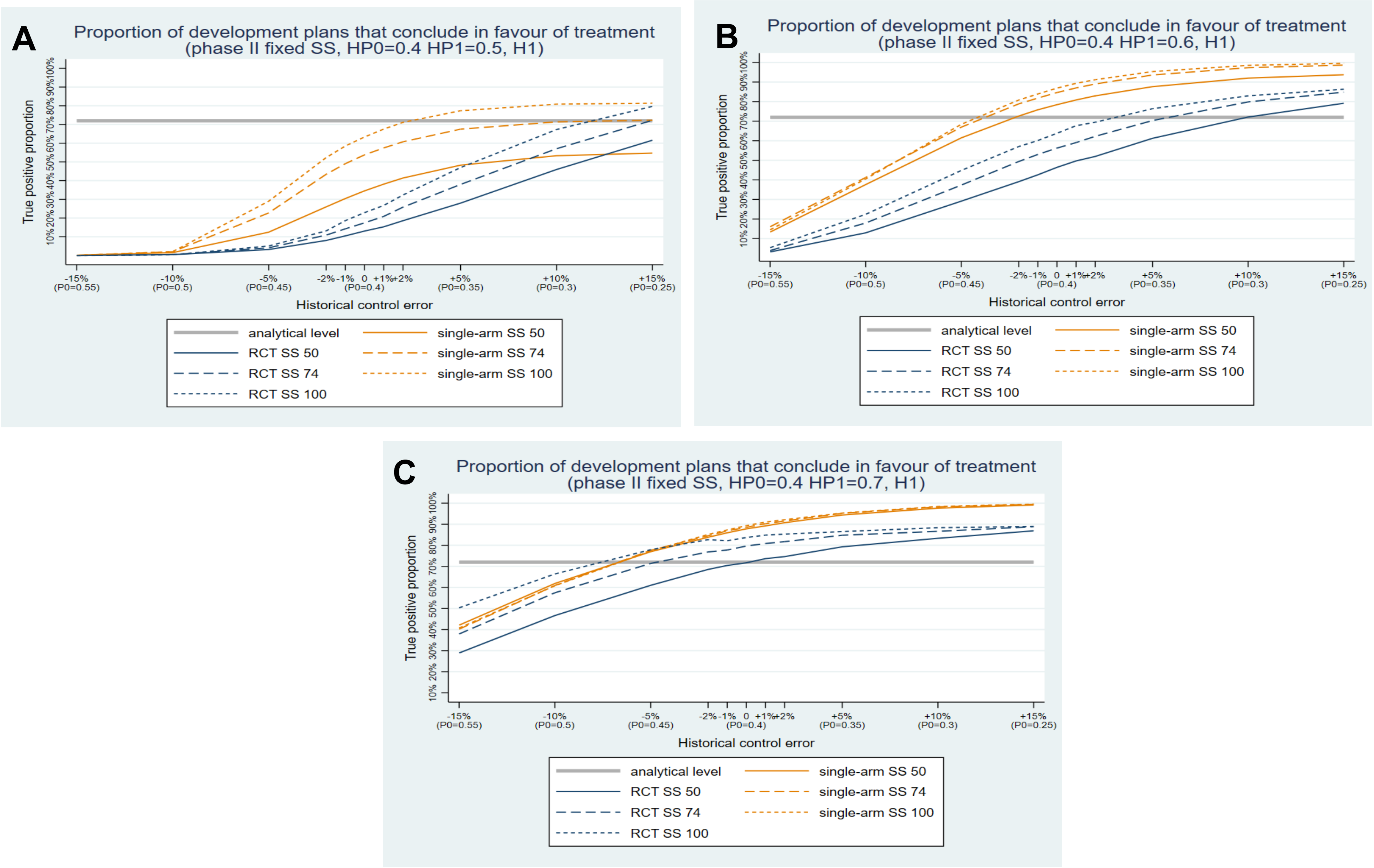
Figure 10a shows that when HP_1=0.5, development plans that used single-arm phase II trials outperformed those with randomised phase II trials when historical control error ranged from −10% to +5%-points. For example, when there was no historical control error (HP_0 = true P_0), development plans with single-arm phase II trials and sample sizes of 50 achieved true positive proportions of 34.5%. However, development plans with randomised phase II trials and sample sizes of 100 achieved true positive proportions of 22.9%. For development plans with single-arm phase II trials, performance tapered in extreme positive historical control error (HP_0 > true P_0). For example, the range of true positive proportions for these development plans when sample sizes were 50 ranged from 48.3% to 54.7% when positive historical control error was between +5% and +15%-points. However, for a development plan with a randomised phase II trial, linear improvement continued throughout. An example of this can be seen when these development plans had a sample size of 100 and true positive proportions ranged from 47% to 79.7% when positive historical control error ranged from +5% to +15%-points. These trends suggest that when phase II sample size is 50, development plans with randomised phase II trials can perform better than those with single-arm phase II trials in extreme positive historical control error (HP_0 > true P_0).

Figure 10b shows that when HP_1=0.6, development plans with single-arm phase II trials always outperformed those with randomised phase II trials. When there was no historical control error (HP_0 = true P_0), development plans with single-arm phase II trials with a sample size of 50 achieved true positive proportions of 78.4%, however, those that used randomised phase II trials with sample sizes of 100 achieved true positive proportions of 63.8%.

Figure 10c shows that when HP_1=0.7 and there was no error or positive historical control error (HP_0 ≥ true P_0), development plans with single-arm phase II trials were optimal. Take the example of when there was no historical control error (HP_0 = true P_0). Development plans with single-arm phase II trials and sample sizes of 50, and those with randomised phase II trials and sample sizes of 100, achieved true positive proportions of 87.8% and 83.7% respectively. However, when phase II sample sizes were 100, development plans with randomised phase II trials outperformed those with single-arm phase II trials in extreme negative historical control error (HP_0 < true P_0). This is evident in −15%-points of historical control error. Development plans with single-arm phase II trials and sample sizes of 50 achieved true positive proportions of 42.1%, while those with randomised phase II trials and sample sizes of 100 achieved true positive proportions of 50.3%.

Figure 11 represents results under the null hypothesis. The y-axis represents true negative proportion and x-axis displays the range of historical control error. Results for when HP_0=0.4 with HP_1=0.5, 0.6 and 0.7 were identical, and therefore displayed in one graph. They are identical because the data generating mechanisms remained the same as they all used the same fixed phase II sample sizes, and under the null hypothesis the truth simulated was P_0=P_1=0.4.

Figure 11 shows that under the null hypothesis, the performance of development plans with randomised phase II trials remained stable across all levels of historical control error and varied between 99.4-99.7%. However, development plans with single-arm phase II trials had reduced true positive proportion in the presence of negative historical control error (HP_0 < true P_0) at 97.2%. This increased to 100% in the presence of extreme positive historical control error (HP_0 > true P_0). All differences in performance between development plans were marginal as results ranged from 97.2% to 100%.



Proportion of development plans that conclude in favour of treatment (phase II fixed SS, HP0=0.4 HP1=0.5, H1)

Proportion of development plans that conclude in favour of treatment (phase II fixed SS, HP0=0.4 HP1=0.6, H1)

Proportion of development plans that conclude in favour of treatment (phase II fixed SS, HP0=0.4 HP1=0.7, H1)

Figure 10 – Study 4 true positive proportions using phase II fixed sample size of 50, 74 and 100 when HP_0=0.4.

Figure 10a, Figure 10b and Figure 10c display results under the alternative hypothesis when HP_1=0.5, 0.6 and 0.7 respectively.

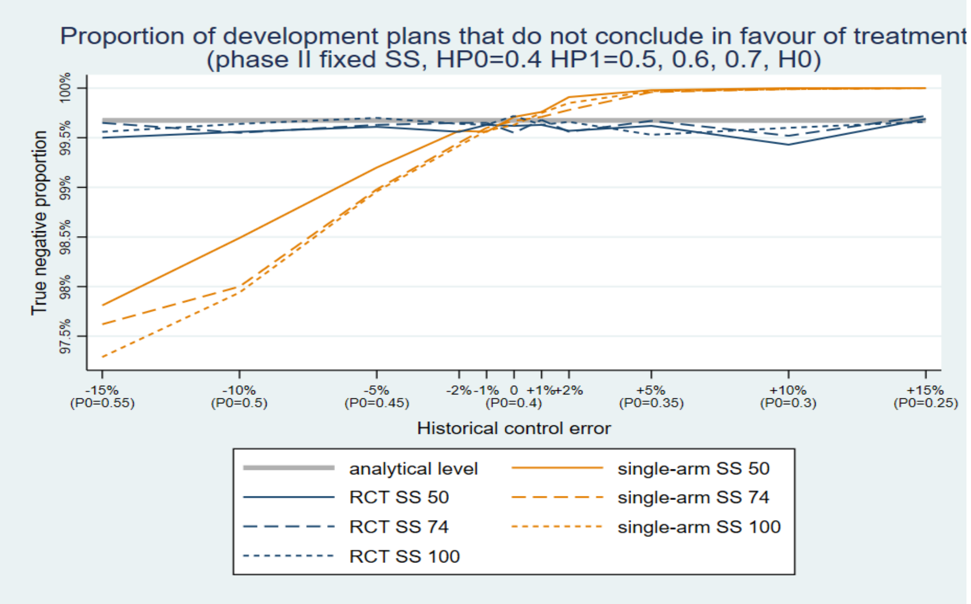


Figure 11 – Study 4 true negative proportions using phase II fixed sample size of 50, 74 and 100 for HP_1=0.2, 0.3 and 0.4 under the null hypothesis.

3.3.4.2 fixed alpha and beta

Figure 12 represents results of Study 4 when phase II sample size was calculated using α=0.15 and $1 - \beta$=0.8. In all scenarios, HP_0=0.4 and HP_1=0.5, 0.6 and 0.7.

Figure 12a represents results under the alternative hypothesis with true positive proportion on the y-axis and range of historical control error on the x-axis.

For the majority of cases considered, development plans with randomised phase II trials outperformed development plans with single-arm phase II trials. In the presence of positive historical control error (HP_0 > true P_0), development plans with randomised phase II trials performed best when HP_1=0.5. For example, when there was +15%-points of positive historical control error, true positive proportions were 50% for development plans with single-arm phase II trials and 88.1% for those with randomised phase II trials. Both development plans achieved very low true positive proportions in −10% and −15%-points of negative historical control error (HP_0 < true P_0). However, in these instances low proportions should not be interpreted as poor performance as true P_0≥0.5 and true P_1=0.5, therefore representing true negatives in which the null hypothesis should not have been rejected.

Figure 12b represents results under the null hypothesis with true negative proportion on the y-axis.

Figure 12b shows that under the null hypothesis, development plans with randomised phase II trials always outperformed those with single-arm phase II trials. Development plans with randomised phase II trials had true negative proportions that ranged from 99.5-99.7% in all instances of HP_1 and historical control error. However, development plans with single-arm phase II trials had true negative proportions that ranged from 97% to 98.6% across all levels of HP_1 and historical control error.

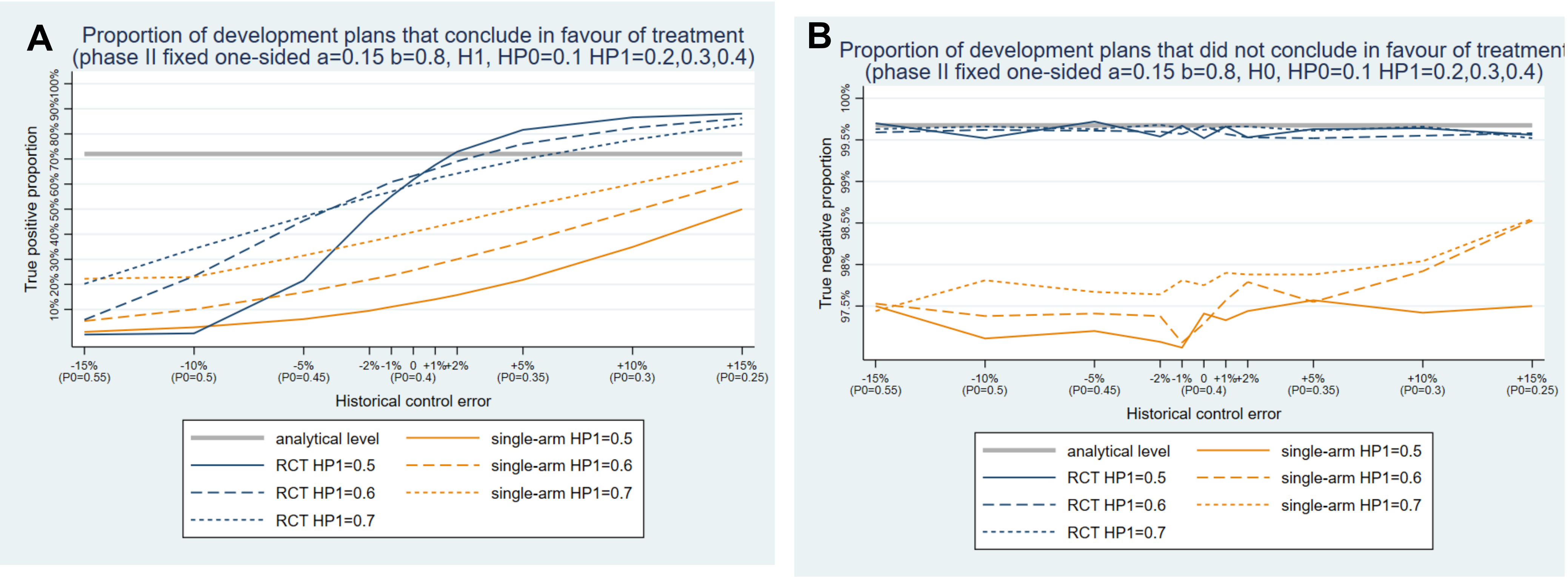


Figure 12 – Study 4 true negative and true positive proportions when phase II sample size is calculated using α=0.15 and $1-\beta$=0.8 with HP_0=0.4 and HP_1=0.5, 0.6, 0.7

Figure 12a and Figure 12b represent results under the alternative hypothesis and null hypothesis respectively.

3.3 Discussion

Studies 1-4 involved simulations which attempted to answer questions about optimal phase II design choice. These studies compared the performance of randomised and single-arm trials at phase II while considering three key elements identified in the narrative synthesis: impact on subsequent phase III trial conclusions, both alternative and null hypotheses, and historical control error. Each study brought a new layer of sophistication and external validity to reflect real life practice as an attempt to reflect concerns of over-simplification. With this in mind, results of Study 3 and Study 4 represent the most true-to-life simulations.

Overall, the results of the studies suggest that in most cases, development plans with single-arm phase II trials are preferred over development plans with randomised phase II trials. This is particularly the case when there is risk of negative historical control error, or if a phase II investigator has limited resources to patients. Therefore, it should be recommended that if an investigator is unsure of true P_0, a conservative value should be chosen for hypothesised P_0 to minimise risk of wrongful conclusions. However, if investigators have the flexibility to base sample size on fixed α and β, and conservative value of hypothesised P_0 was chosen, randomised phase II trials can be used.

One advantage of the studies was that they assessed performance of development plans with randomised phase II trials and single-arm phase II trials in two different ways; with fixed phase II sample size, and fixed one-sided α=0.15 and $1-\beta$=0.8 used calculate phase II sample size. Each of these were conducted to inform phase II investigators who may have different access to patient resources for their trials.

Another advantage to this research was that a wide range of treatment effects and response rates were investigated. This allows investigators to use the results of these studies if they have similar anticipated clinical trial parameters to guide their choice of phase II design. On the other hand, the wide range of response rates may be limiting to assess wider trends, as it is uncertain which response rates are likely to represent values seen in real practice. Therefore, using sources from real clinical trials may enhance this research.

These studies have provided a good foundation for investigating optimal phase II trial design. However, more research needs to be conducted. For example, some simulations were conducted based on limitless access to patient resources when an investigator could choose a phase II sample size based on a one-sided α=0.15 and a $1 - \beta$=0.8. In practice, it is likely an investigator would choose their phase II trial design based on the number of participants they anticipate they are likely to recruit.

Additionally, my recommendations are only useful to investigators which use binary outcomes at both phase II and phase III. This has limited applicability, as phase III studies usually use time-to-event endpoints. Not only this, but the studies also assume that the binary outcomes collected in phase II and phase III were from the same patient populations with the same treatment effect as phase III, which we would not anticipate in practice. My narrative synthesis identified that correspondence between phase II binary response rates and phase III time-to-event survival outcomes is a key issue in current practice. These assumptions may have influenced the conclusions which generally recommended the use of single-arm phase II trials over randomised phase II trials. Therefore, for the next set of studies, I will consider phase III time-to-event endpoints in addition to correspondence between phase II and phase III treatment effects. This is what I will explore in the next chapter.

4 Simulation Studies with Time-to-Event Outcomes and Imperfect Correspondence

4.1 Introduction

In Chapter 3, I assessed the performance of phase II-phase III development plans that contained the most commonly used phase II trial designs in oncology: single-arm trials and randomised trials. Three key elements were included to assess the performance of the development plans that were identified in my narrative synthesis presented in Chapter 2. These three elements included the assessment of: phase III trial conclusions, both alternative and null hypotheses, and historical control error. Performance was captured by true positive proportions and true negative proportions of the phase II–III development plans, which showed the proportion of simulations which either correctly concluded in favour of effective treatment, or correctly failed to reject the null hypothesis when the treatment was ineffective. These development plans were assessed in a variety of conditions including baseline treatment efficacy, treatment effect sizes and degrees of historical control error.

The results of these simulations showed that, in most circumstances, development plans with single-arm phase II trial designs were favoured over randomised phase II trial designs. However, one major limitation of these simulations is that they only considered binary outcomes for both phase II and phase III trials. In the context of oncology, it is common for phase II trials to assess interventions through binary response rate, and for phase III trials to assess interventions through time-to-event measures (40, 89). These can include overall survival or progression-free survival. Therefore, it is unknown whether the results of my simulations in Chapter 3 can be applied to development plans which use binary response rates at phase II and time-to-event survival outcomes at phase III. This is the fourth of the five key elements identified in my narrative synthesis.

Chapter 3 simulations also did not consider imperfect correspondence of treatment effect between phase II and phase III trials. The assumption of perfect correspondence between phase II and phase III treatment effects in Chapter 3 simulation studies may have also contributed to the results that favoured the use of single-arm phase II trials in most simulated circumstances. Imperfect correspondence is an important consideration, as treatment effects observed in phase II trials are commonly seen to be diluted in subsequent phase III trials (90). This could be for many reasons, such as the differences in patients recruited for phase II and III trials, differences in follow-up periods, and evidence that suggests that response rate (commonly measured by tumour shrinkage) may not directly map to improved survival (91). Imperfect correspondence between phase II and phase III treatment effects is the last of the five key elements identified in my narrative synthesis.

Overall, the aim of this chapter is to develop an exploratory simulation method which can demonstrate empirical correspondence between phase II binary outcomes and phase III time-to-event outcomes. Due to its exploratory nature, binary outcomes do not necessarily have to pertain to response rates. However, as simulated phase II trials would use binary outcomes and simulated phase III would use time-to-event outcomes, this would automatically introduce some lack of phase II-phase III correspondence. Not only this, but phase III design needs to be influenced by the results of phase II, despite the differing endpoints.

Implementing key elements four and five into simulations simultaneously is not straight forward. Therefore, the purpose of this chapter is to develop methods to introduce each key element within a simulation study and explain the possible statistical implications. As the focus was not to compare development plans with differing phase II trial designs, for simplicity, only development plans with randomised phase II trials were considered for this chapter. The methods developed could then be built upon in later chapters to assess impact of phase II trial design.

This chapter is therefore broken up into the following sections:

1. Create a simulation study which demonstrates a method to link phase II trials with binary outcomes to phase III trials with time-to-event outcomes.
2. Illustrate how imperfect correspondence occurs within the simulation study.
3. Motivation behind the next simulation studies for Chapter 5.

4.2 Time-to-Event Outcomes at Phase III

In Chapter 3, the same binary outcomes were used in phase II and phase III trials across the same development plan. Not only this, but treatment effect remained stable across both phases, which assumed the same population was tested throughout. The binary outcomes were interpreted as “proportion of patients responding to treatment” i.e., “response rate” in the control arm and experimental arm respectively. Using the same outcome between phase II and phase III allowed for easy translation of treatment effect between the two trials. However, creating a simulation that allows for differing outcomes between phase II and phase III means that there is no direct translation of treatment effect, except for when the same outcome exists but is measured in different ways. To allow for this, I have redefined phase II binary outcomes in this chapter. As such, phase II binary outcomes are interpreted as “proportion of patients who survive up to 12 months” and phase III time-to-event outcomes are interpreted as “overall survival”. To allow phase II results to influence phase III trial design, the results of phase II binary 12-month survival will be extrapolated into survival outcomes assuming the exponential survival function, which will be used for phase III sample size calculation.

To create these simulation studies, I started with the simplest case scenario where there was perfect correspondence between phase II and phase III treatment effects. This means that within the same development plan, phase II and phase III trials shared the same underlying outcome and patient population, implying that the same survival curves existed in experimental and control arms across the phases. Additionally, proportional hazards were correctly assumed. Phase II trials used a binary version of the outcome measured after 12 months, and subsequent phase III trials used the time-to-event outcome with administrative censoring at 60 months (five years). Similar to the simulated development plans in Chapter 3, I also simulated the simplest forms of each design, i.e. each design was conducted as a single-stage trial with only one experimental arm. Additionally, I did not distinguish simulated phase II trials as either phase IIA or phase IIB trials, rather they encompassed the phase II process as a whole. Furthermore, I have continued to assume that there is no missing data.

To give an example of how treatment effect was directly translated between the two phases, I will explain the scenario where phase II binary outcomes i.e., proportion of patients surviving 12-months, in the control arm and experimental arm were 40% and 70% respectively. These values were chosen as an illustration.

A visual representation of the translation of phase II binary outcomes and phase III time-to-event outcomes can be seen in Figure 13. The survival curve for the control arm is green, and the survival curve for the experimental arm is red. Survival curves were chosen such that proportion of patients who survived up to 12-months were 40% in the control arm and 70% in the experimental arm assuming the exponential curve and proportional hazards (using formulas (*1*) and (*3*)). This represents the binary outcomes that would be collected in phase II trials. This information was used to simulate phase III trials which collected overall survival with administrative censoring at 60 months.

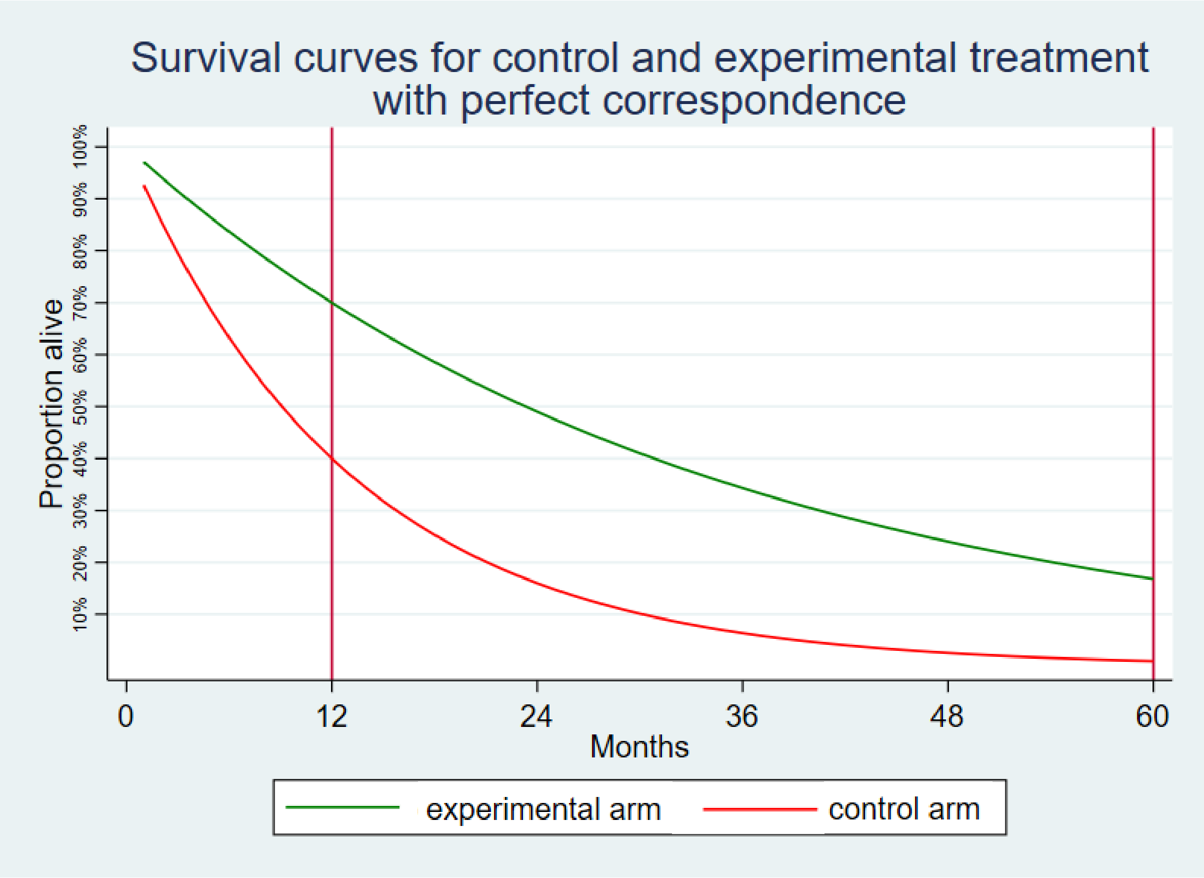


Figure 13 - Survival curves of control arm and experimental arm within a phase II-phase III development plan.

I will now describe the formulas used to calculate phase II and phase III outcomes in more detail.

To estimate survival time at 60 months, the following formula can be used:

$$S(t) = e^{-\lambda_x \mathrm{t}} \tag{1}$$

Where:

t is specific time point

$S(t)$ is survival proportion at time t

As this model assumes proportional hazards, hazard rates of control and experimental arms, λ_0 and λ_1, remain constant throughout the development plan. Hazard rates can be calculated using phase II survival proportions at t=12 months. This is defined as:

$$\lambda_x = \left(\frac{-\ln\left(S(t)_x\right)}{t}\right) \tag{2}$$

For the control arm, we know $S(12)_0$=0.4. From this, we can calculate the associated hazard rate using (3), which is λ_0 = 0.076. For the experimental arm, $S(12)_1$=0.7 with an associated hazard rate of λ_1= 0.058. We can plug these values of λ_0 and λ_1 into equation (1) to calculate expected survival time at 60-months. As such, 60-month survival proportions for control and experimental arms are expected to be 0.01 and 0.17 respectively. Hazard ratio (HR) can also be calculated to compare hazard of death between the two arms. HR is calculated by:

$$HR = \frac{\lambda_1}{\lambda_0} \tag{3}$$

The hazard ratio between the control and experimental arm is calculated as 0.39. This suggests that receiving experimental treatment reduces risk of death by 61% compared to control.

The HR and 60-month survival proportions can now be used to calculate phase III sample size. In the next section I will describe 1) logrank sample size calculation for time-to-event endpoints 2) statistical impact of using time-to-event endpoints by comparing two conceptual development plans.

4.2.1 Time-to-Event Sample Size and Statistical Implications

4.2.1.1 Sample size

As phase II and phase III trials collect different endpoints, different sample size calculations need to be used. As in Chapter 3, the likelihood ratio test is used to calculate sample size for phase II binary outcomes. For time-to-event outcomes the non-parametric logrank test was chosen to reflect common practice (92).

While the likelihood ratio test is a parametric sample size calculation, a non-parametric test was used for phase III sample size calculation, as survival data violates assumptions of normally distributed data. The exponential model is used as a simple way to simulate time-to-event data, such as overall survival, for which the logrank test is used to compare equivalence between the two survival curves. Another noted difference between the likelihood ratio test and logrank test to calculate sample size is the additional dimension of time that is collected from participants. Additionally, the proportional hazards assumptions also influences the logrank test which generates smaller required sample size than when non-proportional hazards is present.

For a trial that uses overall survival as an outcome, there is a number of deaths that need to occur to achieve a pre-specified level of power for a sample size calculation. Assuming there is equal participant allocation between control and experimental groups, number of deaths is given as:

$$d = \frac{4\left(z_{1-\beta} + z_{1-\alpha/2}\right)^2}{[\log(HR)]^2} \tag{4}$$

Where:

d = number of deaths required

$z_{1-\beta}$ – percentile of the normal distribution associated with power level, $1-\beta$

$z_{1-\alpha}$– percentile of the normal distribution associates with significance level, 1–$^{\alpha}/_{2}$

HR – hazard ratio

As stated, the specified number of deaths need to be met within the trial to achieve the level of desired power. However, because of administrative censoring at 60 months, it is not known if the required number of deaths can be achieved. Therefore, equation (4) is used as a basis to calculate number of patients needed across a 60-month period. Number of deaths required can therefore be rearranged as: (5)

$$d = \#events\ in\ control\ group + \#\ events\ in\ experimental\ group$$ (6)

$$d = \frac{n}{2}P(event\ by\ time\ t\ in\ control\ group) + \frac{n}{2}P(event\ by\ time\ t\ in\ experimental\ group)$$

Where:

n = number of patients required for the trial.

This can be rearranged to:

$$n = \frac{2d}{\pi_0 + \pi_1} \tag{7}$$

Where:

π_x = Probability an individual in treatment group x will die in the trial (0=control group, 1 = experimental group)

We assume the survival curves for the control groups and experimental groups follow the exponential distribution, which can be written as: (8)

$$T_i |\ X_i = x \sim Exp(\mu_x)$$

Where hazard is defined as:

(9)

$$\lambda_x = \frac{1}{\mu_x}$$

To calculate probability that an individual in treatment group x will die in the trial, we have:

(10)

$$\pi_x = P(T < \tau | \lambda_x) \qquad (10)$$

$$\pi_x = \int_0^{\tau} \lambda_x e^{-t\lambda_x}\, dt \qquad (11)$$

$$\pi_x = 1 - e^{-\tau\lambda_x} \qquad (12)$$

Substituting (10) into (7) means the number of patients required for a trial is:

(13)

$$n = \frac{2d}{2 - e^{-\tau\lambda_0} - e^{-\tau\lambda_1}}$$

Where:

τ = end-of-trial time

λ_0 = instantaneous risk of death in the control arm, i.e., control arm hazard rate

λ_1 = instantaneous risk of death in the experimental arm, i.e., experimental arm hazard rate

Using our illustrative example we have 12-month control survival proportion, P_0=40% and 12-month experimental survival proportion, P_1=70%. Putting appropriate values in equations (1), (3) and (13), we know that when the total number of patients required in the 60-month phase III trial is 46, with 23 patients in each control and experimental arm.

Introducing phase III time-to-event outcomes within a phase II-phase III development plan means different sample size calculations were used for each phase: likelihood ratio test for phase II, and logrank test for phase III. The use of the different sample size calculations have statistical implications compared to the development plans used in Chapter 3 which used binary outcomes throughout.

For example, when phase III trials follow phase II trials with the same 12-month treatment effect, those that use time-to-event outcomes require a smaller sample size due to the proportional hazards assumption. This is evident in Figure 14.

Here the x-axis represents sample size that was calculated using the likelihood ratio test for development plans with phase III binary outcomes, titled SS[bin]. The y-axis represents the sample size that was calculated for development plans using the logrank test with phase III time-to-event outcomes, titles SS[HR]. There were six development plans for which development plan sample size was calculated, which all used the same phase II P_0=0.4 but had differing levels of phase II P_1=0.5, 0.6 and 0.7. For each level of P_1, there were two different development plans for which the subsequent phase III trial collected different types of data, either binary, or time-to-event. For development plans with phase III binary outcomes, the same values of P_0 and P_1 were used across the phases. For development plans with phase III time-to-event outcomes, 60-month phase III survival outcomes were collected after phase II trials with 12-month binary outcomes. Phase II sample sizes were calculated based on a one-sided α=0.15, and $1-\beta$=0.8. Phase III sample sizes were calculated based on a one-sided α=0.025, and $1-\beta$=0.9.

In Figure 14 the red identity line represents equal development plan sample sizes for the two phase III sample size calculations used. The blue dotted line represents the results between the two sample size calculation methods given the different levels of phase II P_1. Not only does Figure 14 demonstrate that using time-to-event outcomes in phase III reduces the development plan sample size, but that the efficacy in which the logrank test uses the participants increases as the gap between P_0 and P_1 gets wider.

This means that with the same number of participants in a phase III trial, development plans that use time-to-event outcomes at phase III return a higher level of power. This is evident in Figure 15. Here, the same development plans described in Figure 15 were simulated, fixing the sample size calculation at the number given for when binary outcomes were used at both phases. Observed power was recorded given the results of likelihood ratio tests for phase II and phase III binary outcomes, and logrank test for development plans with phase III time-to-event outcomes. 1000 repetitions were run for each development plan.

Figure 15 displays simulated P_1 value on the x-axis and observed power across development plans on the y-axis. The figure shows that, when sample size is fixed, development plans that use binary outcomes at phase III reach observed power at 72% represented by the blue line labelled binary outcome. This is as expected for the levels of β used in sample size calculations (0.8*0.9=0.72), However, for the same sample size, development plans that use time-to-event outcomes at phase III reach nearly ~80% observed power, represented by the orange line labelled TTE outcome. It should be noted that impact on sample size and power is likely due to the proportional-hazards assumption rather than the use of time-to-event outcomes.

Creating a simulation study using time-to-event outcomes in a phase III trial is more clinically relevant to oncology trials in practice than binary outcomes, and comes with the additional benefit of requiring less participants for the same level of power under the assumption of proportional hazards.

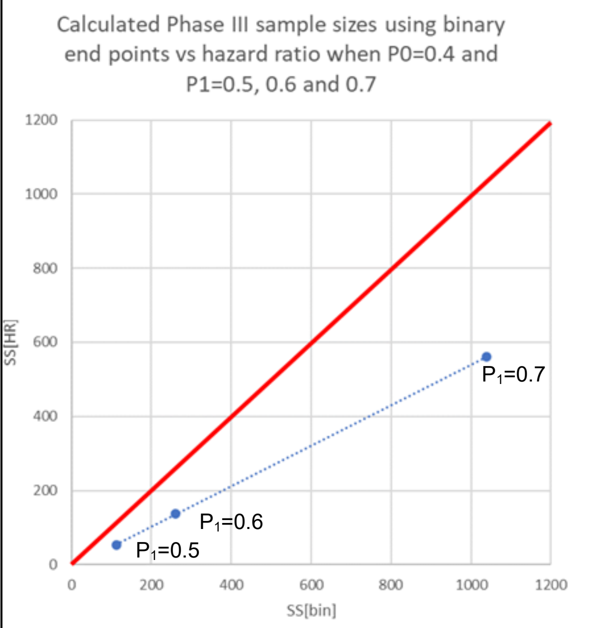


Figure 14 – Graph to show total development plan sample size given phase III binary end points (x-axis) and phase III hazard ratio outcome (y-axis).

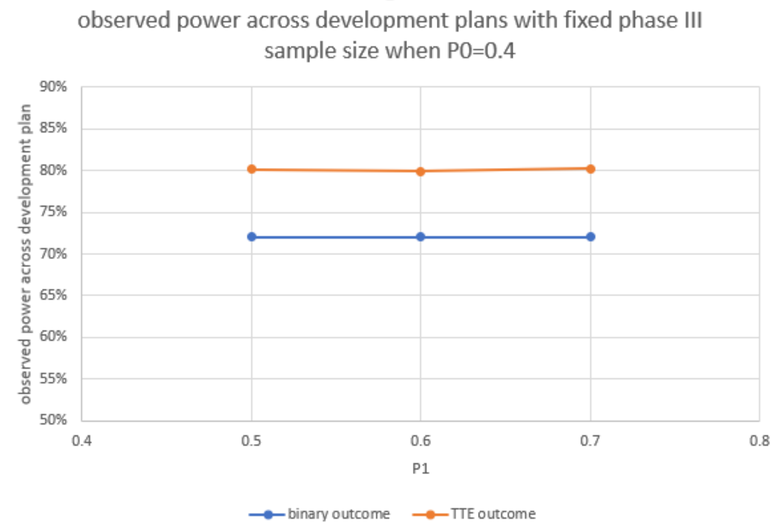


Figure 15 – Graph comparing observed power across a phase II-phase III development plan considering the phase III type of outcome and the method used for determining sample size.

The next section introduces the concept of the dispersion of data between phase II and phase III X^2 statistics to further understand impact of treatment effect on phase III trial conclusions, the first key element identified in my narrative synthesis.

4.2.1.2 X^2 Test-statistics

An alternative way to visualise the connection between phase II and phase III endpoints is to create a scatter plot of X^2 test-statistics at each phase within a phase II-phase III development plan. Specifically, the X^2 test-statistics that are reported from the phase II likelihood ratio test, and the X^2 test-statistics that are reported from the phase III logrank test. The purpose of this plot was to assess proportion of X^2 test-statistics that fell within each of the four quadrants: those that are statistically significant at both phases, those that are only statistically significant at phase II, those that are only statistically significant at phase III and those that are not statistically significant at either phase.

X^2 test-statistics were generated from the following a development plan: a randomised phase II trial with binary outcomes followed by a randomised phase III trial with time-to-event outcomes. The control arm and the experimental arm each share the same survival curves across both phases, and at 12-months the proportion surviving in the control arm is 40% and in the experimental arm is 70%. Phase II sample size is calculated using 12-month proportions with one-sided α=0.15 and $1-\beta$=0.8, and phase III sample size is calculated using extrapolated 60-month proportions with one-sided α=0.025 and $1-\beta$=0.9. The phase II null and alternate hypotheses tested are written as $H_0: P_1 \leq P_0$ and $H_1: P_1 > P_0$, where P_0 and P_1 are 12-month proportions in the control and experimental arms. The phase III null and alternate hypotheses are written as $H_0: HR \geq 1$ and $H_1: HR < 1$. 10000 repetitions were generated.

Of the 10000 repetitions, six were significant in the wrong direction at phase II, i.e. when the control arm was better than the experimental arm. As I was only interested in the one-sided tests where the experimental arm is better than the control, these six results would not have been one-sided significant. Ideally, these six results would have been reclassified as insignificant, however, it was not clear which X^2 value they held; therefore they were removed from the dataset. Figure 16 presents results of the simulations. The plot is split into four quadrants which are determined by X^2 significance threshold at each phase II and phase III (1.074 at phase II and 3.84 at phase III). The proportion of the remaining X^2 test-statistics which exist in each quadrant are listed to the right of the figure in each respective colour.

The observed power for the phase II trials is slightly lower than the designed 80% level at 77.5% (71.7% + 5.8%) which is calculated when green and yellow quadrant proportions are added together.

The observed power for the phase III trials is larger than the designed 90% and is 92.7% (71.7% + 21%) which is calculated when green and blue quadrant proportions are added together. The reason why observed power is greater than 90% is because of the associated sample size; 30 allocated to each group. When sample sizes are small, each patient represents a larger proportion of results; therefore, it is difficult to provide a required sample size with the exact requirements of α and β for a large treatment effect. Therefore, α and β are treated as minimum limits that the sample size must reach, and along with treatment effect, the smallest possible sample size is provided.

It is interesting to note that the plot shows very little correlation: -0.0039. This is as expected as the development plans are simulated in such a way that phase II trials do not influence the phase III design and are therefore independent from each other. Therefore, it follows that true correlation must be 0.

What is important is the proportions of X^2 test-statistics which lie in each of the four quadrants. We expect the dispersion (but not correlation) of X^2 data points to change as imperfect correspondence is introduced to the simulation study.

Figure 16 depicts the results of development plans such that they are divided into quadrants. Development plans that rejected H_0 at phase II and phase III are in the green quadrant, and those that rejected H_0 only at phase II are in the blue quadrant. Additionally, development plans that rejected H_0 only at phase III are in the yellow quadrant, and those that failed to reject the null hypothesis at phase II and phase III are in the red quadrant. Associated proportions for each quadrant are 71.7%, 5.8%, 21% and 1.5% for green, yellow, blue and red quadrants respectively.

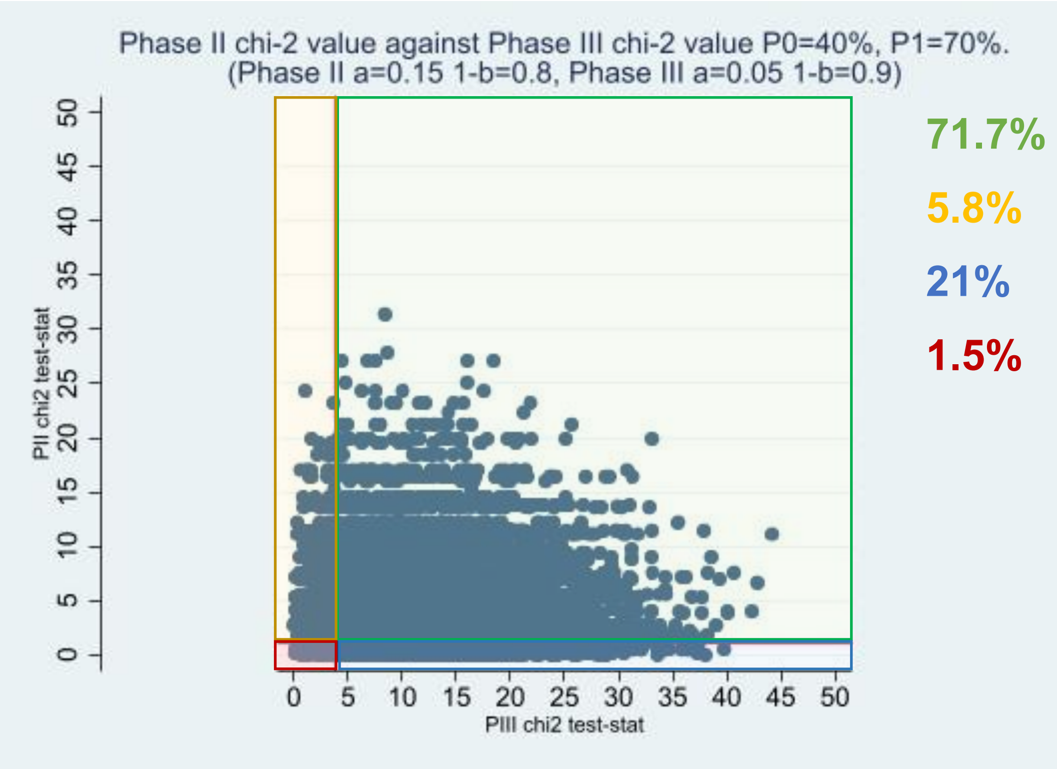


Figure 16 – Scatter plot for X^2 statistics for phase II & III trials within a development plan.

4.3 Imperfect Correspondence Between Phase II and Phase III: Example

Imperfect correspondence between phase II and phase III trials in oncology has been well documented (90). There are many possible reasons as to what causes this. One of the possibilities is that binary response rate often used in phase II trials are not a good surrogate for overall survival often used in phase III (93). Additionally, as phase II trials are often shorter than phase III trials, the response data collected within the limited time frame may not accurately capture true response rates that would exist in in a longer phase III trial. Furthermore, phase II trials often involve stricter inclusion criteria than phase III, therefore the demographic between the two trials is not wholly comparable.

The multiple reasons which may contribute to imperfect correspondence make it difficult to determine a single mechanism within the simulation study that would produce differences between phase II and phase III treatment effects. Therefore, to illustrate how imperfect correspondence could be present within my study, I will present an example of how it could occur: in this instance, with a shift of the hazard ratio towards the null from phase II to phase III via the experimental arm. The rationale for this was that phase II trials that happen to estimate more optimistic treatment effects are more likely to lead to phase III trials, than phase II trials that happen to estimate more pessimistic treatment effects. I will explain how the example will be incorporated into the simulation study, then discuss potential statistical implications of imperfect correspondence whilst using my example.

Using the same illustrative development plans from section 4.2.1.2, phase II estimates of 12-month survival in control and experimental arms were P_0=40% and P_1=70% respectively, which gave an associated hazard ratio of 0.39. It was thought that these values represented the latest evidence available to phase III investigators, and therefore were used to extrapolate 60-month survival times to calculate phase III sample size.

Subsequently, I assumed that within a development plan, the control arm survival curve remained constant between phase II and phase III. However, as previously mentioned, I assumed the phase III hazard ratio was now diluted from the phase II HR of 0.39. This was represented by a multiplicative factor, defined at 1.5. As this is a proof-of-concept simulation study, the value of 1.5 was chosen to demonstrate a noticeable shift in the hazard ratio towards the null. For example, multiplying the phase II hazard ratio, 0.39, by the multiplicative factor, 1.5, resulted in 0.58. 0.58 defined the subsequent phase III hazard ratio which, demonstrating a shift towards the null of 1 from phase II to phase III. This also can be seen as a shift in experimental survival curves within the same development plan between phase II and phase III.

A visual representation of this imperfect correspondence mechanism is provided in Figure 17. The red line represents the survival curve in the control arm across both phases with λ_0. The solid green line represents the survival curve in the experimental arm in phase II with λ_1. The dashed green line represents the survival curve in the experimental arm in phase III with imperfect correspondence, λ_{IC1}. The two vertical lines represent the lengths of each of the trials, 12 months for the phase II trial, and

60 months for the phase III. As demonstrated, the difference between the control and experimental survival curves is lessened from phase II to phase III.

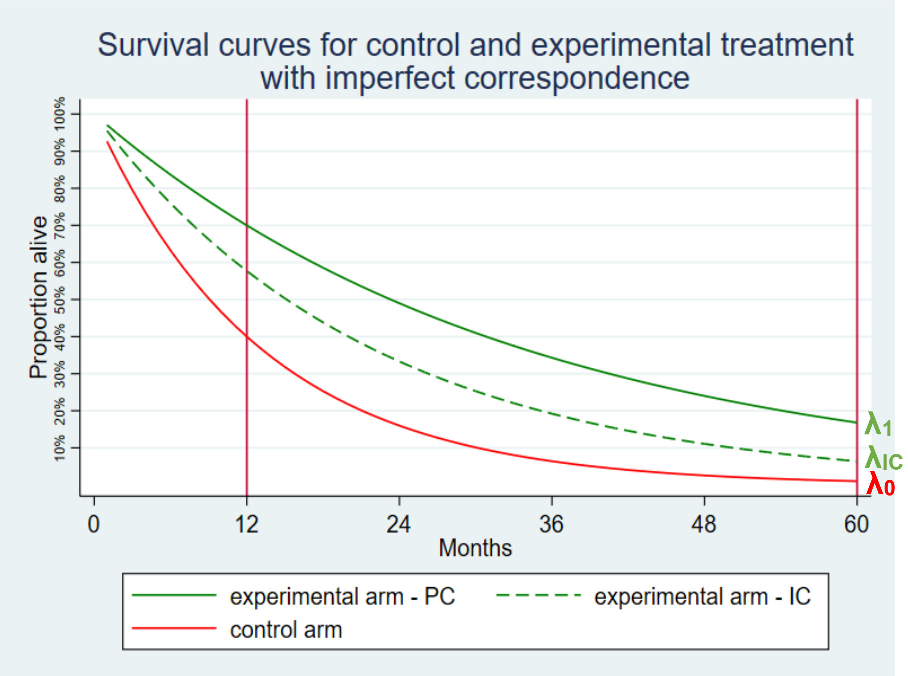


Figure 17 – Survival curves of control and treatment arms when imperfect correspondence is present in phase III

Before phase III trials were simulated, sample sizes were based on logrank calculations using λ_0 and λ_1. Therefore, it was calculated assuming the phase II treatment effect was present in the phase III trial. However, for the phase III trial itself, the true values were simulated based on λ_0 and λ_{IC1}, generating a diluted treatment effect in a phase III setting.

Equations (1), (3) and (2) were used to calculate the new imperfect correspondence parameters for the experimental arm in a phase III setting to use as “the truth” in simulations. The imperfect correspondence λ in phase III is denoted as λ_{IC1} and was calculated by:

$$\lambda_{IC1} = \lambda_0 \times HR_{IC} \qquad (14)$$

4.3.2 Imperfect Correlation Impact on X^2 Statistics

The statistical implications of the illustrative example of imperfect correspondence were assessed by plotting the phase II and phase III X^2 test-statistics from the development plans. From these scatter plots, proportions of trials which were deemed significant at each phase can be compared to the ideal scenario presented in section 4.2.1.2.

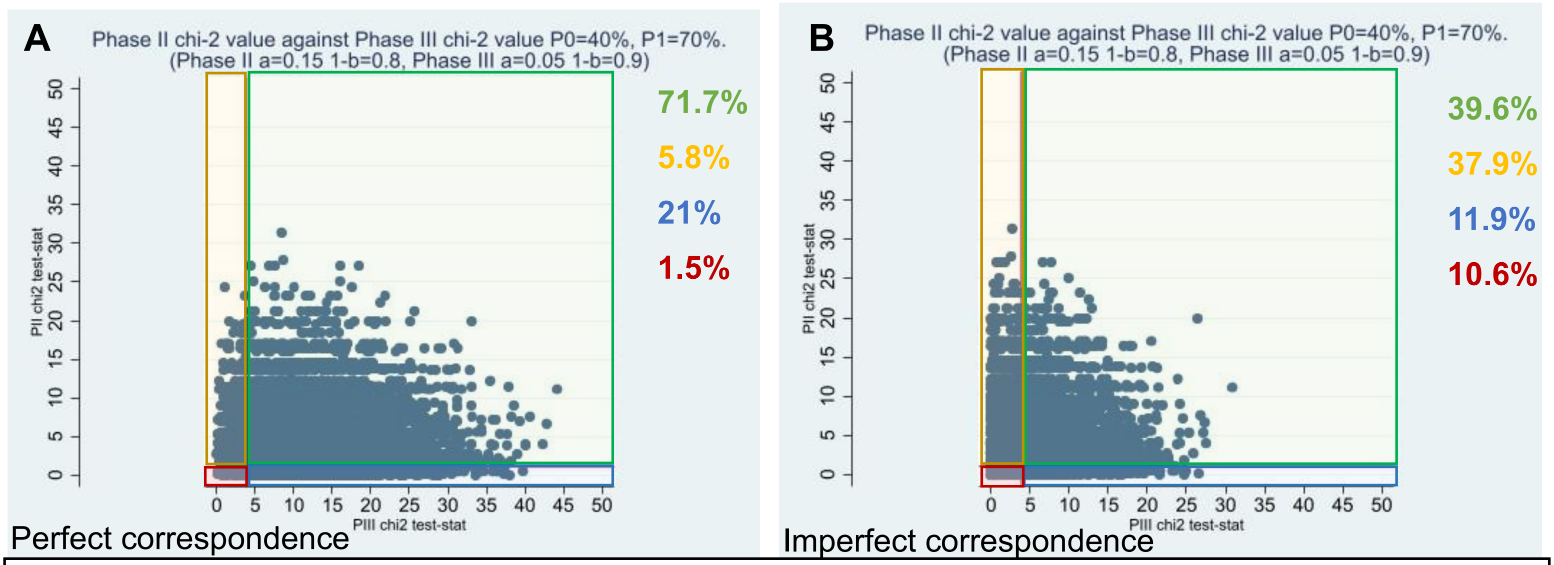
For the illustrative example of imperfect correspondence, the same method was used to simulate development plans and extract X^2 test-statistics as described in section 4.2.1.2. The difference being that imperfect correspondence was introduced in phase III trials, where λ_1 and λ_0 were used to design phase III sample size, but λ_{IC1} and λ_0 were used to generate the "truth".

Out of the 10000 repetitions of each development plan simulated, six X^2 test-statistics were removed from the analysis due to being significant in the wrong direction (all six were significant in the wrong direction only at phase II). The correlation between the phase II and phase III X^2 test-statistics was -0.0028. This may go against what is expected. Intuitively, if the alternative hypothesis is true and correct treatment effect is used for phase II and phase III sample sizes, you may expect there to be a positive correlation between X^2 plots of the two trials. However, this was not the case for these simulations. Not only were the two phases simulated independently from each other, but the treatment effect was also incorrect for phase III sample size calculation.

Figure 18A displays the X^2 test-statistics plots from perfect correspondence as seen in section 4.2.1.2, and Figure 18B displays the X^2 test-statistics plots of imperfect correspondence. Figure 18B shows that in imperfect correspondence, the proportion of phase II trials which rejected the null hypothesis is 77.5% (39.6%+37.9%); the same proportion seen when there is perfect correspondence (seen in Figure 18A). However, the proportion of phase III trials which reject the null hypothesis is 51.5% (39.6%+11.9%), much lower than the designed power level of 90%.

In comparing the two graphs, we can see that when imperfect correspondence is introduced, a large proportion of X^2 data points shift from the green and blue quadrants into the yellow and red quadrants. This demonstrates the impact of imperfect correspondence on statistical power when phase II hazard ratio shifts from 0.39 to a phase III hazard ratio of 0.58, which is reduced by 41.2%-points (92.7%-51.5%). Therefore, imperfect correspondence has the potential to have huge repercussions on phase III trials and their ability to effectively conclude in favour of truly effective treatment.

All analysis was conducted using Stata version 16.0, and example code of this simulation is provided in the appendix.



Phase II chi-2 value against Phase III chi-2 value P0=40%, P1=70%.
(Phase II a=0.15 1-b=0.8, Phase III a=0.05 1-b=0.9)

Perfect correspondence

Phase II chi-2 value against Phase III chi-2 value P0=40%, P1=70%.
(Phase II a=0.15 1-b=0.8, Phase III a=0.05 1-b=0.9)

Imperfect correspondence

Figure 18 – X^2 plots of phase II and phase III trials within the same development plan, and associated proportions within each quadrant determined by X^2 significance threshold at each phase: 1.07 at phase II and 3.84 at phase III.

4.4 Remarks and Next Steps

The simulation methods developed throughout this chapter illustrate how phase II-phase III development plans can be assessed with phase II binary & phase III time-to-event outcomes in the presence of imperfect correspondence. The methods can be combined with the ones developed in Chapter 3 to robustly assess impact of phase II trial design on a phase II-phase III development plan that considers all five key elements. However, it should be noted that this chapter did not consider phase II binary response rates which are commonly used in phase II trials, will be discussed later.

Imperfect correspondence has been a prevailing issue within oncology clinical trials, with evidence reporting that phase II treatment effects are diluted in subsequent phase III trials (90). This chapter has introduced a mechanism for imperfect correspondence to be implemented in simulation studies, but a limitation is that it may not be a realistic way imperfect correspondence behaves in practice. Furthermore, it is impossible to characterise all the ways in which this may occur, as many of them are unknown. Not only this, but without looking at pairings of phase II-phase III trials that have already been conducted, it is difficult to make an assessment on what “reasonable” or “unreasonable” levels of correspondence would be, or even what perfect correspondence would look like between binary response rates and overall survival. Therefore, it is difficult to make definitive statements on the impact of imperfect correspondence on a phase II-phase III development plan using the results from this chapter.

In an attempt to simulate realistic imperfect correspondence, the next simulation studies could be based on real data extracted from phase II-phase III development plans. The simplest circumstance to simulate would be based on phase II randomised trials that collected the same primary outcome as the phase III trial. Ideally, the phase II trials would provide phase III sample size calculations with good quality evidence about the survival in the experimental and control arm. However, this thesis also aims to explore the impact of single-arm phase II trials on subsequent phase III conclusions. Therefore, the next level of complexity would be to also simulate real data from single-arm phase II trials that collect the same primary outcome as the phase III trial. In these circumstances, the only information that can be provided to the phase III trial sample size calculation is in the experimental arm. For the estimate of phase III control survival, the value that was used to calculate phase II sample size could be used. However, as discussed, phase II trials often use binary response rate as their primary outcome, unlike the binary endpoints simulated in this chapter. This makes it difficult to directly inform phase III sample size that collects survival outcomes. As the fourth key element defined that phase II binary response rates need to be considered, in addition to phase III time-to-event survival outcomes, this element is yet to be fully addressed.

As a solution, real examples of phase II trials that have collected both binary response rates and time-to-event survival can be used in which to extract data to inform realistic simulations of development plans. From here, simulated phase II response rates can be treated as a primary outcome. This can be done by using the results of the statistical test comparing control and experimental response rates to inform the decision to proceed to a phase III trial. Then, phase II time-to-event survival outcomes can be used to influence phase III sample size, thereby linking the two phases. This offers a way to include realistic imperfect correspondence between phase II response rates and phase III survival that have been seen in practice. This is a pragmatic, empirical way to define ‘correspondence’ and is what will be explored in Chapter 5.

5. Simulation Studies with Phase II Response Rates and Phase III Survival Endpoints

5.1 Introduction

In chapters 3-4, I reported simulation studies which addressed the five key elements separately. Chapter 3 focussed on key elements one, two and three, and Chapter 4 focussed on key elements four and five. I will now briefly recap the simulations developed in Chapter 3 and 4, and then describe how I aim to combine the work in order to create a simulation study which can address all five key elements at once.

5.1.1. Chapter 3 Recap

Chapter 3 focused on a simulation study of a phase II-phase III development plan which considered the following key elements identified in my narrative synthesis:

- The alternative and null hypothesis
- Phase II design impact on phase III trials
- Historical control error

I did this by assessing the performance of two different phase II-phase III development plans: one with a single-arm phase II trial and one with a randomised phase II trial. For simplicity, binary outcomes were simulated at both phases. Various phase II sample sizes were investigated, including three levels of fixed sample size: 50, 74 and 100, and a required sample size obtained via sample size calculation.

The performance measure was the proportion of times the correct decision was made on treatment. Under the alternative hypothesis, this was the proportion of times a development plan concluded in favour of treatment at the end of a phase III trial. Under the null hypothesis, this was the proportion of times a development plan did not conclude in favour of treatment, either at the end of phase II or at phase III.

In these studies, it was found that generally, development plans with single-arm phase II trials were more likely to make correct conclusions on the treatment. This likelihood increased in two particular scenarios: 1) when the hypothesised $P_{0,}$ used for sample size calculation and historical control error, was less than the true P_0, and 2) when there was not access to a large sample size.

However, when the phase II sample size was large and hypothesised P_0 was greater than true P_0, then the development plans with either phase II trial design performed similarly.

5.1.2 Chapter 4 Recap

In Chapter 4, I created a proof-of-concept simulation study which focussed on the remaining two key elements identified in my narrative synthesis:

- Different phase II and phase III endpoints
- Imperfect correspondence between phase II and phase III treatment effect

For this purpose, only development plans with randomised phase II trials were considered.

Firstly, to implement different phase II-phase III endpoints, phase II had binary endpoints and phase III had time-to-event endpoints. For this simulation study, phase II binary outcome was defined as 12-month survival proportion instead of binary response rate which was used in the previous chapter. Phase III time-to-event outcome was overall survival across a 60-month period. If a phase II trial concluded in favour of treatment, the subsequent phase III sample size was calculated by taking the 12-month survival proportion in each arm and extrapolating them to suit a 60-month trial.

Under perfect correspondence, it was assumed that across a phase II-phase III development plan the two phases would share the same hazard ratio under their relative survival curves. Under imperfect correspondence, it was assumed the hazard ratio moved towards the null in the phase III trials. For this, true phase II hazard ratio was multiplied by an ‘imperfect correspondence factor’ to create the true phase III hazard ratio used to simulate phase III trial results. To reflect real-life practice, imperfect correspondence was not accounted for in the design of the phase III trial, but was used when simulating the “truth” for trial results.

At the end of the simulation studies, it was found that development plans with randomised controlled trials correctly made decisions on treatment 72% of the time when there was perfect correspondence between phase II and phase III trials. This was expected, as the $1-\beta$ levels chosen for phase II and phase III sample size calculations were 80% and 90% respectively (0.8 * 0.9 = 0.72). However, it was found that if the imperfect correspondence factor was 1.5, it reduced the proportion of times the development plan made a correct decision on treatment to 40%.

The five key elements have been implemented separately in simulation studies seen in Chapter 3 and Chapter 4. In the present chapter, I design and report a final simulation study that combines all five.

5.1.3 Combining Five Key Elements.

Choosing *data generating mechanisms* to enact a simulation study that accounts for all key elements is not simple.

One of the issues addressed previously is how to define realistic imperfect correspondence within a simulation, given that phase II trials often collect binary data, and phase III trials often collect time-to-event data. Chapter 4 illustrated one way for imperfect correspondence to be incorporated in simulations. However, the binary data collected in Chapter 4 simulated phase II trials was “proportion of survival at 12 months”, which could also be treated as a time-to-event outcome, allowing the implementation of imperfect correspondence. In reality, the most common phase II binary endpoint is “response rate”, which cannot be treated as a time-to-event outcome (94). Therefore, to allow these simulations to reflect common practice while incorporating all five key elements, the phase II trials need to collect both binary response rate and time-to-event survival data.

Although it is not uncommon for phase II trials to collect time-to-event data, due to the different research aims of phase II and phase III trials, it is rare that the same type of time-to-event data is collected across a development plan. Without collecting the same type of outcomes across both phases, imperfect correspondence cannot be calculated. Therefore, without an abundance of development plans which have collected the same outcomes across phase II and phase III trials, it is very difficult to define a realistic set of imperfect correspondence parameters to implement in a simulation study.

However, in the rare instances where real-life development plans *have* collected the same time-to-event data across phases, and where the phase II trial has collected both response rate and survival outcomes, data can be extracted to be used for illustrative purposes for simulation studies. Using real-life development plans will not define a ‘typical’ range of imperfect correspondence factors for investigators to account for in practice, but instead, represent instances of imperfect correspondence that have been seen. Extracting data from these real-life development plans will define the *data generating mechanisms* within the simulation studies in this chapter. This will allow us to create simulation studies while considering all five key elements and reflect a handful of interactions seen in practice. This is preferred to simulating a wide range of values, which is computationally demanding and without the benefit of knowing which ones are reflective of real life.

The simulation studies will compare the performance of real phase II-phase III development plans with the phase II design actually used, e.g. a randomised trial, and one had a different phase II trial been used e.g. single-arm.

5.2 Methods

As discussed in the introduction, real-life clinical trial reports were used to inform the simulation studies in this chapter. This methods section discusses the reasoning behind this methodology, then details the systematic literature review used to identify the real-life phase II-phase III trial pairings. Next, this section discusses how these examples informed the creation of the simulation studies, and finally, explains the design of the simulation study itself.

As previously mentioned, to conduct a simulation that incorporates key elements four and five (consideration of differing phase endpoints and imperfect correspondence) while simulating realistic development plans, simulated phase II trials would require collection of both binary response rates and time-to-event survival data.

After consideration, the use of existing pairs of phase II-phase III trials was determined to be the best way to choose *data generating mechanisms* to reflect realism, both in terms of imperfect correspondence and the relationship between binary and time-to-event outcomes.

Initially, other methods were considered. For example, a wide range of imperfect correspondence *data generating mechanisms* could have been chosen to reflect possible phase III scenarios following a phase II trial e.g. perfect correspondence, moderate correspondence and mild correspondence between phase II and phase III time-to-event outcomes.

However, the issue would still persist on how response rates would correlate to time-to-event outcomes within the same phase II trial in a realistic way. This is particularly difficult as correlation between response rates and survival has not been well defined (69-71, 95-98).

One solution to this is to simulate many scenarios of phase II response rate and survival combinations, followed by the three levels of imperfect correspondence for the subsequent phase III trial. From here, performance of development plans could be assessed to determine optimal phase II design choice for each simulated circumstance.

From these results, phase II investigators could find which simulated circumstance is the closest match to their trial parameters to use as a guide for optimal design choice.

Using this methodology, the results would have limited usefulness to phase II investigators attempting to choose an optimal phase II design. Considering the difficulty in defining correlation between response rates and survival, it is unlikely investigators will know which of the simulated relationships between the two endpoints will apply to them ahead of the phase II trial. In the same vein, its unlikely investigators will be able to anticipate the level of imperfect correspondence that should be expected in the subsequent phase III trial. Not only this, but it would be difficult to determine which of the simulated circumstances would be unlikely to reflect real practice.

Extracting *data generating mechanisms* of phase II and phase III endpoints from existing development plans grounds the simulation studies in scenarios that have been seen in practice before.

Not only does this method offer realism with imperfect correspondence and relationships between response rate and survival, but also provides further realism for simulated phase II treatment effects and phase III hazard ratios.

Only published phase II-phase III pairings with randomised phase II trials could be used for this simulation study. This is because trial pairings with a single-arm phase II trial could not provide the information needed to simulate the control arm of an equivalent randomised trial. There is no way to determine what the concurrent control arm estimate would have been. However, published phase II-phase III pairings with a randomised phase II trial could be realistically simulated as if it was a single-arm phase II trial. Specifically, the historical control that a single-arm trial would likely use could be extracted from the hypothetical control arm value used to calculate the randomised phase II sample size.

Therefore, the research question for this chapter becomes: how would the performance of development plans differ had the choice of phase II trial design been a single-arm trial instead of a randomised trial?

The literature review details how real-life pairs of trials in a development plan with randomised phase II trials were selected.

5.2.1 Literature Review of Clinical Trial Reports

To find real-life development plans for the simulation studies, phase III trials were first selected and, if appropriate, the previous phase II trials were then found in the citation of the paper and scanned for suitability. This backwards approach was chosen as a preceding phase II trial could be expected, whereas a forward approach starting from phase II clinical trials would not guarantee a subsequent phase III. It should be noted that once a phase III trial was identified, the preceding phase II could be identified as either a phase IIA or phase IIB to still be included for my simulation study.

Table 8 describes the requirements for each published pairing of phase II-phase III trials, why, and when relevant how this requirement fulfilled the key elements. As a reminder, the key elements are for the simulations to consider 1) both phase II and phase III trials 2) null and alternative hypothesis 3) historical control error 4) differing phase endpoints and 5) imperfect correspondence of treatment effects between phases. It should be noted that by identifying both phase II and phase III pairings, key element #1 is fulfilled.

Development plan phase	Requirements	Why
Published phase III	Trials identified from clinicaltrials.gov search results had to have a full paper of published results linked on its own clinicaltrials.gov page	To guarantee phase III trials fulfilled the clinicaltrial.gov inclusion criteria. These criteria ensured standardization of high-quality published phase III trials with useable results needed for the simulation study. Phase III trials that were identified through the clinicaltrial.gov search fulfilled the following criteria: they were interventional, randomised, completed trials with results and provided access to statistical analysis plans
	Trials conducted in the context of cancer	To ensure published phase II-phase III pairings chosen for the simulations reflected real cancer clinical trial environments
	Trials were efficacy studies	To ensure published phase II-phase III pairings chosen were relevant to the research aims. More specifically, I was only interested in trials which investigate treatments that aim to improve response rates or survival outcomes.
	Trials tested a superiority hypothesis as primary outcome (not equivalence or non-inferiority)	To ensure published phase II-phase III pairings chosen assessed treatment effects, and relevant endpoints were collected. Specifically, survival outcomes with hazard ratios

	Trials had to have at least one control arm and one experimental arm	This would ensure that phase III trials used the gold-standard randomised-controlled trial design. Due to the high-quality trial design, the treatment effect estimates of these trials could be considered close estimations of the truth, and could help determine whether it was under the alternative or null hypothesis (key element #2)
	Trials collected time-to-event endpoints, for which there was graphical representation of the data using a Kaplan-Meier curve	This would ensure the simulation study had access to published phase III survival outcomes which could be compared against phase II survival outcomes to determine imperfect correspondence (key element #5). The graphical representation of the Kaplan-Meier curve would be used to determine a digitized pool of patients to sample simulated phase III results from (more details in section 5.2.2.1 – Data Generating Mechanisms)
	Trial report had to have a previous randomised phase II trial cited in the introduction	This was to identify the preceding phase II trial, and ensure the phase II trial is within the same disease area and used the same treatment arms (same dosages/ dosing schedules not necessary).
Once an appropriate phase III trial fulfilled all requirements, the preceding phase II trial was identified. This phase II trial then had to fulfil another set of requirements before using these phase II-phase III development plan as an example in my simulations:		
Published phase II	Trial had to be in the same disease area as the subsequent phase III trial	This was to ensure the phase II and phase III trials were within the same development plan. This also ensures that simulated phase II trial estimates could be reasonably used for subsequent phase III sample size calculations

	Trial had to have at least one control arm and at least one experimental arm that matched the treatment arms in the subsequent phase III trial	This would guarantee the previous phase II trial that was within the same development plan. Estimates from the control arm and the experimental arm also determined the simulation treatment effect "truth" for each simulated phase II
	Trial had to be a stand-alone trial, i.e. not a meta-analysis	This would clearly identify one set of response rate values for the hypothesised control arm, hypothesised experimental arm, estimated control arm and estimated experimental arm. Hypothesised values would be used to determine simulated phase II sample size. Post-trial estimates from the control arm and the experimental arm would be used to represent the simulation treatment effect "truth" for each simulated phase II. The hypothesised control arm value would also be used to determine the historical control for the simulation of the equivalent single-arm phase II trial.
	Trial had to collect the same type of time-to-event endpoint as the phase III trial, for which there was a graphical representation of the data using a Kaplan-Meier curve.	This would guarantee that the survival outcomes between phase II and phase III trials can be compared, from which imperfect correspondence between the two phases can be inferred (key element #5). The inclusion of the Kaplan-Meier curve ensured that the data could be digitized to mimic patient-level data to infer survival proportion at the end of the phase II trial. These datapoints would be adapted by the results of the simulated phase II trial to determine hypothesised survival proportion needed for the subsequent simulated phase III sample size calculation (more details in section 5.2.2.1 - Data Generating Mechanisms).

	Trial had to collect binary data	This fulfilled the criteria that the simulation study had to collect differing phase endpoints, specifically binary endpoints in phase II and time-to-event endpoints in phase III (key element #4)
	Trial had to report values used in sample size calculation	To identify values for hypothesised control arm and hypothesised experimental arm. These values would be used to determine simulation phase II sample size. Additionally, hypothesised value in the control arm would be used as a historical control for the simulations of the equivalent single-arm trial (key element #3)

Table 8 – Description of requirements from published phase II-phase III pairings to be used for simulation studies

30th August 2022, I searched for trials through clinicaltrial.gov under the condition or disease “cancer” with the following filters:

- Randomized
- Completed studies
- Studies with results
- Interventional studies
- Phase III
- With statistical analysis plans

This search term generated 293 trials. Each of the 20 phase II-phase III pairings which progressed onto the second stage of screening was given a ‘pairing number’ from #1 - #20, detailed in Figure 19.

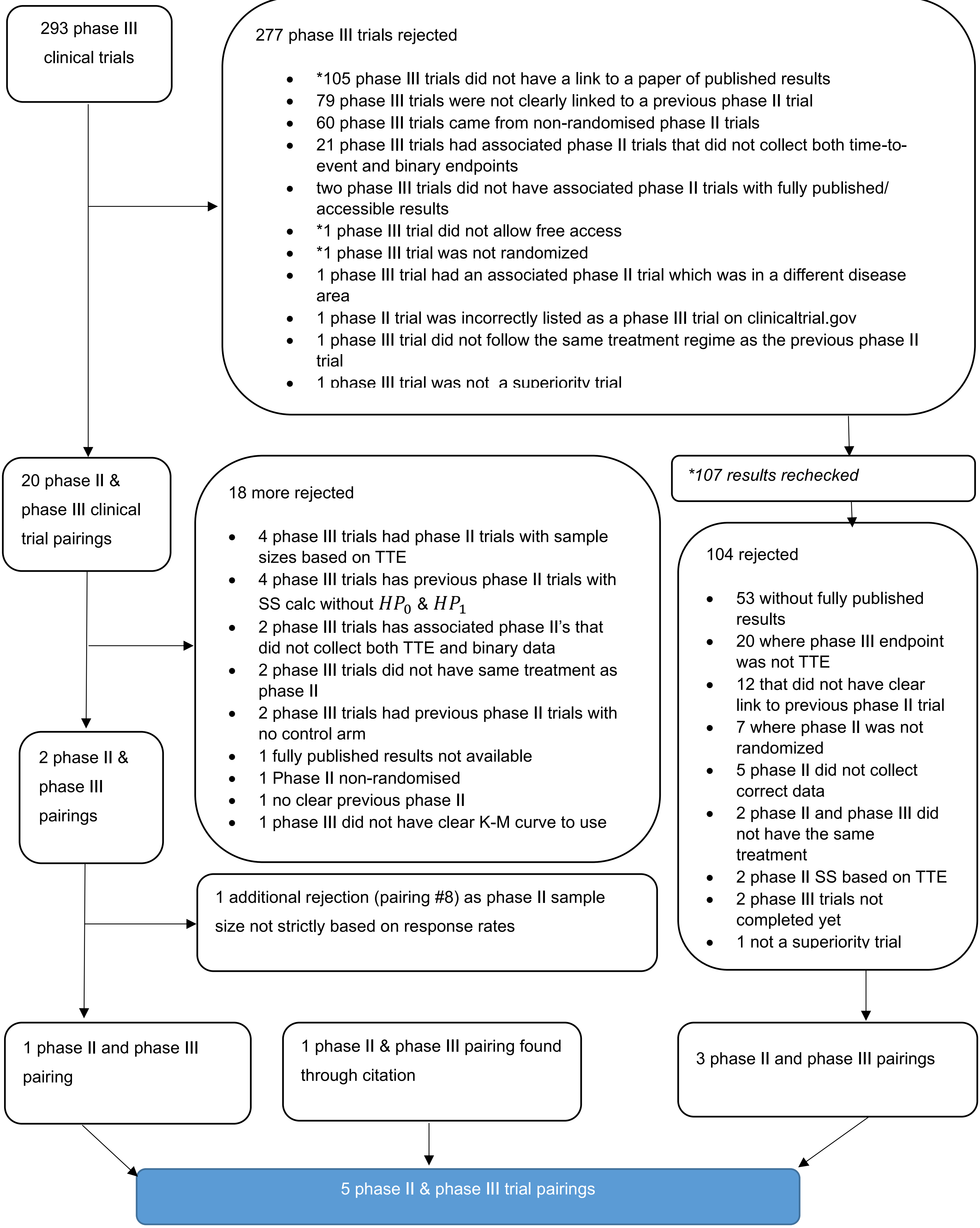
However, as the requirements for the phase II and phase III trials were so specific, initially only one phase II-phase III pairing, pairing #3, was found to be suitable for the simulations. Therefore 107 of the originally rejected phase III trials were rechecked for suitability:

- 105 phase III trials that did not have their final trial reports linked on their clinicaltrial.gov pages were searched again via google scholar.
- One phase III trial had a final report that did not allow access through UCL. Access was attempted through contacting the authors.
- One phase III trial was incorrectly identified as non-randomised.

After rechecking, three more phase II-phase III pairings were identified and were each assigned a pairing number of #21, #22 and #23. One other phase II-phase III trial pairing was found through citations of another trial report and was given the pairing number #24. In total, five phase II-phase III trial pairings were found for the simulation studies.

A PRISMA diagram is presented in Figure 19 to represent the literature review process of selecting papers (99).

A summary of each of the five phase II and phase III pairings is found in Table 9.



293 phase III clinical trials

277 phase III trials rejected

- *105 phase III trials did not have a link to a paper of published results
- 79 phase III trials were not clearly linked to a previous phase II trial
- 60 phase III trials came from non-randomised phase II trials
- 21 phase III trials had associated phase II trials that did not collect both time-to-event and binary endpoints
- two phase III trials did not have associated phase II trials with fully published/ accessible results
- *1 phase III trial did not allow free access
- *1 phase III trial was not randomized
- 1 phase III trial had an associated phase II trial which was in a different disease area
- 1 phase II trial was incorrectly listed as a phase III trial on clinicaltrial.gov
- 1 phase III trial did not follow the same treatment regime as the previous phase II trial
- 1 phase III trial was not a superiority trial

107 results rechecked

20 phase II & phase III clinical trial pairings

18 more rejected

- 4 phase III trials had phase II trials with sample sizes based on TTE
- 4 phase III trials has previous phase II trials with SS calc without HP_0 & HP_1
- 2 phase III trials has associated phase II's that did not collect both TTE and binary data
- 2 phase III trials did not have same treatment as phase II
- 2 phase III trials had previous phase II trials with no control arm
- 1 fully published results not available
- 1 Phase II non-randomised
- 1 no clear previous phase II
- 1 phase III did not have clear K-M curve to use

104 rejected

- 53 without fully published results
- 20 where phase III endpoint was not TTE
- 12 that did not have clear link to previous phase II trial
- 7 where phase II was not randomized
- 5 phase II did not collect correct data
- 2 phase II and phase III did not have the same treatment
- 2 phase II SS based on TTE
- 2 phase III trials not completed yet
- 1 not a superiority trial

2 phase II & phase III pairings

1 additional rejection (pairing #8) as phase II sample size not strictly based on response rates

1 phase II and phase III pairing

1 phase II & phase III pairing found through citation

3 phase II and phase III pairings

5 phase II & phase III trial pairings

Figure 19– CONSORT diagram of selecting phase II & phase III pairings for simulation study

Pairing #	Phase	Disease	Date Conducted	Trial treatment arms	NCT identifier	Reported SS	HP_0	HP_1	$\widehat{P}_0$	$\widehat{P}_1$
#3	II (7)	Advanced gastric cancer	Oct 2011 - Dec 2012	Control: S-1+civaplatin Experimental: S-1+leucovorin+oxaliplatin additional arm: S-1+leucovorin	Not available	Total SS: 145 Control arm: 49 Experimental arm: 47 Additional arm: 49	0.5	0.65	0.46	0.66
	III (8)	Advanced gastric cancer	Jan 2015 - Dec 2016	Control: S-1+civaplatin Experimental: S-1+leucovorin+oxaliplatin	NCT 02322593	Total SS: 711 Control arm: 355 Experimental arm: 356	N/A	N/A	N/A	N/A
#21	II (1)	Squamous Cell Carcinoma of the Head and Neck and low or no PD-	Apr 2015 - Mar 2016	Control: durvalumab Experimental: durvalumab & tremelimumab	NCT 02319044	Total SS: 267 Control arm: 67 Experimental arm: 133	0.13	0.27	0.092	0.078

		L1 tumour cell expression		Additional: tremelimumab						
	III (2)	Squamous Cell Carcinoma of the Head and Neck and high PD-L1 tumour cell expression	2015 - 2017	Control: durvalumab Experimental: durvalumab & tremelimumab Additional: SOC	NCT 02551159	Total SS: 1084 Control arm: 465 Experimental: 413 Additional arm: 206	N/A	N/A	N/A	N/A
#22	II (9)	Previously Untreated Locally Advanced or Metastatic Non–Small-Cell Lung Cancer	Not reported	Control: carboplatin/ paclitaxel alone Experimental: carboplatin/ paclitaxel plus 15 mg/kg bevacizumab	Not available	Total SS: 99 Control arm: 32 Experimental arm: 35 Additional arm: 32	0.27	0.52	0.188	0.315

				Additional: carboplatin/ paclitaxel plus 7.5 mg/kg bevacizumab						
	III (10)	Non-small cell lung cancer	2001 - 2004	Control: carboplatin/paclitaxel alone Experimental: carboplatin/paclitaxel plus bevacizumab	NCT 00021060	Total SS: 878 Control arm: 444 Experimental arm: 434	N/A	N/A	N/A	N/A
#23	II (3)	Advanced melanoma	Sep 2013 - Feb 2014	Control: ipilimumab 3 mg/kg plus placebo Experimental: ipilimumab 3 mg/kg plus nivolumab 1 mg/kg	NCT 01927419	Total SS: 142 Control arm: 47 Experimental arm: 95	0.1	0.4	0.11	0.56
	III (4)	Advanced melanoma	Jul 2013 - Mar 2014	Control: nivolumab plus placebo	NCT 01844505	Total SS: 945 Control arm: 316 Experimental arm: 314	N/A	N/A	N/A	N/A

				Experimental: nivolumab plus ipilimumab Additional: ipilimumab plus placebo		Additional arm: 315				
#24	II (5)	Metastatic Colorectal Cancer	Aug 2006 - Mar 2007	Control: FOLFOX-4 Experimental: cetuximab plus FOLFOX-4	Not available	Total SS: 344 Control arm: 168 Experimental arm: 169	0.5	0.7	0.36	0.46
	III (6)	RAS Wild-Type Metastatic Colorectal Cancer	Sept 2010 - Jan 2016	Control: FOLFOX-4 Experimental: cetuximab plus FOLFOX-5	NCT 01228734	Total SS: 393 Control arm: 200 Experimental arm: 193	N/A	N/A	N/A	N/A

Table 9 – Table to describe each of five selected phase II-phase III pairings

These five pairings provide a range of phase II and phase III development plan examples that can be seen in real life:

- Pairing #3 provides an example of a real-life situation where hypothesized P_0 and P_1 are estimated with a high level of accuracy.
- Pairing #21 provides a null example, where the phase III $\widehat{P_0}$ and $\widehat{P_1}$ are very similar values.
- Pairing #22 provides an example of when hypothesized values are over-estimated in both the control and experimental arm (positive historical control error).
- Pairing #23 provides an example of when hypothesized values are too modest, and the estimated treatment effect is larger than anticipated (mild negative historical control error).
- Pairing #24 provides an example of when the hypothesized values are too optimistic, and real treatment effect is smaller than anticipated.

It should be noted that while the literature review allowed for either phase IIA or phase IIB trials to be found, none of the five development plans distinguished between the two, and instead, the general label of "phase II" was used for each one. Therefore, like previous chapters, simulated phase II trials can be generalised to represent the overall phase II process.

Conducting a simulation study using data from these real-life phase II & phase III trial pairings will help inform guidance on what phase II design is most likely to lead to correct conclusions on treatment, given similar trial parameters. Details of the simulation study methods are explained in the next section.

5.2.2 Simulation Studies

I will discuss the methods of the simulation studies, and how the five phase II-phase III pairings are incorporated.

Two development plans were simulated given the parameters seen in each of the five phase II-phase III published pairings:

- A single-arm phase II trial followed by a randomized phase III trial.

- A randomised phase II trial followed by a randomized phase III trial.

Similar to previous chapters, for simplicity, only single-stage designs were considered and contained only one experimental arm. Additionally, as previously mentioned there was no distinction between either phase IIA or phase IIB trials, but rather they encompassed the whole phase II process. These made up five simulation studies, each one comparing the performance of two development plans. Furthermore, I have continued to assume no missing data.

Overall, the aim of the simulation studies was to assess the methodology of phase II trial designs by addressing the five key elements. Specifically, how would the performance compare of a development plan with a randomised phase II trial, if it had used a single-arm phase II trial instead?

The first key element, accounting for both phase II and phase III success rates, was addressed by considering the performance of the whole phase II-phase III development plan. This included assessing the development plan sample size, and the proportion of times the development plan made a correct decision on treatment.

The second key element, considering both the alternative and null hypothesis, was assessed by the parameters extracted from the five phase II-phase III trial pairings. Four simulation studies were conducted under the alternative hypothesis, as four pairings concluded in favour of the experimental treatment in their published phase III trial reports (pairings #3, #22, #23 and #24). The simulation study that used parameters from pairing #21 was considered to be conducted under the null hypothesis, as its published phase III trial report did not conclude in favour of the experimental arm and had an estimated hazard ratio of 1.0.

All simulated pairings had some degree of historical control error when run as a development plan with a single-arm phase II trial. The hypothesized P_0 was determined by the control value used in the published phase II sample size calculation, and the true P_0 was determined by the published phase II trial estimate. No hypothesised values and estimated values within the same trial were the same. This accounted for the third key element, considering historical control error.

The fourth key element accounted for the differing endpoints between phase II and phase III trials. For each development plan simulated, the phase II trials generated binary data, with parameters extracted from the published phase II trials. The simulated phase III trials resampled time-to-event endpoints using patient outcomes from the published phase III data.

Finally, the fifth element is correspondence between phase II and phase III treatment effect. This was difficult to implement as there is wide debate surrounding the correlation between phase II response rates and phase III survival (69-71, 96-98, 100-102). Therefore, I did not attempt to define a ‘translation’ between phase II binary response rates and phase III time-to-event survival endpoints that can be applied to all development plans. Instead, I compared the time-to-event endpoints collected in both published phase II and phase III trials, to demonstrate five empirical examples of correspondence seen in real-life.

Table 10 describes all the data that was involved in the simulation studies and defines the notation used in this chapter. Some parameters of the *data generating mechanisms* were extracted from the published trials (1-10). Others depend on the simulated data within that repetition to reflect the decision-making that would happen following a phase II trial. Therefore, there are two data sources used in the simulation; derived from published trials and derived from simulations. Table 10 is split into two: the orange section with circles describes data extracted from the published trials, and the purple section with diamonds describes the data that were calculated within a simulation study run itself.

Reasoning behind the choices of each *data generating mechanism* is detailed in Section 5.2.2.1 ADEMP(S) – Data Generating Mechanisms.

Data type	Notation	Definition	Used for
Data derived from published trials	HP_0	Hypothesised proportion of responders in the control arm used to calculate sample size in the published phase II trial	For simulated randomised phase II trials - Used to calculate sample size For simulated single-arm phase II trials - used to calculate sample size, used as historical control, and used to represent control proportion when calculating $sim\ RR$
	HP_1	Hypothesised proportion of responders in the experimental arm used to calculate sample size in the published phase II trial	Used as the hypothesised experimental proportion for the simulated phase II sample size
	$\widehat{P_0}$	Estimate of proportion of responders in the control arm from the published phase II trial	For simulated randomised phase II trials: used to represent the “truth” for proportion of responders in the control arm
	$\widehat{P_1}$	Estimate of proportion of responders in the experimental arm from the published phase II trial	Used to represent the “truth” for proportion of responders in the experimental arm
	PII_t	Length of published phase II trial	Used as a reference for proportion of survival at PII_t months. This will be used as a reference to extrapolate the hypothesised survival proportions at the end of the subsequent phase III trial - used to calculate simulated phase III sample size
	$PIII_t$	Length of published phase III trial	Used as a reference for proportion of survival at $PIII_t$ months. This will be used as a target to extrapolate the

			hypothesised survival proportions at the end of the subsequent phase III trial - used to calculate simulated phase III sample size
	PII_S_0	Estimate of survival proportion in the control arm at the end of the published phase II trial	For simulated phase III trials following a randomised trial: Used to extrapolate $HPIII\_S_0$, hypothesised proportion of survivors in the control arm at the end of the subsequent phase III trial
	PII_S_1	Estimate of proportion of survivors in the experimental arm at the end of the published phase II trial	For simulated phase III trials following a single-arm trials: Used to extrapolate $HPIII\_S_1$, hypothesised proportion of survivors in the experimental arm at the end of the subsequent phase III trial
	$HPIII\_S_0$	Hypothesised survival proportion in the control arm at the end of the subsequent phase III trial. For simulated phase III trials following a randomised trial: $HPIII\_S_0$ is based on extrapolations of PII_S_0 by fitting an exponential survival model. For simulated phase III trials following a single-arm trial: See alternative $HPIII\_S_0$ definition below	For simulated phase III trials following a randomised trial Used to represent the control arm in the logrank test to calculate simulated phase III sample size. $HPIII\_S_0$ is also multiplied by $sim\ RR$ to produce an estimate of $HPIII\_S_1$ For simulated phase III trials following a single-arm trial: See alternative $HPIII\_S_0$ definition below
	$HPIII\_S_1$	Hypothesised survival proportion in the experimental arm at the end of the subsequent phase III trial.	For simulated phase III trials following a single-arm trial: used to represent the control arm in the logrank test to calculate simulated phase III sample

		For simulated phase III trials following a single-arm trials: $HPIII\_S_1$ is based on extrapolations of PII_S_1 by fitting an exponential survival model. For simulated phase III trials following a randomised trial: See alternative $HPIII\_S_1$ definition below	size. $HPIII\_S_1$ is also multiplied $\left(\frac{1}{sim\ RR}\right)$ to produce an estimate of $HPIII\_S_0$ For simulated phase III trials following a randomised trial: See alternative $HPIII\_S_1$ definition below
	PIII patient - level survival	Patient-level data derived from digitizing published Kaplan-Meier curves from the published phase III trial	Used to generate outcomes of the simulated phase III trial using sample with replacement
Data derived from simulated trials	$sim\ PII\ SS$	Simulated phase II sample size determined by a likelihood ratio sample size calculation using HP_0 and HP_1 with one-sided α=0.15, and $(1-\beta)$=0.8	Used to set the number of observations for a simulated phase II trial
	$sim\ RR$	For simulated phase II randomised trials: estimate of relative risk from the simulated phase II trial. For simulated phase II single-arm trials: this is the estimate of the experimental response rate from the simulated trial divided by HP_0	For simulated phase III trials that follow a randomised trial: $sim\ RR$ is multiplied with $HPIII\_S_0$ to produce an estimate of $HPIII\_S_1$ For simulated phase III trials that follow a single-arm trial: $\left(\frac{1}{sim\ RR}\right)$ is multiplied with $HPIII\_S_1$ to produce an estimate of $HPIII\_S_0$
	$HPIII\_S_0$	Hypothesised survival proportion in the control arm at the end of the subsequent phase III trial. For simulated phase III trials	For simulated phase III trials that follow a single-arm trial: this is used to represent the hypothesised survival

		following a randomised trial: $HPIII\_S_0$ is based on $\left(\frac{1}{sim\ RR}\right)$ multiplied by $HPIII\_S_1$	proportion in the control arm in the logrank test to calculate sample size
	$HPIII\_S_1$	Hypothesised survival proportion in the experimental arm at the end of the subsequent phase III trial. For simulated phase III trials following a single-arm trials: $HPIII\_S_1$ is based on $sim\ RR$ multiplied by $HPIII\_S_0$	For simulated phase III trials that follow a randomised trial: this is used to represent hypothesised survival proportion in the experimental arm in the logrank test to calculate sample size
	$sim\ PIII\ SS$	Simulated phase III sample size calculated by logrank test with one-sided α=0.025 and $(1-\beta)$=0.9	Used to determine the number of set the number of outcomes to be sampled from PIII patient-level survival
	PIII survival estimates	Predicted survival estimates using a Weibull model from the published phase III trial	Used to generate additional PIII patient-level data if $sim\ PIII\ SS$ is greater than the number of observations in PIII patient-level survival
	Sim PIII patient - level survival	Additional simulated phase III patient-level data	If $sim\ PIII\ SS$ is greater than the number of observations in the PIII patient-level survival, this is used to provide more observations to sample from (with replacement)

Table 10 – Table to describe sources of data involved in Chapter 5 simulation studies depicting the phase II-phase II development plans.

How these data were implemented in the simulation studies can now be seen in the flow diagrams, represented in Figure 20 and Figure 21. Figure 20 describes the simulations of development plans with a randomised phase II trial, and Figure 21 describes simulations of development plans with a single-arm phase II trial. In these flow diagrams, data that were derived from published pairings are represented by a white circle (seen as orange in Table 10), and data that are derived from the simulations themselves are represented by a white diamond (seen as purple in Table 10).

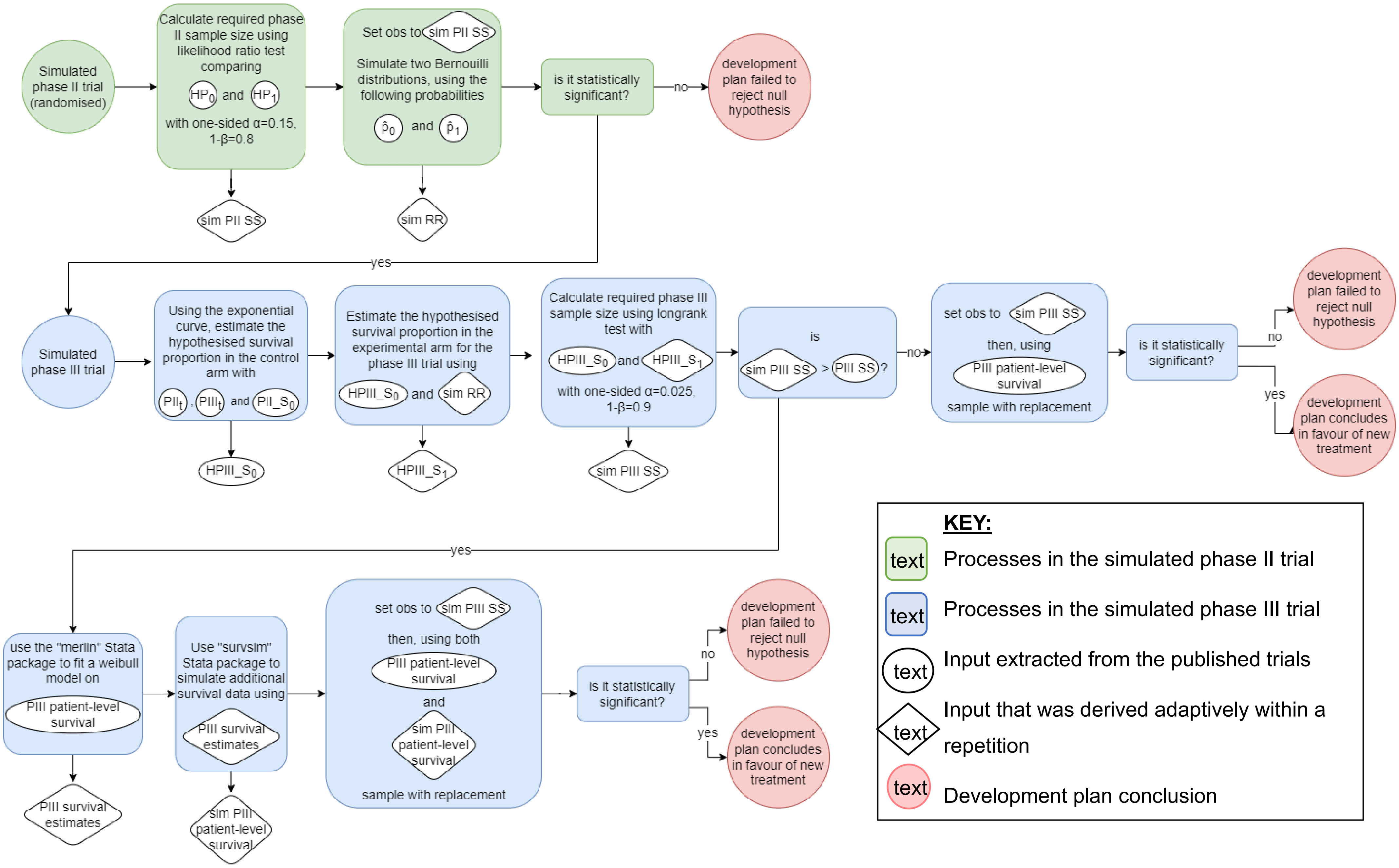

Simulated phase II trial (randomised)
Calculate required phase II sample size using likelihood ratio test comparing
HP_0 and HP_1
with one-sided α=0.15, 1-β=0.8
sim PII SS
Set obs to sim PII SS
Simulate two Bernouilli distributions, using the following probabilities
$\hat{p}_0$ and $\hat{p}_1$
sim RR
is it statistically significant?
no
development plan failed to reject null hypothesis
yes
Simulated phase III trial
Using the exponential curve, estimate the hypothesised survival proportion in the control arm with
PII_t , $PIII_t$ and PII_S_0
$HPIII\_S_0$
Estimate the hypothesised survival proportion in the experimental arm for the phase III trial using
$HPIII\_S_0$ and sim RR
$HPIII\_S_1$
Calculate required phase III sample size using longrank test with
$HPIII\_S_0$ and $HPIII\_S_1$
with one-sided α=0.025, 1-β=0.9
sim PIII SS
is
sim PIII SS > PIII SS ?
no
set obs to sim PIII SS
then, using
PIII patient-level survival
sample with replacement
is it statistically significant?
no
development plan failed to reject null hypothesis
yes
development plan concludes in favour of new treatment
yes
use the "merlin" Stata package to fit a weibull model on
PIII patient-level survival
PIII survival estimates
Use "survsim" Stata package to simulate additional survival data using
PIII survival estimates
sim PIII patient-level survival
set obs to sim PIII SS
then, using both
PIII patient-level survival
and
sim PIII patient-level survival
sample with replacement
is it statistically significant?
no
development plan failed to reject null hypothesis
yes
development plan concludes in favour of new treatment
KEY:
text Processes in the simulated phase II trial
text Processes in the simulated phase III trial
text Input extracted from the published trials
text Input that was derived adaptively within a repetition
text Development plan conclusion

Figure 20- Flow diagram to depict the simulation study for a phase II-phase III development plan with a randomised phase II trial

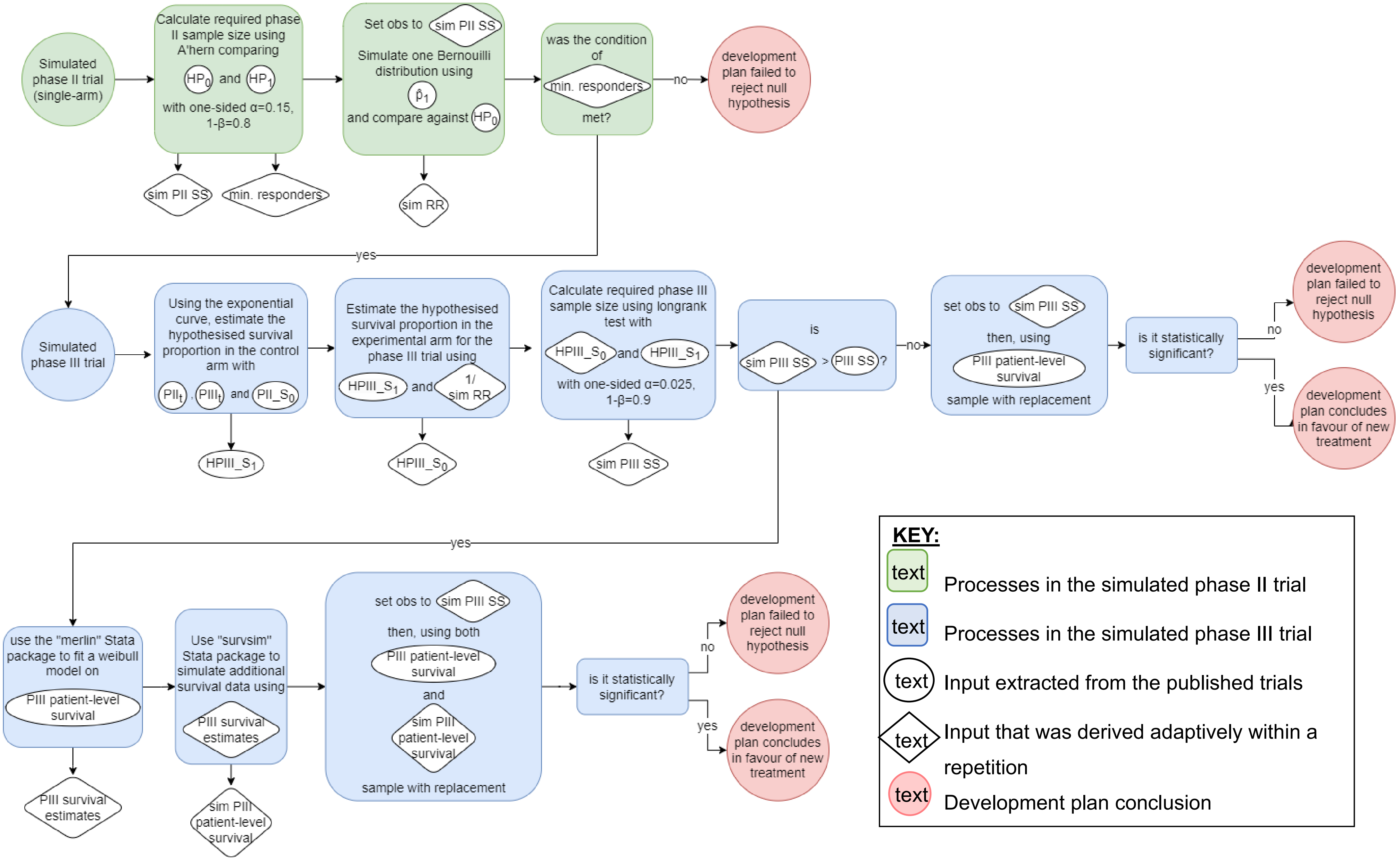

Simulated phase II trial (single-arm)
Calculate required phase II sample size using A'hern comparing
HP$_0$ and HP$_1$
with one-sided α=0.15, 1-β=0.8
sim PII SS
min. responders
Set obs to sim PII SS
Simulate one Bernouilli distribution using
$\hat{p}_1$
and compare against HP$_0$
sim RR
was the condition of min. responders met?
no
development plan failed to reject null hypothesis
yes
Simulated phase III trial
Using the exponential curve, estimate the hypothesised survival proportion in the control arm with
PII$_t$, PIII$_t$ and PII_S$_0$
HPIII_S$_1$
Estimate the hypothesised survival proportion in the experimental arm for the phase III trial using
HPIII_S$_1$ and 1/sim RR
HPIII_S$_0$
Calculate required phase III sample size using longrank test with
HPIII_S$_0$ and HPIII_S$_1$
with one-sided α=0.025, 1-β=0.9
sim PIII SS
is sim PIII SS > PIII SS?
no
set obs to sim PIII SS
then, using
PIII patient-level survival
sample with replacement
is it statistically significant?
no
development plan failed to reject null hypothesis
yes
development plan concludes in favour of new treatment
yes
use the "merlin" Stata package to fit a weibull model on
PIII patient-level survival
PIII survival estimates
Use "survsim" Stata package to simulate additional survival data using
PIII survival estimates
sim PIII patient-level survival
set obs to sim PIII SS
then, using both
PIII patient-level survival
and
sim PIII patient-level survival
sample with replacement
is it statistically significant?
no
development plan failed to reject null hypothesis
yes
development plan concludes in favour of new treatment
KEY:
text Processes in the simulated phase II trial
text Processes in the simulated phase III trial
text Input extracted from the published trials
text Input that was derived adaptively within a repetition
text Development plan conclusion

Figure 21 - Flow diagram to depict the simulation study for a phase II-phase III development plan with a single-arm phase II trial

The rest of the methods section will describe the simulation plan for the five simulation studies using the ADEMP(s) structure by Morris et al which describes the Aims, Data generating mechanisms, Estimands, Methods of analysis, Performance measure and Simulation sample size (79).

5.2.2.1 ADEMP(S)

Aims

The aim of the simulation studies was to investigate the impact of phase II trial design choice on decisions made over the whole development plan based on five examples of real-life settings.

Data Generating Mechanisms

As a reminder, a *data generating mechanism* refers to the models and set of parameter values used to simulate a dataset. Parameters extracted from the five published phase II-phase III trial pairings informed the settings for scenarios: a simulated development plan with a single-arm phase II trial, and a simulated development plan with a randomised phase II trial. This gave five total simulation studies.

I will first describe the *data generating mechanisms* used in each development plan in the setting of pairing #3 in detail. Specifically, a description of a simulated development plan with a randomised phase II trial will be detailed first, followed by a simulated development plan with a single-arm phase II trial. It should be noted that *data generating mechanisms* described will be ***highlighted in bold and italicised***.

The *data generating mechanisms* of pairing #21, #22, #23 and #24 will be described with less detail. The inputs are different but the way they are handled is the same.

Table 11 below describes all data extracted from trial pairing #3 for each of the two development plans simulated.

	HP_0	HP_1	$\widehat{P_0}$	$\widehat{P_1}$	PII_t (months)	$PIII_t$ (months)	PII_S_0	$HPIII\_S_0$	PII_S_1	$HPIII\_S_1$
randomised phase II trial	0.5	0.65	0.46	0.66	27	36	0.1204	0.059	N/A	N/A
single-arm phase II trial	0.5	0.65	N/A	0.66	27	36	N/A	N/A	0.2462	0.1543

Table 11 – Table of data that are extracted from pairing #3

Hypothetical randomised phase II trial followed by phase III RCT

First, the phase II simulation sample size was calculated. For this, hypothesised values of control response rate and experimental response rate, HP_0 and HP_1, were extracted from the sample size calculation of the published phase II trial. Values of HP_0 and HP_1 were inputted into a likelihood ratio sample size calculation with a one-sided α=0.15 and $1-\beta$=0.8. Because this uses information from the published study only, the sample size does not change over repetitions. If any published phase II trial used unequal allocation between treatment groups, this was also reflected in the simulated phase II sample size calculation. In the setting of pairing #3, $\boldsymbol{HP_0}$***=0.5*** and $\boldsymbol{HP_1}$***=0.65*** which produced a phase II simulation sample size of 152 (76 in each arm as the phase II trial from pairing #3 used equal allocation).

Next, the outcomes for the arms in the simulated phase II trial were generated. True response rates for the simulated phase II trial were taken from the published phase II post-trial estimates, i.e. $\hat{P}_0$ and $\hat{P}_1$. In pairing #3, published estimates of control and experimental response rates in the phase II trial were $\boldsymbol{\hat{P}_0}$***=0.46*** and $\boldsymbol{\hat{P}_1}$***=0.66***. Therefore, for the simulated randomised phase II trial, two Bernoulli distributions were simulated with 76 observations each: $Y \sim Bernoulli\ (0.46)$ in the control arm and $Y \sim Bernoulli\ (0.66)$ in the experimental arm. A likelihood ratio test was used to determine statistical significance, with the null hypothesis written as $H_0\colon P_1 \leq P_0$, and the alternate hypothesis written as $H_1\colon P_1 > P_0$.

It was then determined if the experimental treatment in the simulated phase II trial should continue to phase III. For this, a phase II trial had to fulfil two criteria: 1) the result of the simulated phase II is statistically significant with a one-sided p-value<0.15, and 2) the result is clinically meaningful, which I operationalised as observed simulation relative risk ($sim\ RR$) >1.25. This value based from the GRADE recommendations of the optimal information size for relative risk (103). To assess the first criteria, a chi-squared test was performed. For illustrative purposes, suppose that the simulated randomised phase II trial generated 30/76 responders in the control arm, and 46/76 responders in the experimental arm. In this instance, a chi-squared test would produce a one-sided p-value of 0.009, fulfilling the first criterion. This would also produce an associated $\boldsymbol{sim\ RR}$ ***= 1.53***, fulfilling the second criterion. In this example, the simulated phase II randomised trial would continue to a simulated phase III. It is important to note that the $\boldsymbol{sim\ RR}$ ***= 1.53*** would be used to aid the calculation of simulated phase III sample size.

There were multiple steps to determine the sample size for the simulated phase III trial. Hypothesised values for the survival proportions in each arm needed to be determined, relative to the length of the phase III trial. For simulated development plans with randomised phase II trials, the hypothesised value of the phase III control survival proportion, $HPIII\_S_0$, was determined first. For this, the Kaplan-Meier curve from the published phase II trial was digitized. This allowed me to obtain a proxy of patient-level survival data to estimate the published phase II survival proportions for both the control and experimental arms. Digitizing the Kaplan-Meier curves was done using the software “DigitizeIt” (104). For pairing #3, after digitization, the control survival proportion from the published phase II paper was found to be $\boldsymbol{PII\_S_0}$= ***0.1204.*** However, the published phase II trial was 27 months long ($\boldsymbol{PII_t = 27}$). As the subsequent published phase III trial was 36 months long, ($\boldsymbol{PIII_t = 36}$), the phase II published survival proportions needed to be extrapolated to suit the longer trial. For the purpose of sample size calculations, this was done by assuming exponentially distributed survival times, fitting an exponential survival model as described in Chapter 3. Through the exponential survival model the control arm hazard, λ_0, was obtained, to extrapolate the control survival proportion for a 36-month trial. For pairing #3, $\boldsymbol{HPIII\_S_0}$ ***was found to be equal to 0.059.***

The hypothesised phase III experimental survival proportion, $HPIII\_S_1$, was calculated by multiplying hypothesised phase III control survival proportion by the simulated phase II relative risk, i.e., $HPIII\_S_1$ = ($HPIII\_S_0$ * $sim\ RR$). This was done for two reasons: 1) In superiority trials, investigators often focus on how much more efficient the experimental arm is from the perspective of the control, rather than separate estimates of survival proportions for each arm, 2) This provided an adaptive element to the methods to allow simulated phase II trials to influence simulated phase III sample size. This reflects choices that investigators make, adapting the design of a phase III trial based on preceding phase II results. Although simulated relative risk is not equivalent to hazard ratio, it represents how much better the experimental arm is from control. As the main point of interest is usually how much better a treatment is than the standard of care, it was determined that generating $HPIII\_S_1$ from $HPIII\_S_0$ and $sim\ RR$ would best reflect real-life practice. Therefore, in this example, hypothesised phase III experimental survival would be equal to $\boldsymbol{HPIII\_S_1}$***= (0.059*1.53)=0.090***.

The two hypothesised phase III survival proportions would be used in a logrank sample size calculation. If any published phase III trial used unequal allocation this was reflected in the simulated phase III sample size. With our illustrative example using pairing #3, a logrank sample size calculation would be used to compare survival proportions of 0.059 and 0.090, with equal allocation, a one-sided α=0.025 and a $1-\beta$=0.9. This gave a simulation phase III sample size of $\boldsymbol{sim\ PIII\ SS}$ ***= 1748*** (874 in each arm).

In chapter 4, simulated phase III sample size determined the number of generated data from a simple exponential survival model. However, in this chapter, data are sampled from a Digitized dataset to allow baseline hazard function and time-varying effects to be flexible, giving more realism than the outcomes generated previously. Therefore, in these simulations, simulated phase III sample size determined the number of observations that would be sampled with replacement from the published phase III digitized dataset, also obtained using “DigitizeIt”. This method also allowed imperfect correspondence to be collected as an output of the simulation, as opposed to defining it ourselves. Therefore, the difference in the survival proportions from the published phase II trial and the survival observations sampled from the published phase III trial would represent the imperfect correspondence between the phases in a simulated development plan.

After simulated phase III sample size was calculated, the simulated phase III sample size could be less than or equal to the number of observations within the published phase III digitized dataset, or greater. The latter needed consideration in terms of how to generate the data. In our illustrative example, simulated phase III sample size is 1748, and the number of observations in the published phase III digitized dataset for pairing #3 is only 681. It should be noted that observations in the digitized phase III dataset is different to the published phase III sample size, 711, as the digitized dataset is not an exact replication of the data, rather a close approximation of the numerical data. I will first describe the scenario where simulated phase III sample size is less than or equal to the observations in the published phase III digitized dataset, before describing the extra process when the simulation phase III sample size is larger than the published trial.

In the instance where simulated phase III sample size was less than or equal to the number of observations in the published phase III digitized dataset, outcomes would be sampled with replacement using the `bsample` Stata command. Sampling with replacement was chosen so no parametric assumptions were made about the distribution of survival times in the digitized dataset by using the observed survival censoring time distribution. This method allows for non-proportional hazards should they exist in the digitized dataset. The number of outcomes sampled was determined by the simulated phase III sample size. Once sampled, a logrank test was performed comparing the sampled outcomes of the control and experimental arm. Formally the null hypothesis is written as $H_0: HR \geq 1$ and the alternate hypothesis is written as $H_1: HR < 1$. If the one-sided p-value was less than the pre-determined threshold of 0.025 then it was considered that the simulated development plan concluded in favour of the new treatment. If not, the simulated development plan failed to reject the null hypothesis, and therefore failed to conclude in favour of the treatment. Additionally, simulated phase III hazard ratio was taken from a Cox model.

I will now discuss what happens within the simulation when the simulated phase III sample size is greater than the number of observations in the published phase III digitized dataset.

We saw in our illustrative example for pairing #3, that the simulated phase III sample size of 1748 (874 per arm) was greater than the number of observations in the phase III digitized dataset, 681. In instances like these, it would be inappropriate to keep resampling from the digitized datasets. This is because it would not properly propagate uncertainty as the required sample size is greater than the number of observations in the digitized dataset.

Therefore, I need to model the phase III digitized dataset to simulate additional datapoints. Although the Cox model is used in the simulation study to estimate a hazard ratio, it would be a poor model choice to simulate more data. The main reason for this is that the baseline hazard function is not estimated and is instead treated as a nuisance parameter. This makes it difficult to simulate data from. Therefore, I needed an alternative model to predict additional survival data from.

The Royston–Parmar model is a flexible parametric model for censored survival data. It builds on other less flexible parametric models, such as loglogistic or Weibull models, and augments their hazard functions via splines to create flexible models to fit onto data and to estimate baseline hazards smoothly (105).

To describe this, I will first explain how to fit a Weibull model for the log-cumulative hazard function, $\ln H(t)$, with proportional hazards. μ and ρ define the Weibull distribution with 'characteristic life' μ and shape parameter ρ (where σ = ρ^{-1}).

$$\ln H(t) = \ln\left[\left(\frac{t}{\mu}\right)^{p}\right] = p\ln t - p\ln \mu = \frac{\ln t - \ln \mu}{\sigma} = \gamma_0 + \gamma_1 \ln t$$

Where:

$$\gamma_0 = \frac{-(\ln \mu)}{\sigma}$$

$$\gamma_1 = \frac{1}{\sigma}$$

Let $s(\ln t;\, \gamma)$ represent some non-linear function of $\ln t$ and having an adjustable parameter γ. To approximate $s(\ln t;\, \gamma)$ by natural cubic splines, this can be written as:

$$s(\ln t\,; \gamma) = \gamma_0 + \gamma_1 \ln t + \gamma_2 v_1(\ln t) + \cdots + \gamma_{m+1} v_m(\ln t)$$

Where:

$v(\ln t)$ is a ‘basis function’ of x which fits splines between the internal knots.

m defines the number of internal knots.

j then represents the j^{th} basis function such that $j = 1, \ldots, m$ and:

$$v_j(\ln t) = (\ln t - k_j)_+{}^3 - \lambda_j(\ln t - k_{min})_+{}^3 - (1 - \lambda_j)(\ln t - k_{max})_+{}^3$$

Where:

Boundary knots are defined as k_{min} and k_{max}. It follows that $k_1 > k_{min}$ and $k_m < k_{max}$.

λ_j determines the knot for the j^{th} basis function and is defined as:

$$\lambda_j = \frac{k_{max} - k_j}{k_{max} - k_{min}}$$

Finally, to fit a natural cubic spline model onto a cumulative hazard function, we start with the cumulative hazard function which is:

$$\ln H(t; \boldsymbol{z}) = \gamma_0 + \boldsymbol{\gamma}^T \boldsymbol{v}(\ln t) + \boldsymbol{\beta}^T \boldsymbol{z}$$

Where:

$\boldsymbol{\beta}$ is a vector of parameters to be estimated for covariates $\boldsymbol{z}$

$\boldsymbol{v}(\ln t)$ is the matrix of the j number of $v(\ln t)$ basis functions

This is a *proportional* hazards model because $\boldsymbol{\beta}$ is additive on the log-cumulative hazard scale; that is, it is not a function of $\ln t$. The difference between applying proportional hazards and non-proportional hazards can be seen below when describing the j^{th} component of γ as:

$$\gamma_j = \begin{cases} \gamma_{j0}, & l = 0 \qquad \text{(Proportional hazards model)} \\ \gamma_{j0} + \sum \gamma_j z, & l \geq 1 \qquad \text{(Non-proportional hazards model for } z_1, \dots, z_k\text{)} \end{cases}$$

Where:

l is the dimension of $\boldsymbol{z}$. In this instance, the dimension of $\boldsymbol{z}$ is 1 as the only covariate is a dummy variable to determine treatment (0=control, 1=experimental)

The Royston–Parmar model can be easily extended to include time-dependent covariate effects i.e., non-proportional hazards, by incorporating the natural cubic splines. The cubic spline function allows for easy estimation of the baseline hazard, which makes it simpler to predict data, and additionally allows an easy extension of time-dependant covariates i.e., non-proportional hazards. This is because different estimates of the hazards can be calculated between each knot. The complexity of the curve is determined by the number of knots chosen, where degrees of freedom are equal to $(k_{max} - 1)$.

I use the Stata package, '`merlin`', to fit Royston–Parmar models to the existing published data (106). For degrees of freedom, Royston *et al.* chose three degrees of freedom for baseline log-cumulative hazard and two for non-proportional hazards (107). However, the paper concluded that more could be chosen as it allows for more flexibility without much additional computation cost. Therefore, I chose four degrees of freedom for both. This equates to three internal knots being used to fit the model; two representing the extremes of the data, and then the remaining three represent the 25%, 50% and 75% quartiles.

Another package, `survsim`, is then used to simulate survival data directly from the fitted Royston–Parmar model, with a maximum trial time of 36 months, mirroring the published data (108). Sampling with replacement using '`bsample`' was drawn from the digitized dataset to prioritize outcomes from the published phase III trial. Then, the additional Royston–Parmar generated data was appended to the digitized dataset to account for the remaining sample that the simulated phase III trial required. The parametrically simulated data are a ‘second best’ compared with the resampled data, but given the flexibility afforded by Royston–Parmar models, they should provide a close (smooth) approximation to the empirical survival distribution.

Figure 22 demonstrates and example of Royston–Parmar model fitting the digitized survival data using *data generating mechanisms* described in the illustrative example, for one particular repetition. The Kaplan-Meier curve displays the survival outcomes from four groups: pairing #3 digitized phase III control arm, pairing #3 digitized phase III experimental arm, Royston–Parmar-generated control arm and Royston–Parmar-generated experimental arm. Each group is represented by control (blue), experimental (red), RP-control (green) and RP-experimental (yellow) respectively.

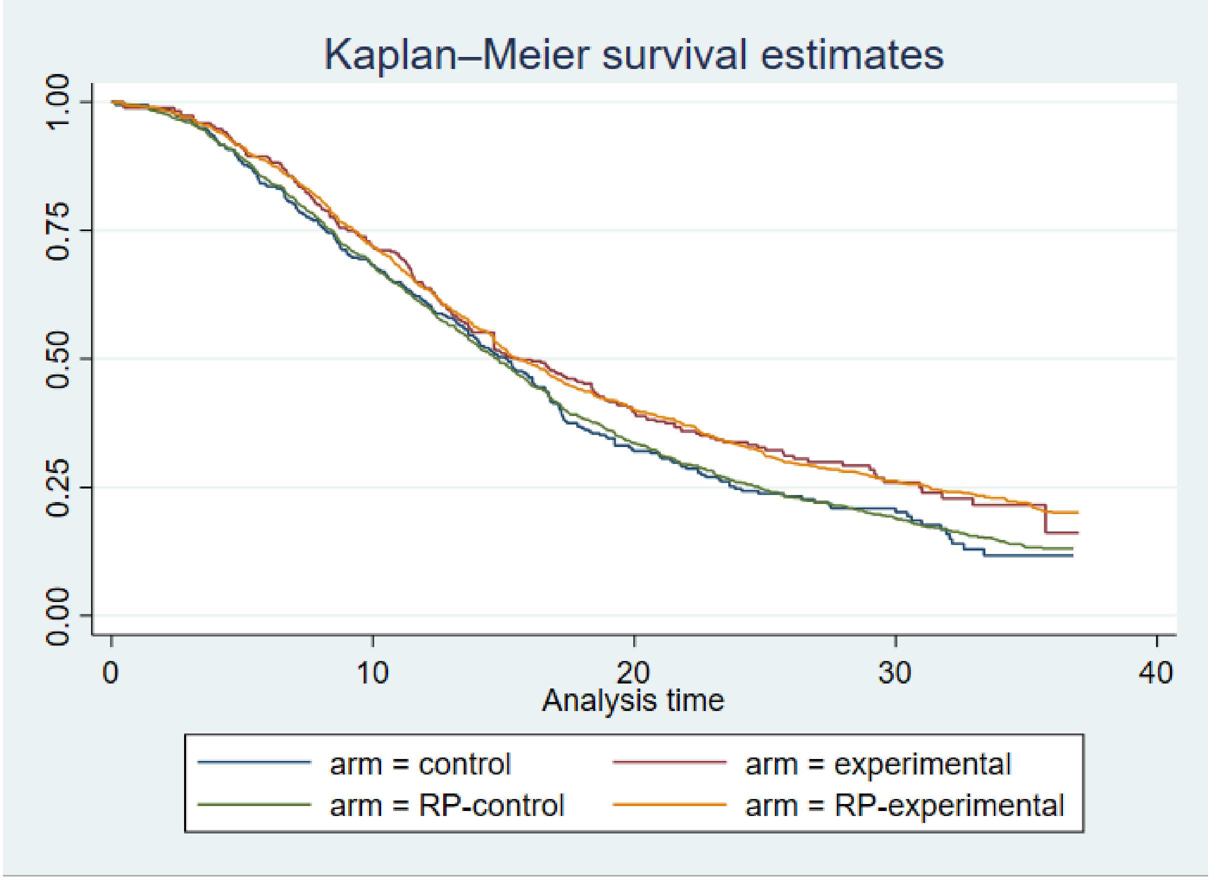


Figure 22 – Kaplan-Meier curve to compare survival outcomes from phase III digitized dataset of pairing #3 with Royston–Parmar generated survival outcomes.

As described above, a logrank test would then be performed to determine whether the phase III trial concluded in favour of the treatment.

It should be noted that the analysis performed on simulated phase III trials, logrank test and hazard ratio from a Cox model, are valid to use in conjunction with each other. Even though the Cox model and the logrank test calculate test-statistics differently, the hazard ratio calculated in the logrank test would be very close to identical to the hazard ratio calculated in the Cox-model. Additionally, a logrank test is appropriate to use on all simulated phase III trials as it is a non-parametric test, therefore performs well if non-proportional hazards exist within the phase III sampled outcomes.

Each simulated development plan consisted of 10000 repetitions, using the same simulation sample size calculation as Chapter 3 to minimise Monte Carlo error (79). The number of times the development plan correctly concluded in favour of treatment was recorded.

Single-arm phase II trial followed by phase III RCT

To begin simulating published phase II-phase III pairing #3 as if it had used a single-arm phase II trial, simulated phase II sample size was calculated. The same values of HP_0 and HP_1 were extracted from the published phase II trial as previously described. However, HP_0 and HP_1 were used in a sample size calculation using the A'hern method with a one-sided α and a $1-\beta$ level of as 0.15 and 0.8 respectively. Not only did the A'hern method provide the sample size, it also provided the minimum number of responders needed in the experimental arm to conclude in favour of treatment. It should be noted than the A'hern method was not simple using Stata, but a user-written A'hern package was found for the statistical software, R (109). Therefore, all simulated phase II single-arm sample sizes were calculated using R software. As seen in the pairing #3 setting, HP_0***=0.5*** and HP_1***=0.65***. With these values the A'hern method gave a sample size of 42, with a threshold of 25 responders needed to conclude in favour of treatment.

To generate the data for the simulated phase II single-arm trial, data was drawn from one Bernoulli distribution which represented outcomes from the experimental arm, and therefore used $\boldsymbol{\widehat{P}_1}$***=0.66*** extracted from the published phase II trial. As seen previously, the Bernoulli distribution simulated for the experimental arm was $Y\sim Bernoulli(0.66)$.

To allow the single-arm phase II trial to continue to phase III, the minimum number of responders in the experimental arm needed to be achieved. Otherwise, the simulated phase II trial failed to reject the null hypothesis, and the simulated development plan stopped here. For illustrative purposes, let's assume that 26 responders were generated in the experimental arm. In this case, as 26 is greater than the minimum threshold of 25, the simulated phase II single-arm trial would continue to phase III.

Next, the simulated phase III sample size was calculated. We have seen previously that the hypothesised survival proportion in the experimental arm, $HPIII\_S_1$, was calculated by multiplying the hypothesised survival proportion in the control arm, $HPIII_0$, and simulated phase II relative risk, $sim\ RR$. To derive $HPIII\_S_0$, published phase II survival proportion in the control arm, PII_S_0, was extrapolated along the exponential survival curve to suit the length of the subsequent simulated phase III trial. However, development plans that conduct a single-arm phase II trial do not have access to control survival proportions. Therefore, published phase II experimental survival proportions, PII_S_1, were used to derive hypothesised phase III survival proportions in the experimental arm $HPIII\_S_1$.

However, a value for $HPIII\_S_0$ was still missing, and an adaptive element was needed to allow simulated phase II trials to influence subsequent simulated phase III sample size. To replicate the decisions of phase II single-arm investigators, $HPIII\_S_0$ was based on the best information available: estimation of simulated phase II experimental response rate, hypothesised phase III survival proportions in the experimental arm ($HPIII\_S_1$) and historical control (HP_0). First, a pseudo-relative risk was obtained by dividing the estimate of the simulated experimental response rate by HP_0. Subsequently, $HPIII\_S_0$ was calculated by multiplying $HPIII\_S_1$ with the inverse of phase II simulated relative risk, $\frac{1}{sim\ RR}$.

Therefore, the published phase II experimental survival proportion for a 27-month trial for pairing #3 was extracted, which was $\boldsymbol{PII\_S_1}$**= 0.2462**. To suit a longer 36-

month phase III trial, the estimation is then extrapolated. The hypothesised phase III experimental survival proportion for a 36-month trial then becomes ***$HPIII\_S_1$=0.1543.***

The phase II pseudo-relative risk was calculated by dividing the proportion of responders in the experimental arm by the simulated historical control, HP_0. With our illustrative example, we assumed there were 26/42 responders in the simulated single-arm phase II trial in the setting of pairing #3. The proportion of responders is therefore 60.5%. As HP_0=0.5, this would mean the pseudo-relative risk for the simulated phase II single-arm trial would be (0.605/0.5)=1.21, therefore ***$sim\ RR$ = 1.21***.

To obtain the hypothesised phase III control survival proportion, the hypothesised phase III experimental survival proportion was multiplied by the inverse of the simulated phase II pseudo-relative risk. In the illustrative example, the inverse of the simulated phase II pseudo-relative risk is 1/1.21 = 0.826. Then, this inverse value is multiplied by the hypothesised phase III experimental survival proportion, i.e., ***$HPIII\_S_0$= 0.1543*0.826 = 0.1275.***

Simulated phase III sample size was then based on a logrank sample size calculation using the values of the hypothesised phase III survival proportions with a one-sided α=0.025 and $1-\beta$=0.9. In this instance, hypothesised survival values of 0.1275 and 0.1543 gives a required simulation phase III sample size of ***$sim\ PIII\ SS$ = 5186*** (2593 in each arm). The simulated phase III sample size determined the number of observations that needed to be sampled from the phase III digitized dataset.

As with a simulated development plan with a randomised phase II trial, when simulated phase III sample size was less than or equal to the number of observations in the digitized phase III dataset, I sampled with replacement to obtain simulated phase III outcomes. When simulated phase III sample size was greater than the number of observations in the digitized phase III dataset, I fitted a Royston–Parmar model with 4 degrees of freedom for both baseline log-cumulative hazard and two for non-proportional hazards to simulate extra data after resampling 681 datapoints from the digitized phase III dataset.

Finally, a logrank test was conducted to determine if the simulated development plan concluded in favour of the new treatment, and a Cox-model was performed to obtain a hazard ratio.

This simulated development plan consisted of 10000 repetitions. The number of times the development plan correctly concluded in favour of treatment was recorded.

The two simulated development plans, one with a randomised phase II trial and one with a single-arm phase II trial, were repeated for each of the parameters taken from the four remaining pairings (#21, #22, #23 and #24). It should be noted that, for pairing #21, as it represents the null hypothesis, the proportion of times the 10000 repetitions of each simulated development plan correctly failed to reject the null hypothesis was recorded. This was either at phase II or phase III within the development plan.

Table 12 describes the parameters taken from published pairings that were used for data generation.

Pairing name	phase	HP_0	HP_1	$\widehat{P_0}$	$\widehat{P_1}$	phase length (months)	S_0	S_1	$sim\ SS$
#3	II	0.5	0.65	0.46	0.66	27	0.12	0.246	phase II randomised: 152 (76 in each) phase II single-arm: 42 (25 responders needed)
	III	N/A	N/A	N/A	N/A	36	0.059	0.154	*Based on data within simulation iteration*
#21	II	0.13	0.27	0.092	0.078	17	0.283	0.323	phase II randomised: 129 (43 in the control arm, 86 in the experimental arm) phase II single-arm: 28 (6 responders needed)
	III	N/A	N/A	N/A	N/A	48	0.028	0.41	*Based on data within simulation iteration*
#22	II	0.27	0.52	0.188	0.315	42	0.188	0.036	phase II randomised: 54 (27 in each) phase II single-arm: 16 (7 responders needed)
	III	N/A	N/A	N/A	N/A	42	0.188	0.036	*Based on data within simulation iteration*
#23	II	0.1	0.4	0.11	0.56	27	0.486	0.598	phase II randomised: 33 (11 in the control arm and 22 in the experimental arm)

									phase II single-arm: 7 (2 responders needed)
	III	N/A	N/A	N/A	N/A	45	0.3	0.424	*Based on data within simulation iteration*
#24	II	0.5	0.7	0.36	0.046	10	0.221	0.309	phase II randomised: 84 (42 in each) phase II single-arm: 22 (14 responders needed)
	III	N/A	N/A	N/A	N/A	33	0.007	0.021	*Based on data within simulation iteration*

Table 12 – Table of data extracted from each published phase II-phase III pairing. S_0 and S_1 denote survival proportions

Targets

The targets of this simulation study are the proportion of correct conclusions made about treatment at the end of each type of development plan.

Method of analysis

The target in my simulations is hypothesis rejection at the end of the simulated development plan – either at phase II or phase III.

For a simulated randomised phase II trial, analysis was a likelihood ratio test with one-sided α and $1-\beta$=0.8. Additionally, simulated phase II relative risk was compared against a pre-determined threshold of 1.25 to determine clinical meaningfulness. For a simulated single-arm phase II trial, analysis was comparing the number of responders in the simulated experimental arm to the threshold given by the A'hern method. For simulated phase III trials, analysis performed was a logrank test with a one-sided α=0.025 and $1-\beta$=0.9.

Table 13 provides further details of my methods of analysis for each phase, for each development plan simulated.

Development plan	Phase II analysis	Phase III analysis
Phase II single-arm x Phase III RCT	Compare the number of responders in the trial with minimum number of responders needed to conclude in favour of treatment assuming A'hern method. This calculation determines the minimum number of responders needed where the 85% confidence interval does not include hypothesised P_0.	Compare the two survival curves for each treatment group which have been sampled from the digitized dataset from the published phase III trial for the given pairing. A logrank test is performed with one-sided p-value < 0.025.
Phase II RCT x Phase III RCT	Compare two observed proportions using a likelihood ratio test that assumes $Y_i \sim Binomial$ with one-sided p-value < 0.15.	

Table 13 – Method of analysis for the two development plans

Performance Measures

There are two performance measures in which I assessed each of the 10 development plans (two development plans for each of the five pairings). One of them is the proportion of times correct decision has been made on treatment, and the other is total development plan sample size.

Proportion of times a correct decision is made on treatment

In chapter 3, true positive proportion and true negative proportion was used, which depended on whether the simulation occurred under the alternative or null hypothesis respectively.

True positive proportion was the proportion of times a development plan concluded in favour of truly effective treatment i.e., if a development plan was run where there was a true treatment effect, and in 7762/10000 repetitions the development plan rejected the null hypothesis at phase III, then the true positive proportion would be 77.62%.

True negative proportion was the proportion of times a development plan correctly failed to reject the null hypothesis at either phase II or phase III i.e., if a development plan was run where the control and experimental arm were truly equivalent, and in 9988/10000 repetitions the development plan failed to reject the null hypothesis at either phase II or phase III, then the true negative proportion would be 99.88%

Of the five published phase II-phase III pairings found, one is under the null hypothesis, whereby the control arm is exactly as effective as the experimental arm and the given published phase III HR=1.00 (pairing #21).

To compare the performance of all five pairings against each other, the performance measure of "proportion of times correct decision is made on treatment" was used. This means true positive proportion was collected for pairings #3, #22, #23 and #24, and true negative proportion for pairing #21.

Total development plan expected sample size

The total development plan sample size was also obtained.

If a development plan stops at phase II, the total development plan sample size is the simulated phase II sample size. If a development plan continues to phase III, the total development plan sample size is simulated phase II sample size plus simulated phase III sample size.

For each of the 10 development plans, the mean simulated development plan sample size across 10000 repetitions was calculated using the following formula:

$$mean\ development\ plan\ sample\ size$$

$$=$$

$$\begin{pmatrix} proportion\ of \\ development\ plans \\ that\ stopped\ at \\ phase\ II \\ * \\ simulated\ phase\ II \\ sample\ size \end{pmatrix} + \begin{pmatrix} proportion\ of \\ development\ plans \\ that\ reached \\ phase\ III \\ * \\ mean\ simulated\ phase\ III \\ sample\ size \end{pmatrix}$$

Simulation Sample Size

Simulation sample size is 10000 repetitions for each of the 10 development plans to ensure Monte Carlo standard error remains below the desired level of 0.5%. This is the same simulation sample size as calculated in Chapter 3.

All analysis was conducted using Stata version 16.0, except in two instances: 1) Digitizing Kaplan-Meier curves was performed using DigitizeIt software and 2) single-arm sample size calculation was performed with a user-written package “Ahern” in R version 4.2.3. The code of this package is included in the appendix. Example code of the simulation study for pairing #3 is provided in the appendix, simulating a development plan with a randomised phase II trial.

5.3 Results

5.3.1 Data Extracted From Published Trials.

First, I will present the data extracted from the five phase II-phase III trial pairings.

Figure 23, Figure 24, Figure 25, Figure 26 and Figure 27 provide published Kaplan-Meier curves extracted from phase II and phase III trials from each pairing. These were the Kaplan-Meier curves used to digitize each dataset.

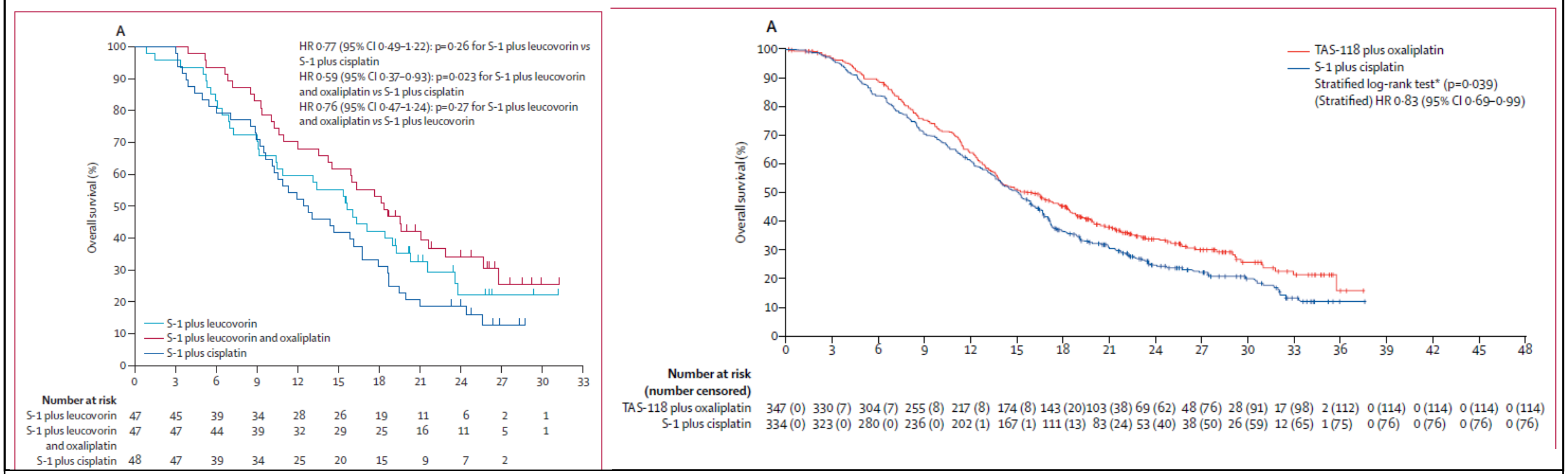
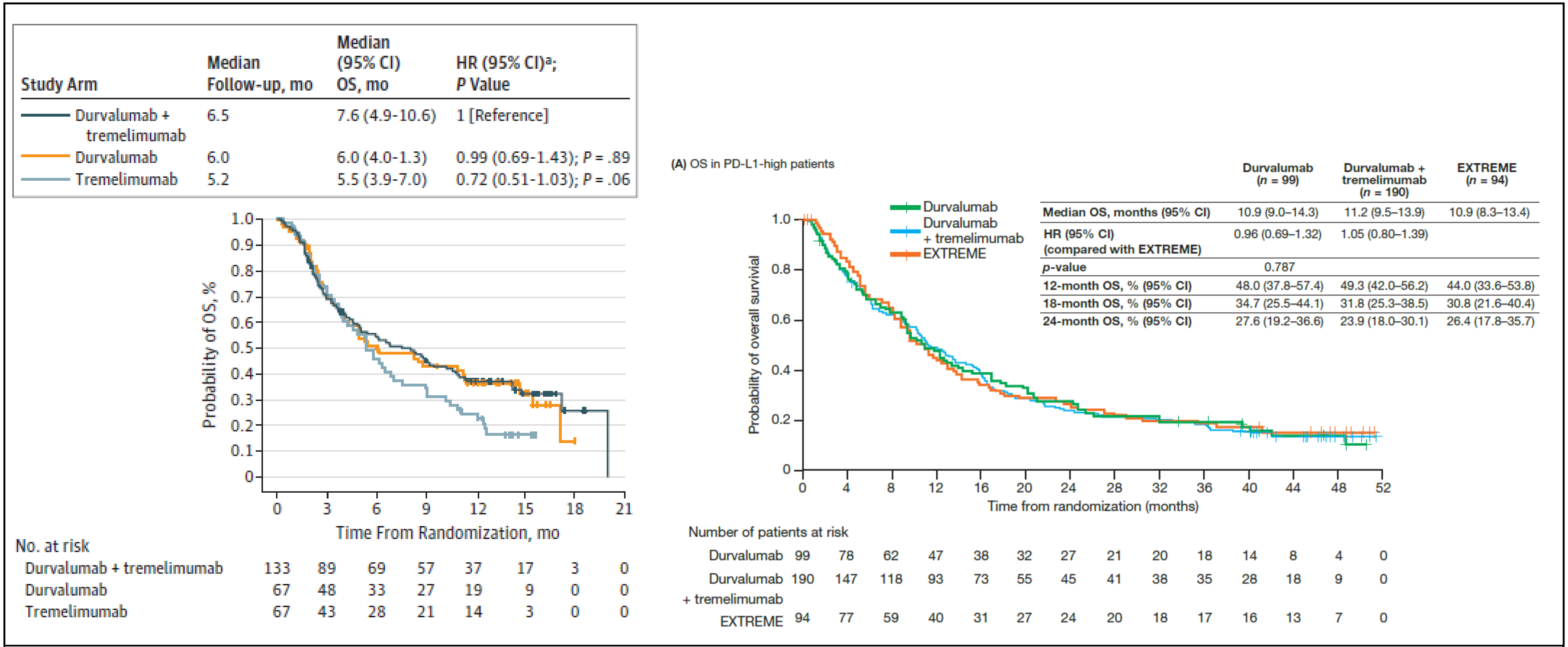


Figure 23 - Published phase II and phase III overall survival curves for pairing #3.

Control and experimental arms used for simulations are S-1 plus cisplatin and S-1 plus leucovorin and oxaliplatin (also known as TAS-118 plus oxaliplatin) respectively. (7, 8)



Study Arm	Median Follow-up, mo	Median (95% CI) OS, mo	HR (95% CI)[a]; P Value
Durvalumab + tremelimumab	6.5	7.6 (4.9-10.6)	1 [Reference]
Durvalumab	6.0	6.0 (4.0-1.3)	0.99 (0.69-1.43); *P* = .89
Tremelimumab	5.2	5.5 (3.9-7.0)	0.72 (0.51-1.03); *P* = .06

No. at risk

	0	3	6	9	12	15	18	21
Durvalumab + tremelimumab	133	89	69	57	37	17	3	0
Durvalumab	67	48	33	27	19	9	0	0
Tremelimumab	67	43	28	21	14	3	0	0

	Durvalumab (*n* = 99)	Durvalumab + tremelimumab (*n* = 190)	EXTREME (*n* = 94)
Median OS, months (95% CI)	10.9 (9.0–14.3)	11.2 (9.5–13.9)	10.9 (8.3–13.4)
HR (95% CI) (compared with EXTREME)	0.96 (0.69–1.32)	1.05 (0.80–1.39)	
***p*-value**	0.787		
12-month OS, % (95% CI)	48.0 (37.8–57.4)	49.3 (42.0–56.2)	44.0 (33.6–53.8)
18-month OS, % (95% CI)	34.7 (25.5–44.1)	31.8 (25.3–38.5)	30.8 (21.6–40.4)
24-month OS, % (95% CI)	27.6 (19.2–36.6)	23.9 (18.0–30.1)	26.4 (17.8–35.7)

Number of patients at risk

	0	4	8	12	16	20	24	28	32	36	40	44	48	52
Durvalumab	99	78	62	47	38	32	27	21	20	18	14	8	4	0
Durvalumab + tremelimumab	190	147	118	93	73	55	45	41	38	35	28	18	9	0
EXTREME	94	77	59	40	31	27	24	20	18	17	16	13	7	0

Figure 24 - Published phase II and phase III overall survival curves for pairing #21.

Control and experimental arms used for simulations are durvalumab and durvalumab + tremelimumab respectively. (1, 2)

]

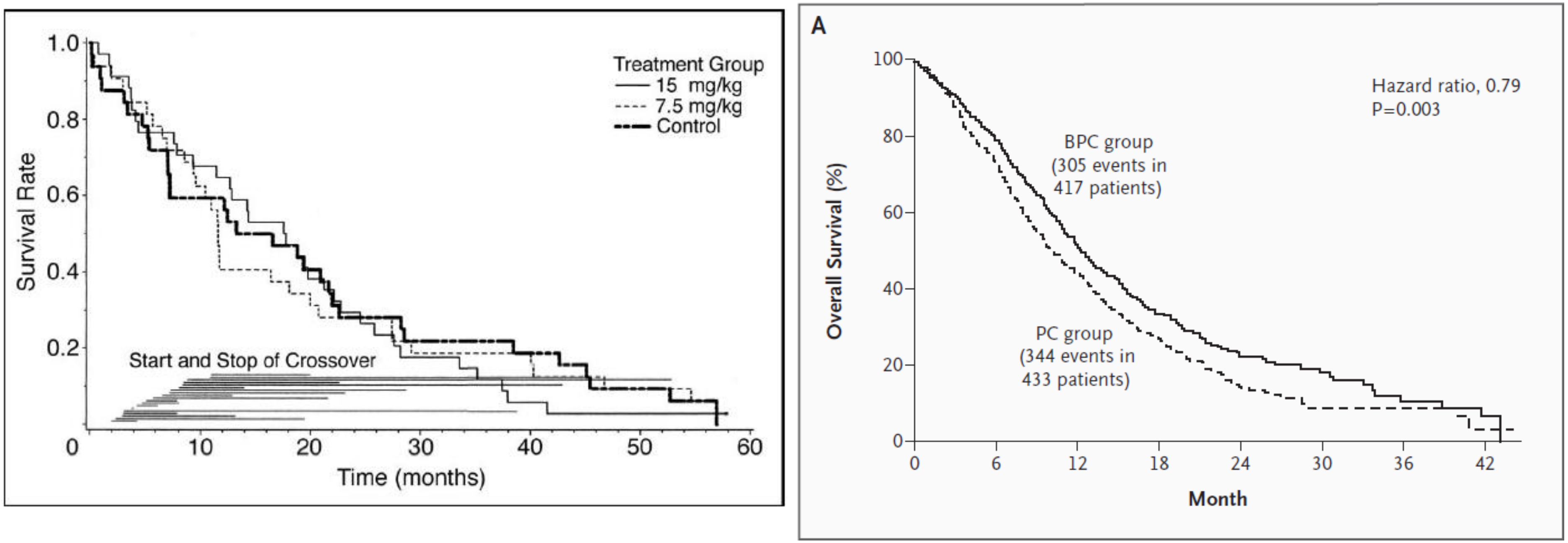
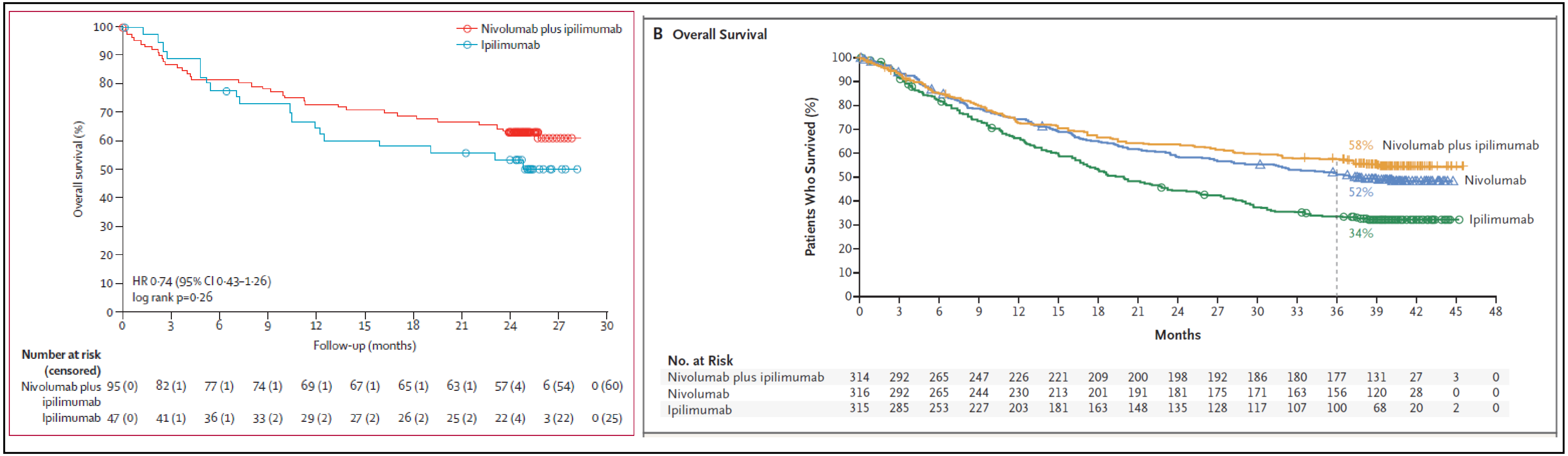


Figure 25 - Published phase II and phase III overall survival curves for pairing #22.

Control arm used for simulations are carboplatin and paclitaxel (Control or PC group) and experimental arm used for simulations are paclitaxel, carboplatin and bevacizumab (15 mg/kg or BPC group) respectively. (9, 10)



Number at risk (censored)											
Nivolumab plus ipilimumab	95 (0)	82 (1)	77 (1)	74 (1)	69 (1)	67 (1)	65 (1)	63 (1)	57 (4)	6 (54)	0 (60)
Ipilimumab	47 (0)	41 (1)	36 (1)	33 (2)	29 (2)	27 (2)	26 (2)	25 (2)	22 (4)	3 (22)	0 (25)

No. at Risk																	
Nivolumab plus ipilimumab	314	292	265	247	226	221	209	200	198	192	186	180	177	131	27	3	0
Nivolumab	316	292	265	244	230	213	201	191	181	175	171	163	156	120	28	0	0
Ipilimumab	315	285	253	227	203	181	163	148	135	128	117	107	100	68	20	2	0

Figure 26 - Published phase II and phase III overall survival curves for pairing #23.

Control and experimental arms used for simulations are ipilimumab and nivolumab plus ipilimumab respectively. (3, 4)

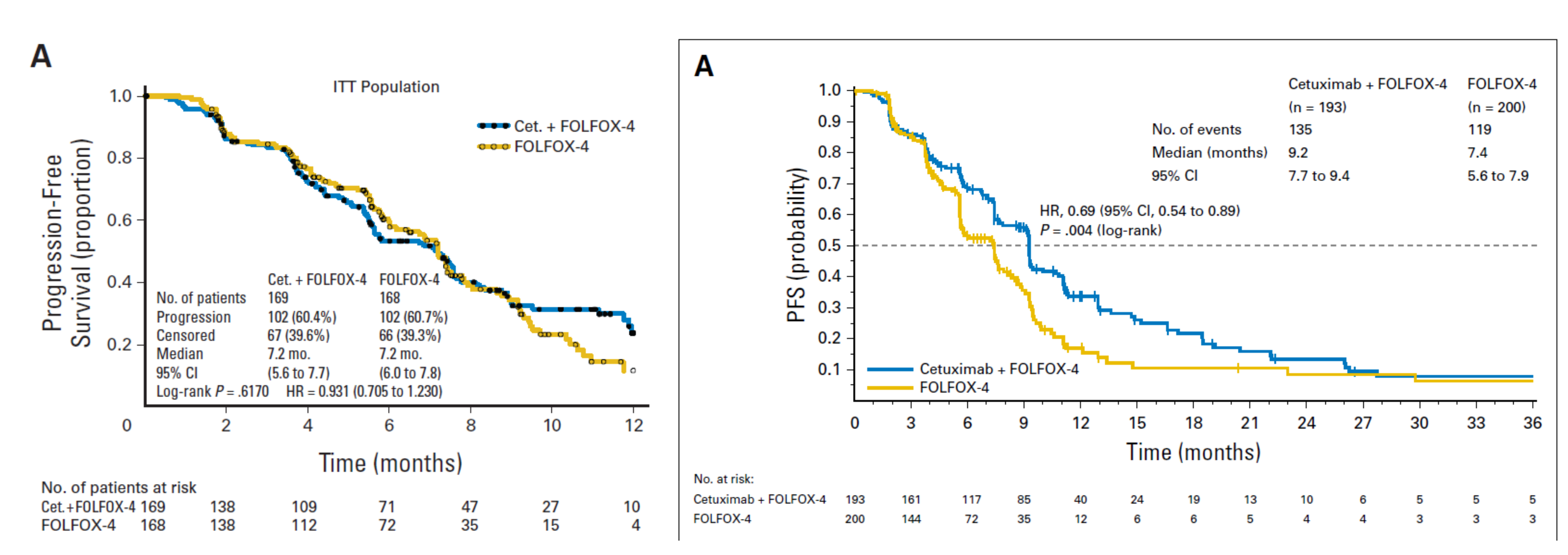


Figure 27 - Published phase II and phase III progression free survival curves for pairing #24.

Control and experimental arms used for simulations are FOLFOX-4 and cetuximab + FOLFOX-4 (also Cet + FOLFOX-4) respectively. (5, 6)

5.3.2 Comparison of Published Estimates, Digitized Datasets and Royston–Parmar Models

Table 14 provides hazard ratios estimates from the published papers and the digitized datasets from each of the five pairings, along with the associated confidence intervals and p-values. Those extracted from the published papers are coloured in blue, those extracted from the digitized dataset are coloured in orange. This is to compare how well the digitized datasets approximated the data from the published papers. Comparisons of Kaplan-Meier curves produced by the published papers and digitized datasets can be found in the appendix (Figure 29, Figure 30, Figure 31, Figure 32, Figure 33, Figure 34, Figure 35, Figure 36, Figure 37, Figure 38)

Additionally, Table 14 provided the imperfect correspondence factor observed from each of the pairings. The calculation for imperfect correspondence is based on the formula established in Chapter 4, i.e., published phase III HR/ published phase II HR. This will give us an idea of the discrepancy in treatment effect between phase II and phase III.

Pairing number	Phase II				Phase III				Imperfect correspondence factor (published phase III HR/ published phase II HR)
	Published phase II HR (95% CI)	Published phase II p-value	Digitized phase II HR (95% CI)	Digitized phase II p-value	Published phase III HR (95% CI)	Published phase III p-value	Digitized phase III HR (95% CI)	Digitized phase III p-value	
Pairing #3	0.59 (95% CI 0.37-0.93)	0.023	0.59 (95% CI 0·37–0·94)	0.028	0.83 (95% CI 0.69-0.99)	0.039	0·82 (95% CI 0·69–0·98)	0·027	1.41
Pairing #21	0.99 (95% CI 0.69-1.43)	0.89	0·96 (95% CI 0·67–1.39)	0·85	Not available	Not available	1.00 (95% CI 0.77-1.3)	0.98	1.01
Pairing #22	Not available	Not available	1.1 (95% CI 0.66-1.83)	0.714	0.79 (95% CI0.67-0.92)	0.003	0.84 (95% CI 0.73-0.96)	0.01	0.95
Pairing #23	0.74 (95% CI 0.43-1.26)	0.26	0.74 (95% CI 0.44-1.24)	0.26	0.55 (95% CI 0.45-0.69)	<0.001	0.57 (95% CI 0.44-0.73)	0.0001	0.74
Pairing #24	Not available	0.66	0.93 (95% CI 0.7-1.23)	0.59	0.69 (95% CI 0.54-0.89)	0.004	0.7 (95% CI 0.55-0.9)	0.006	0.74

Table 14 – Imperfect correspondence factor from each of the published trial pairings, and hazard ratio estimates from both the published papers and digitized datasets from each phase, for each of the trial pairings.

Table 15 provides information on how well the Royston–Parmar models fit the digitized data, and therefore indicates the quality of additional generated survival outcomes. As a reminder, additional survival outcomes were only generated when simulated phase III sample size was greater than the number of observations in the digitized datasets.

Additionally, it should be noted that the number of survival outcomes that needed to be generated depended on the simulated phase II relative risk for that particular repetition. Therefore, the table below demonstrated instances where the Royston–Parmar model generated the same number of survival outcomes than the number of observations in the digitized dataset. The generation of survival outcomes from Royston Royston–Parmar models was repeated 1000 times. The mean hazard-ratio, p-value, 95% CI lower boundary and 95% upper boundary are provided.

Phase III pairing #	Number of observations in digitized dataset	Digitized Hazard Ratio	Digitized p-value	Digitized 95% CI lower boundary	Digitized 95% CI upper boundary	Number of survival outcomes generated by R-P	RP mean Hazard Ratio	RP mean p-value	RP mean 95% CI lower boundary	RP mean 95% CI upper boundary
pairing #3	681	0.82	0.027	0.69	0.98	681	0.8	0.007	0.72	0.9
pairing #21	289	1	0.98	0.77	1.3	289	1.01	0.49	0.84	1.12
pairing #22	878	0.84	0.01	0.73	0.96	878	0.83	0.008	0.75	0.92
pairing #23	629	0.57	0.0001	0.44	0.73	629	0.57	<0.0001	0.49	0.66
pairing #24	393	0.7	0.006	0.55	0.9	393	0.73	0.002	0.63	0.85

Table 15 – Table to compare results of survival outcomes from digitized datasets, and examples generated from Royston–Parmar models

5.3.3 Simulation Results

Table 16 displays the HP_0, HP_1, $\widehat{P_0}$ and $\widehat{P}_1$ extracted from the published phase II trials. Additionally, it displays the simulated sample sizes, and proportion of times a correct decision was made on treatment.

Pairing number	Published phase II HP_0	Published phase II HP_1	Published phase II $\widehat{P_0}$	Published phase II $\widehat{P_1}$	Simulated phase II design	Simulated phase II SS	% simulated dev plans stopped at phase II	Simulated phase III mean SS	% simulated dev plans continued at phase III	Mean of total dev plan SS	% correct decisions made for simulated dev plans
pairing #3	0.50	0.65	0.46	0.66	SAT	42	14.05%	2657	85.95%	2290	42.24%
					RCT	152	7.31%	3144	92.69%	2926	44.99%
pairing #21	0.13	0.27	0.09	0.08	SAT	28	98.24%	1964	1.76%	63	99.96%
					RCT	129	89.51%	419	10.49%	160	99.96%
pairing #22	0.27	0.52	0.19	0.32	SAT	16	7.05%	1905	92.95%	1772	14.35%
					RCT	54	49.19%	167	50.81%	112	33.03%
pairing #23	0.10	0.4	0.11	0.56	SAT	7	3.23%	62	96.77%	61	96.67%
					RCT	33	3.82%	42	96.18%	42	96.10%
pairing #24	0.50	0.7	0.36	0.46	SAT	22	92.37%	9482	7.63%	744	7.63%
					RCT	84	52.79%	5894	47.21%	2827	46.90%

Table 16 – Table of design elements extracted from pairings to use in simulations along with results of simulations

Figure 28 shows the performance of the five trial pairings, with each simulated development plan i.e., with a single-arm phase II trial or a randomised phase II trial. The mean total development plan sample size is along the x-axis, and proportion of times correct decision was made on treatment along the y-axis. Data points in red represent results of development plans which used a randomised phase II trial, and data-points in blue represent results of development plans which used a single-arm phase II trial. Horizontal lines around each data point represent the 95% confidence intervals of the mean development plan sample size, and vertical lines represent the 95% confidence intervals of the simulation Monte Carlo error.

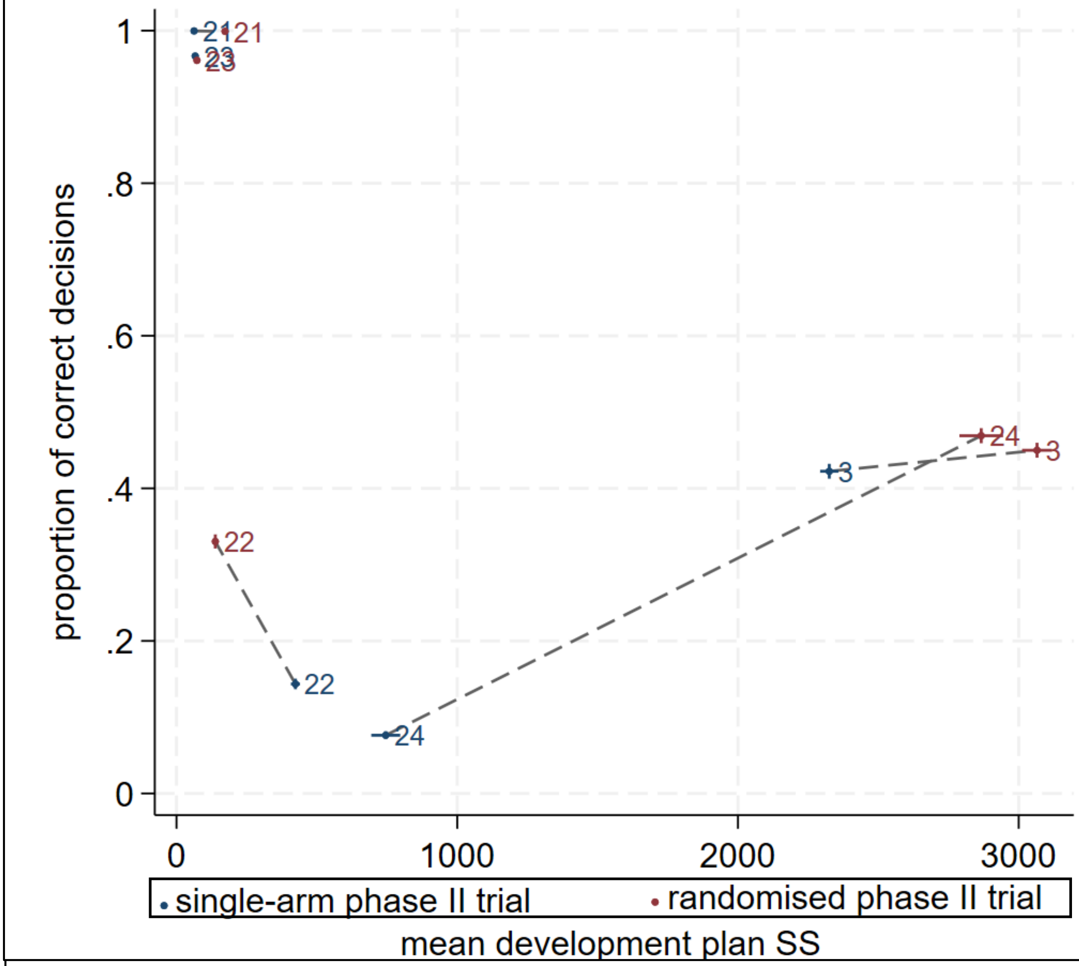


Figure 28 – proportion of correct decisions made against the overall mean development plan sample size.

5.3.3.1 Sample size

Development plans with randomised phase II trials have an overall increased mean SS as a higher % of phase II trials correctly continued into the subsequent phase III trial – for which, required less participants than phase III trials that followed single-arm phase II trials.

Sample sizes across all development plans can be seen in Table 16 and Figure 28. Each simulated phase II trial had a fixed sample size and phase III sample sizes varied for each repetition. The set phase II sample sizes ranged from 7 to 152. It was found that simulated single-arm phase II trials required 70%-79% less participants than simulated randomised phase II trials, comparable to previous literature (67) (64).

Interestingly, in four out of the five pairings, phase III trials that followed randomised phase II trials generated a lower mean sample size (pairing #21, #22, #23 and #24). This can also be seen in Table 16. For pairings #23 and #24, mean simulated phase III sample size that followed a randomised phase II trial was 32-38% less than the sample size of the mean simulated phase III sample size that followed a single-arm trial. In pairings #21 and #22, mean phase III sample sizes for development plans that followed a randomised trial was 79%-91% lower than mean phase III sample size for development plans that followed a single-arm trial.

For pairings #21, #22 and #24, the reduction in mean phase III sample size can be attributed to the randomised phase II trial, as it was able to collect relevant information on the control arm. In these scenarios, the hypothesised P_0 was greater than the simulated truth, $\widehat{P_0}$. This meant the sample size calculation for the subsequent phase III trial could be recalibrated with up-to-date information estimated in the control arm to reflect the larger true treatment effect. However, for the equivalent single-arm phase II trial, this recalibration of the control arm could not be done.

For pairing #23, hypothesised P_0 was close to the observed estimate in the published trial, $\widehat{P_0}$ (0.1 vs 0.11). It should be noted that the 32% reduction in mean phase III sample size equated to an absolute difference of only 20 patients i.e., the mean phase III sample size in development plans that followed a randomised phase II trial was of 42 compared to 62. This could be the result of the small single-arm phase II trial which had a sample size of 7. With so few patients, each additional responder in the experimental arm would have a large impact on the pseudo-relative risk, which would then impact the hypothesised survival estimates for the subsequent phase III sample size calculation. It is therefore possible that there were not enough participants in the single-arm phase II trial to allow for more precise estimates in the pseudo-relative risk, which could have further reduced the mean phase III sample size.

For pairing #3, simulated phase III trials that followed a single-arm phase II trial had a mean sample size that was 15% less than those that followed a randomised phase II trial. At face-value, this seems odd as the hypothesised P_0 is greater than the simulated truth, $\widehat{P_0}$. Similar to pairings #21, #22 and #24, it seems like randomised phase II trials would have the opportunity to recalibrate the phase II control arm to account for the larger than anticipated treatment effect in the subsequent phase III trial. However, the reduction in mean simulated phase III sample size that followed a single-arm phase II trial can be explained by the different processes used to extract hypothesised phase III survival estimates. For example, let's assume both simulated development plans in the setting of pairing #3 calculate a simulated phase II relative risk of 1.5. This would mean a simulated phase III sample size that followed a single-arm trial would use a logrank sample size calculation with hypothesised phase III control survival proportion $HPIII\_S_0$=0.1641 and hypothesised phase III experimental survival proportion $HPIII\_S_1$=0.2462. However, a simulated phase III sample size that followed a randomised trial would use a logrank sample size calculation with $HPIII\_S_0$=0.1204 and $HPIII\_S_1$=0.1806. Although the relative treatment effect is the same between the two calculations, the absolute value is much larger for the simulated phase III sample size that followed a single-arm trial. This is because the published phase II control survival proportion, which is used to extrapolate $HPIII\_S_0$ in development plans with randomised phase II trials, is very close to 0. This contributed to reducing the mean required sample size.

It was found that the mean total development plan sample size was smaller in development plans that used single-arm phase II trials in three out of the five pairings. However, it should be noted that randomised phase II trials more often led to subsequent phase III trials when there was a true difference to be detected at phase III. So, although development plans with single-arm phase II trials had a lower mean development plan sample size overall, this was not necessarily to the benefit of the development plan. Therefore, to fully assess the performance of the two development plans, the proportions of times a correct decision was made on treatment needs addressed.

5.3.3.2 Proportion of Correct Decisions Made and Overall Performance

There are three instances where development plans with either phase II design have similar ability to make the correct decision on treatment. This can be seen in pairing #3, pairing #21 and pairing #23.

In pairing #3, development plans with single-arm phase II trials were able to correctly conclude in favour of treatment 42.2% of the time, and development plans with randomised phase II trials were able to conclude in favour of treatment 45% of the time. This can be seen in Figure 28. Difficulties in detecting treatment effect were due to imperfect correspondence. Table 14 shows that in pairing #3, the published phase II hazard ratio was 0.59 but drifted to 0.83 in the published phase III trial. Given this, the results show that the development plan with a randomised phase II trial had increased ability in detecting true treatment effect in light of imperfect correspondence due to the concurrent control arm used in phase II. The concurrent control arm could detect the larger-than-anticipated treatment effect as hypothesised P_0 was greater than true control rate, $\widehat{P_0}$. This meant it could detect the true treatment effect easier than the single-arm phase II trial, which continued to use the smaller anticipated treatment effect as a benchmark.

Furthermore, the mean total development plan sample size for development plans with randomised phase II trials is 2926, compared with 2290 for development plans with single-arm phase II trials. The question then becomes, is an average of 636 patients worth the increased ability to make a correct decision by 2.8%?

In pairing #21 both development plans performed extremely well, with 99.96% of development plans failing to reject the null hypothesis at either phase II or phase III as it was under the null hypothesis. When looking further into the phase II conclusions, it was clear only 1.8% of single-arm phase II trials unnecessarily continued to phase III, yet 10.5% of randomised phase II trials unnecessarily continued to phase III. However, both are greater than the one-sided type I error threshold of 0.15 used in the sample size calculations. It should also be noted that mean phase III sample size that follows a single-arm trial is 1964, while mean phase III sample size that follows a randomised trial is 419. Yet, because the single-arm phase II trials are so efficient at filtering out futile treatment, the mean of total development plan sample size for those with single-arm phase II trials is just 63 compared to the other development plan with a mean sample size of 160. It could be concluded that a development plan with a single-arm phase II trial is recommended when a trial is conducted under the null hypothesis and there is nearly perfect correspondence between phases (Table 14 shows that phase II HR=0.99, phase III HR=1.00). However, this recommendation is not practical, as no phase II investigator would conduct a trial anticipating the null hypothesis, and level of between-phase-correspondence is difficult to anticipate. However, if investigators do find themselves operating under the null hypothesis, it is useful to know both development plans are able to detect futile treatment well. However, it can be noted that less resources were wasted in funding futile phase III trials when a development plan included a single-arm phase II trial in this study.

In pairing #23, both development plans performed extremely well again, with 96.1%-96.7% of development plans correctly concluding in favour of treatment. Phase II trial success rates are also similar, with 96.2% of randomised phase II trials leading to phase III investigation, and 96.8% of single-arm phase II trials leading to phase III investigation. Success rates are likely to be high for two reasons. The first reason is that the experimental arm performed better than accounted for in the phase II sample size calculations (HP_1=0.4 and $\widehat{P_1}$=0.56 as seen in Table 16). Because this happened in the experimental arm and not the control arm, both single-arm and randomised trials could pick up on this greater treatment effect. The second reason is that a large level of imperfect correspondence was present, which can be seen in Table 14 through the shift in published hazard ratios between each phase (phase II HR=0.74, phase III HR=0.57). Therefore, phase II estimates used to calculate phase

III sample size meant that more participants were recruited than needed, which made it easier for phase III trials to detect the treatment effect. Again, as the proportion of correct decisions made is so similar for each development plan, more weight should be placed on sample size when deciding which development plan is best. As mentioned previously, phase III trials that followed phase II randomised trials produced a smaller mean sample size, which is also reflected in the mean of total development plan sample size. Therefore, in this scenario, the development plan with the randomised phase II trial performs best.

In the remaining two pairings, #22 and #24, the different development plans have very different abilities in being able to make correct conclusions on treatment. It should be noted that, in both instances, hypothesised control proportion and hypothesised experimental proportion proved to be extremely inaccurate for phase II sample size calculations.

In pairing #22, only 33% of development plans with randomised phase II trials correctly concluded in favour of treatment. However, only 14.4% of development plans with single-arm phase II trials were able to conclude in favour of treatment. Table 16 shows that the hypothetical value chosen for the phase II sample size calculation was HP_0=0.27 while the simulated truth was $\widehat{P_0}$= 0.19. Similarly, HP_1=0.52 while $\widehat{P}_1$= 0.32. When looking at the phase II trials specifically, it can also be seen in Table 16 that 93% of single-arm trials did not continue to phase III and, if they did, the mean phase III sample size was 1,905. This is because the single-arm phase II trial was effectively trying to detect a small treatment effect with a hypothesised P_0=0.27 and $\widehat{P}_1$= 0.32, with a sample size of 16. Randomised phase II trials performed better, with 49.2% continuing to a phase III trial. This was due to the benefit of the concurrent control arm, which could attempt to detect a treatment effect between $\widehat{P_0}$ = 0.19 and $\widehat{P}_1$=0.32 with a sample size of 54. Mean phase III sample size that followed a randomised phase II trial was 167.

Pairing #22 also experienced imperfect correspondence, with hazard ratio shifting away from the null between phase II and phase III trials (phase II HR=0.74, phase III HR=0.57). As seen in pairing #23, this can be a benefit to the development plan. However, the inaccurate choices of hypothetical estimates for sample size calculation in phase II mitigate this possible advantage. Although the ultimate recommendation should be to focus efforts on choosing a more representative hypothesised values to calculate phase II sample size, in instances where there not enough information to obtain a reliable estimate a secondary phase I trial could be considered, but ultimately, the trial design at phase II should be randomised.

Finally, pairing #24 presents the biggest difference between development plans in the ability to make correct decisions on treatment. Development plans with randomised phase II trials were able to conclude in favour of treatment 46.9% of the time, and development plans with single-arm randomised trials were able to conclude in favour of treatment 7.6% of the time.

Pairing #24 suffers a similar fate to pairing #22, with inaccurate hypothetical values chosen for phase II sample size calculation. Here, HP_0=0.5 while $\widehat{P_0}$= 0.36 and HP_1=0.7 while $\widehat{P_1}$= 0.46 as seen in Table 16. These values mean that only 7.6% of single-arm phase II trials continue onto phase III, whilst 47.2% of randomised phase II trials continue onto phase III. Again, the poor performance in the single-arm phase II trial is due to the lack of control arm re-estimation, and in this scenario means the single-arm trial is left comparing a historical control which is better than true experimental response rate (HP_0=0.5 vs $\widehat{P_1}$= 0.46). This meant that the mean phase III sample size that followed single-arm phase II trials was calculated as 9482. For phase III trials that followed randomised phase II trials, mean sample size was 5894. The means of the total development plan sample sizes were 744 for those with a single-arm phase II trial, and 2827 for those with a randomised phase II trial. However, this is only because so many single-arm phase II trials failed to detect a treatment effect and is not indicative of sample size efficiency.

Again, like pairing #22, imperfect correspondence was present with hazard ratio shifting away from the null between phase II and phase III trials seen in Table 14 (phase II HR=0.93, phase III HR=0.7). However, any advantage this could have offered either development plan was lost with unrepresentative hypothesised estimates in the phase II sample size calculation. As concluded before, the main benefit to the development plans would be to choose more accurate hypothesised estimates for phase II sample size calculations. However, in the absence of this, a development plan with a randomised phase II trial is recommended in this scenario.

5.4 Discussion

This study improved upon previous studies conducted in chapter 3 and chapter 4 by combining all five key elements identified in the narrative synthesis. This was achieved by using parameters extracted from real-life development plans and sampling survival outcomes from the associated phase III trials.

Five pairings of published phase II and phase III trials within the same development plan were found following a search, which represented a wide variety of clinical trial scenarios. Although this was a smaller number than originally hoped, each of the five development plans represented a wide variety of clinical trial scenarios. This included when a phase II trial had access to accurate representations of true response rates, when the treatment is truly ineffective, when a phase II trial did not have access to accurate representations of true response rates, when the treatment effect is underestimated and when the treatment effect is overestimated. Not only this, there was also a variety of imperfect correspondence seen between phase II and phase III trials. Using these varied pairings as examples of parameters seen in real development plans provide a wealth of results on the impact of phase II trial designs.

Overall, phase II design choices can impact the ability to make correct decisions on treatment at the end of a development plan. Not only is this through the decision on treatment at the end of the phase II trial, but also by influencing the design of the subsequent phase III trial in a constructive way. These simulation studies assessed whether the information available from randomised phase II trials achieves this better than single-arm trials.

It has previously been thought that a randomised trial design can be used when you are unsure of treatment effect as the two concurrent arms are able to estimate the true response rates. This line-of-thinking has been proved correct with pairings #22 and #24, where development plans with randomised phase II trials performed better even though hypothesised values chosen for phase II sample size were inaccurate.

However, one of the advantages of single-arm phase II trials was the reduced sample size required. However, this benefit is commonly traded-off against much larger sample size requirements at phase III. Therefore, when considering sample size over a whole development plan, this advantage can be lost.

It seems more conservative estimates of hypothesised response rates to use in sample size calculations can be beneficial to the overall development plan. This is seen in pairing #23. Here, imperfect correspondence was beneficial to the development plan, as the phase III trial was designed to detect a larger treatment effect seen in phase II than there actually was for phase III. Here both development plans had similar ability in making correct conclusions. Ultimately, the development plan with a single-arm phase II trial required less participants overall.

It also seems that both development plans with a single-arm or randomised phase II trial succeed in detecting truly ineffective treatments, as seen in pairing #3. It therefore seems like high failure rates seen in phase III cancer clinical trials are largely due to difficulty detecting truly effective treatment, as opposed to letting futile treatment through too easily. Having said that, it seems as though randomised phase II trials are marginally more likely to lead to futile phase III trials than single-arm phase II trials.

There are limitations to this study. I was limited to choosing five pairings as examples of real-life development plans as the requirements needed to be included in this study was strict. For future research, to widen the scope, clinical trials could be identified through a journal database search like PubMed, as not all clinical trials link their results their clinicaltrial.gov registration. Additionally, it was impossible to choose real-life development plans with single-arm phase II trials as simulated randomised phase II trials could not be generated without an estimation of a concurrent control arm. Also, estimated response rate seen in the published phase II trials were used as “the truth” in simulated phase II trials, when in reality, true response rates are unknown. One solution to this is to widen the search criteria of published phase II-phase III trials outside of a cancer setting which also use single-arm phase II trials, such as cardiology and endocrinology (110-112). It should also be noted that selection bias would be present in the published phase II-phase III pairings, as it wasn't expected that published phase III trials were preceded by an unpromising phase II (although, pairing #21 proved to be an exception). However, as funding a phase III trial is generally justified through a promising phase II trial, there seems to be no solution to select development plans without this selection bias.

Another limitation is that the simulated phase II-phase III development plans did not necessarily reflect the published phase II-phase III pairings. One difference was in the sample sizes, as the published phase II trials often had a larger sample size than the simulated phase II trials. Additionally, published phase III trials often had a smaller sample size than the simulated phase III trials. There are multiple reasons for this.

In the simulated phase II trials, sample sizes were calculated with a more lenient power and two-sided alpha threshold, using 80% and 15% respectively. However, in at least three of the published phase II trials, 87-90% power thresholds were used with two-sided 5% alphas (1, 7, 9). Additionally, for two of the published phase II trials, they contained an additional experimental arm which was ignored for the purposes of the simulation studies (7, 9). This meant that the published phase II sample size calculations were based on power and alpha thresholds between three arms, not just two. For published phase III trials, sample size calculations were based on time-to-event outcomes from each arm collected in the published phase II trial. However, in the simulation studies, the published phase II time-to-event outcome from one arm was used, and then the simulated phase II relative risk was used to determine the treatment effect between the two arms in order to calculate the simulated phase III sample size. This may have led to smaller treatment effect size estimates that the published phase III trials used, leading to larger simulated phase III sample sizes.

It appears the overall recommendation for phase II trial design depends on whether or not the investigators have confidence in the accuracy of hypothesised control arm response rate and hypothesised experimental arm response rate, i.e., HP_0 and HP_1 , for phase II sample size.

If there is a plethora of information regarding likely outcomes for each arm (particularly the control arm), and there are no expectations for the treatment effect to drift towards the null at phase III, then a single-arm phase II trial should be recommended. This is because, under these circumstances, single-arm phase II trials have the same ability to make correct decisions on treatment while reducing mean overall development plan sample size by up to 61% (like pairing #21).

However, if there is uncertainty around any hypothesised estimates (particularly the control arm) or hazard ratio is expected to drift towards the null in phase III, then randomised phase II trials should be used. Not only are they more likely to make correct decisions on treatment, but in these circumstances would lead to reduced total development plan sample size. Another recommendation could be that phase II trials, in these circumstances, collect phase III survival outcomes as they would provide more informative treatment effect estimates for the design of the subsequent phase III trial.

Even though the overall recommendation is for investigators to use a randomised phase II trial, the stronger recommendation is to place importance on designing the development plan well. The more accurate hypothesised values are to calculate sample size, the more likely you are to make correct decisions on treatment. If investigators are aware of imperfect correspondence, between phase II and phase III trials in either direction, this should be accounted for when calculating subsequent phase III sample size. Examples of when phase II investigators could prepare for imperfect correspondence is if imperfect correspondence was previously seen using the same treatment but tested in a different disease area, or if correlation between phase II response rates and phase III survival outcomes have not been well defined for the particular treatment or disease.

6 Discussion

6.1 Summary of Thesis and Implications of Results

Clinical trials play a pivotal role in drug discovery. Investigations into a new drug are streamlined by dividing the experimental process into component parts starting with pre-clinical studies, and followed by phase 0, phase I, phase II, phase III and phase IV clinical trials (14).

In 2016, a report published by the Biotechnology Innovation Organization found that only 5.1% of agents tested in phase I oncology trials led to regulatory approval in the US (35). This was almost half the average success rate of newly discovered drugs reaching regulatory approval across 14 major disease areas. Roadblocks in the cancer drug development pipeline proved to be within phase II and phase III trials, with only 24.6% of phase II cancer trials proceeding to phase III (compared to the average of 30.7%) and only 41.9% of phase III cancer trials progressing to licensing approval (compared to the average of 59.9%) (35). These high failure rates are a particular cause for concern, not only as costs of the drug development process are estimated at £800 million ($1 billion USD) but also because cancer is predicted to be the leading cause of death worldwide by 2060 (17, 18, 38).

This report served as a key motivator behind my thesis. It initiated my investigation into common phase II oncology trial designs, and how this choice could impact the success rates of development plans as a whole.

I conducted a narrative synthesis to identify quantitative papers which had compared two of the most commonly used phase II trial designs in oncology: single-arm and randomised trials. The publications provided useful insights. One finding was that when historical controls in single-arm trials were not representative of the truth, i.e., when the historical controls were not equal to the population control response rates, errors in treatment effect were made. Not only was this to the detriment of the phase II trial, but continued to negatively impact subsequent phase III trials if undertaken. However, when single-arm trials had representative historical controls, i.e., when historical controls were equal to the population control response rates, this design offered the equivalent power as gold-standard randomised-controlled trials, with the added benefit of requiring a smaller sample size, or higher power for the same sample size.

Given the nature of the two trial designs, these findings are as expected. However, as the studies had different aims and methodologies it was difficult to amalgamate all results to comprehensively determine phase II design impact on a drug development plan. For example, a paper by Taylor *et al.* found that single-arm trials were more likely to lead to wrong conclusions with large variability, and randomised trials were more likely to lead to wrong conclusions in smaller sample sizes (56). Taylor *et al.* was able to determine how likely these phase II trials were able to lead to phase III trials, however the conduct of the phase III trials themselves were not simulated. Therefore, Taylor *et al.* could not provide information on how phase II trials could impact phase III conclusions. One paper that did simulate the conduct of subsequent phase III trials was Grayling *et al.,* which found that a development plan that used a group-sequential design was superior to development plans with either single-arm or randomised trial phase II designs (67). However, it was difficult to apply the results of the study to real single-arm phase II trials as Grayling *et al.* did not account for historical control error.

As a result, the narrative synthesis identified five key elements as crucial for inclusion in a quantitative methodological study in order to determine impact of the two phase II trial designs on a phase II-phase III development plan. These key elements include the consideration of:

1. Phase III trial conclusions
2. Both null and alternative hypotheses
3. Historical control error
4. Phase II binary response rates and phase III time-to-event survival endpoints
5. Imperfect correspondence of treatment effect between phase II and phase III

Therefore, my next aim was to conduct a methodological study which shed light on all these elements, ideally all at once.

To ensure results were understandable and to create a template on which I could later build, I designed preliminary simulation studies which considered key elements 1, 2 and 3 (presented in Chapter 3). Overall, the studies found that both development plans with phase II single-arm and randomised trials had a high ability to detect ineffective treatments. The true negative proportions, i.e., proportion of times the development plans correctly failed to reject the null hypothesis, remained above 97.5% in all circumstances. However, when detecting effective treatment, a development plan with a single-arm phase II trial was more likely to make the right conclusions on treatment in most circumstances. Development plans with randomised phase II trials were only favoured when two conditions were met: 1) more than 75 participants could be recruited and 2) the equivalent single-arm trial had negative historical control error (i.e. underestimated the true control response rate) by 5%-points or more. These findings were comparable to results from the quantitative papers identified through the narrative synthesis (56, 60, 62).

Furthermore, the findings demonstrated benefits of single-arm phase II trials with positive historical control error, i.e. overestimation of true control response rate. In these instances, single-arm phase II trials assumed a smaller treatment effect than the truth, which could not be re-estimated with a concurrent control arm. Therefore, the subsequent phase III trials also assumed a smaller treatment effect than the truth for sample size calculations. This enhanced the statistical power of the phase III trial compared to those that followed randomised phase II trials. In these circumstances, development plans with single-arm phase II trials were more likely to correctly conclude in favour of treatment than equivalent development plans with randomised phase II trials. These findings suggest that phase II investigators should choose a single-arm trial design with conservative estimates of treatment effect. Not only is this because it increases the likelihood that the trial itself could detect effective treatment, but also increases the likelihood for the subsequent phase III trial. Furthermore, equivalent randomised phase II trials are more costly as they would require more participants.

However, as the study lacked consideration of time-to-event endpoints at phase III, the findings are restrained to development plans that use binary endpoints at both phases. Furthermore, parameters chosen were not inherently based on values seen in real clinical trials, so the development plans might not have reflected real practice. Not only this, but imperfect phase II-phase III treatment effect correspondence was not considered, which may have inflated the performance of the development plans;

particularly of development plans with single-arm phase II trials which do not re-estimate their control arm.

The next stage of research was to incorporate key elements 4 and 5, however, it became clear that implementing differing endpoints and differing treatment effects between phases was not straight forward. Therefore, chapter 4 entails a simulation study which is more exploratory in nature, to test how both key elements could be considered simultaneously.

In Chapter 4, I developed a method where simulated phase II trials collected relevant information which could be used in phase III sample size calculations. Specifically, phase II trials collected survival data which was treated as a binary outcome i.e., proportion of survivors at 12-months. These results were used to determine whether the treatment was promising enough to warrant further investigation using likelihood ratio tests. Then, if successful, anticipated phase III time-to-event outcomes for a 60-month trial were translated from the binary outcomes to hazard ratios using the exponential curve equation. These anticipated values were used for phase III sample size calculation, thereby linking the two phases. Imperfect correspondence was also implemented using a multiplicative factor, which shifted the hazard ratios towards the null from phase II to phase III. This shift represented the proportion of phase II trials which happened to collect more optimistic estimates of treatment effect that would likely recommend a subsequent phase III trial. This was in opposition to the proportion of phase II trials which happened to collect more pessimistic estimates, likely stopping the development plan at this point. This phenomenon is seen in practice where treatment effects have been seen to be larger in phase II than in phase III (90).

As expected, under imperfect correspondence between phase II and phase III treatment effects, the power of the development plan was vastly affected. These results could be informative to phase III investigators wishing to test an established drug with a large treatment effect in a new population. The recommendation would be that it is still necessary to test the established drug in a phase II trial in the new population to help design the subsequent phase III. Not only this, but information from the original phase II-phase III trials tested in identical patient populations could inform the amount of imperfect correspondence that may be expected when testing the new population group. However, if phase III trials were to be conducted in new populations, investigators should be cautious with treatment effects used for sample size calculations, as the power may be compromised due to possible imperfect correspondence.

However, as Chapter 4 contained proof-of-concept research, the parameters of the simulation study were simplified and only considered development plans with phase II randomised trials. Furthermore, it is difficult to determine if the levels of imperfect correspondence simulated were reflective of real-life trials. Not only this, but the study also did not consider the most common phase II binary endpoints, response rates, meaning the research had limited applicability.

It could be considered that Chapters 4 and 5 encapsulate early-phase methodological research (113). Early-stage methodological research comprises of proposing new ideas of research and providing theoretical methods to address them. These early-phases can also include the assessment of empirical evidence in limited settings. Good examples of these are the limited development plan settings that were simulated in both Chapter 3 and 4, and Chapter 4 comprising of proof-of-concept research to develop methods that address differing phase endpoints and imperfect correspondence between phase treatment effects.

Therefore, the aims of the next chapter would be to implement key elements 4 and 5 which is more reflective of real-life practice, specifically by using phase II response rates and phase III survival. This brought to light the ongoing debate among trialists surrounding how well phase II response rates act as a proxy for phase III overall survival in cancer clinical trials, adding to the complexity of the research (69-71, 96-98, 100-102).

To address these issues, Chapter 5 involved using real phase II trials which had collected both response rate and survival outcomes. The extracted data then informed the parameters of the simulated development plans. In the simulated development plans, the decision to move to a phase III trial was made using the results of the binary response rates. Then, if successful, a combination of the phase II simulated relative risk and survival outcomes informed the anticipated phase III survival for sample size calculation. Survival outcomes of the real subsequent phase III trial were then sampled with replacement to generate simulated phase III results. Therefore, imperfect correspondence was implemented through the difference in hazard ratios seen between the real-life phase II and phase III survival curves.

Chapter 5 aimed to combine methods developed through chapter 3 and chapter 4 to facilitate the consideration of all key elements at once. As discussed, to ensure that the simulation parameters were reflective of real-life practice, data was extracted from five pairings of published phase II-phase III trials (1-10). Due to the requirements needed to inform the simulation, all published phase II trials were conducted with randomised designs, therefore, the main research question of this chapter became "would the performance of these development plans have differed had they used single-arm trials for phase II?"

This simulation study showed several insights into the impact of phase II trial design. Firstly, it identified that both phase II trial designs were highly resistant to making false positive decisions about moving to phase III, similar to the conclusions made in Chapter 3. This seems to suggest that the high failure rate in phase III cancer trials is caused by the failure at phase II in detecting effective treatments, not false positive results from futile treatments.

The simulation study further identified a key "tipping-zone", a grey area where it was uncertain that one phase II trial design would perform better than the other. In one example (pairing #3), a development plan with a single-arm phase II trial could only identify a true treatment effect 42.2% of the time, and a development plan with a randomised phase II trial could only identify the effect 45% of the time. This 2.8%-point increase in the ability to detect true treatment effect was at the cost of recruiting 636 more participants. This prompts the question, *at what point is the additional cost of participants worth the slightly increased ability to detect a true treatment effect?*

Another pairing (pairing #24) also demonstrated a lack of ability in detecting true treatment effect, with only 47.2% of development plans with randomised phase II trials correctly concluding in favour of treatment and only 7.6% of development plans with single-arm phase II trials correctly concluding in favour of treatment. This reduced performance was largely due to the lack of accurate information for anticipated trial parameters when designing the phase II trial, which was further compounded by imperfect correspondence between phase II and phase III. This result demonstrates the need for investigators to conduct thorough investigation to underpin the choice of anticipated response rates for phase II sample size calculation.

Additionally, this study highlighted the ramifications for overestimating historical control response rates in single-arm trials compared to the true population control response rate. As a reminder, in Chapter 5 the value used to represent control response rate for single-arm sample size calculations was the same value used for the historical control. It was found that when anticipated control response rates were overestimated in phase II, the subsequent phase III trials that followed single-arm phase II trials demanded far more patients than those that followed a randomised phase II trial. Therefore, investigators could view the additional participants phase II randomised trials require as an investment to minimize total patient demand for the whole development plan, assuming it continues to phase III. This is in direct contrast to the recommendations in Chapter 3.

Overall, the findings from Chapter 5 highlight the need for investigators to place weight on collecting relevant descriptive data before designing phase II trials. It is particularly important to choose an accurate value for anticipated control arm response rate, especially when opting for a single-arm trial design. If a new drug is being developed with no prior history of testing in the population of interest, then there is no relevant information available, and a randomised phase II trial should be conducted. This would maximise the chance of making the correct decision about the treatment. Only when the quantified uncertainty is satisfactory surrounding the anticipated control response rate should a single-arm trial be considered at phase II. In instances where anticipated control response rate is well researched and understood, and access to participants is limited, a single-arm phase II trial is recommended.

Chapter 5 is an example of late-phase methodological research, following early-stage research from Chapters 3 and 4 (113). Late-stage methodological research provides a more expansive range of settings simulated, evidence of the method's validity, in addition to applications of the method itself. This is represented by the five real-life phase II-phase III trial pairings that were used to inform the simulations, and the comparison of the original phase III outcomes, the equivalent digitized datasets, and the Royston–Parmar models used to mimic the published phase III survival data. Progressing through various stages of methodological research demonstrates that the research aims have been thoroughly investigated and can be used as a trustworthy source for further research.

6.2 Strengths and Limitations

My research has shed light on the impact of phase II design choice in oncology, however, it does not provide a definitive answer. There are various strengths and limitations to my approach which I will discuss in this section.

6.2.1 Strengths

As stated throughout my thesis, one major strength is the identification of the five key elements needed for a study to fully assess the impact of phase II design choice on phase II-phase III oncology development plans. These elements were identified through a narrative synthesis which had a primary aim to find quantitative methodology papers that had compared single-arm and randomised trials. The original purpose was to combine the results from these papers to form a basis for guidelines for phase II investigators. As mentioned previously, the paper identified through the narrative synthesis written by Taylor *et al.* compared the likelihood that each design would lead to a subsequent phase III trial (56). This was a useful starting point. However, in practice, the influence of phase II trials does not stop here, as often phase II estimates are used to calculate phase III sample size. Therefore, this limited the practical relevance of their conclusions. Another paper identified through the narrative synthesis by Hunsberger *et al.* did compare these designs while considering the full dimensions of a phase III trial (60). However, the results have limited applicability as their simulated phase II trials collected time-to-event endpoints. It became clear that the results of each quantitative methodology paper identified in this narrative synthesis could be enhanced with consideration of at least one of the five key elements, thereby identifying a gap in the literature.

Using the key elements as a basis to build a study, I was able to develop a simulation study comparing phase II-phase III development plans which reflected real-life practice to maximise the relevance of the results. This gave me a solid foundation to investigate how phase III cancer success rates could be improved from the choice of phase II trial design.

Another strength of this simulation study was the empirical method used to include both phase II binary response rates to phase III survival outcomes within the same development plan. As previously mentioned, how well response rates predict overall survival is difficult to quantify (69-71, 96-98, 100-102). Therefore, allowing phase II trials to collect both binary and time-to-event data gave the flexibility for each phase II-phase III development plan to simulate estimates of response rates and overall survival to varying degrees. Additionally, allowing simulated phase II trials to collect time-to-event data allowed the implementation of imperfect correspondence; through shifting the hazard ratio between the two phases.

In theory, the recommendations from this thesis could be applied across any phase II-phase III development plan setting which uses either single-arm or randomised trials in phase II, such as such as cardiology or endocrinology (110-112). This is due to the variety of phase II-phase III development plans simulated based on the differing scenarios of the published trials. However, these published phase II-phase III trial pairings may not encompass all typical scenarios seen in all disease areas, impacting the generalisability of the recommendations. Therefore, for disease-areas that have a vastly different disease profile to oncology, the simulation study methods could be applied to alternative disease settings. For example, if researchers wanted to investigate the impact of each of the phase II designs on a cardiology phase II-phase III development plan, published pairings of phase II-phase III cardiology trials could inform data generating mechanisms within the code used in Chapter 5's simulation studies. The results of these cardiology-specific simulation results could then inform recommendations on the scenarios in which each phase II design is preferred in a cardiology setting.

6.2.2 Limitations

One of the main limitations of this thesis is that it only considered two phase II trial designs. Furthermore, the trial designs were both relatively simple, single-stage, stand-alone trials. In practice, Simon's two-stage design is the most commonly conducted single-arm trial design (72, 114-116). Two-stage designs are also commonly seen in randomised phase II trials (117). Two-stage trials are frequently used as they allow early-assessment of treatment effect, with an opportunity to stop the trial early in cases of futility (118). Three-stage single-arm trials have also been used in practice (119). Not only does this benefit patients recruited on the trial, but prevents unnecessary waste of resources. The inclusion of multi-stage designs may have affected the performance of the simulated development plans. For example, allowing for early-stopping at phase II may reduce the overall development plan sample size. Additionally, some simulated development plans that reached phase III testing may have otherwise detected futility early. As a result, the proportion of times each development plan concluded in favour of treatment may have also decreased.

Not only this, but alternative randomised trial designs exist. Throughout this thesis, the randomised design has been defined as a two-armed, controlled trial. In practice, randomised noncomparative trial designs are also used where each arm is assigned a new experimental drug (120). Other alternate types of randomised designs include randomised discontinuation. Additionally, there are alternate non-randomised controlled designs, such as the "pick-the-winner" design (120). Furthermore, phase II trials are not limited to either single-arm or randomised designs. As previously mentioned, two papers identified in the narrative synthesis also considered alternate designs, including Grayling *et al.* who included a group sequential design, and Hunsberger *et al.* who included an integrated phase II/III design (60, 67, 121). Other types of phase II cancer trial designs also include multi-arm, multi-stage (MAMS) designs, adaptive designs and Bayesian designs (24, 43, 122, 123). Therefore, development plans simulated throughout this thesis could be considered too simplistic, which limits the applicability of the results.

In the same vein, this thesis assumes a simplified version of a development plan. It assumes that only one phase III trial follows from one phase II trial, when often this is not the case. Of course, both phase IIA and phase IIB trials can be used within the same development plan (25). Not only this, but phase III trials use multiple sources of evidence for treatment effect to justify the large cost, including evidence from other phase III trials. An example of this was seen in one of the published phase II-phase III pairings identified for Chapter 5 (pairing #21). These clinical trials investigated the use of durvalumab with or without tremelimumab against standard of care (noted as the EXTREME regimen) in patients with squamous cell carcinoma (1, 2). However, the phase III trial was not solely motivated by the previous phase II trial, but also two additional phase III trials where benefits of the treatment arm were also seen in hepatocellular carcinoma patients and metastatic non-small-cell lung carcinoma patients (124, 125). Given this, the structure of my simulated development plans where phase III trials only use the information provided by the previous phase II trial may not always hold true. Therefore, in reality, the drug development pipeline is not as linear as this thesis presents. Similarly, there has been a recent increase in approvals of new cancer drugs directly following phase II trials (126). This means that simulated phase II development plans that demonstrated a large treatment effect may not have always led to phase III trial designs in practice. However, the instances where approvals follow phase II trial designs is relatively low (126, 127).

One other limitation is the small number of published phase II-phase III pairings found to replicate real clinical trial environments. As only five were identified, results from the final study are limited to clinical trials represented by one of the five scenarios that were recreated. This limits the usefulness of the results, especially with uncertainty surrounding tipping-zones i.e. the trade-off between the increased likelihood of making the correct decision on treatment with the increased sample size needed. In these instances, judgements will have to be left to the phase II investigator.

This leads onto one of the other limitations of the thesis. With more time and resources, I would have liked to identify more examples of published phase II-phase III trials. In order to identify the five published pairings used for the study in Chapter 5, I conducted a systematic search of phase III trials using clinicaltrial.gov. If I was to repeat this, I would also use a secondary source such as a citation review to identify further pairings. Not only this, but I would use the opportunity to explore published phase II-phase III pairings beyond the disease of cancer. A wider breadth of clinical trial scenarios could have given more insight into tipping-zones in a wider set of clinical trial parameters. This would improve the generalisability of the recommendations for any development plan in a phase II-phase III setting than often uses either single-arm or randomised phase II trials.

Similarly, another limitation of this project is that not all parameter values that are relevant to oncology trials where exhausted in the simulation studies. For example, in Chapter 3, further fixed sample sizes could have been simulated, in addition to alternative thresholds for power and alpha in sample size calculations. Additionally, different values of P_0 and P_1, treatment effect and historical control error could have been investigated. For Chapter 4, alternative imperfect correspondence factors could have been explored. Furthermore, this thesis primarily assesses development plans based on proportion of correct decisions made on treatment and overall development plan sample size, when other optimal criteria could have been considered such as development plan length or cost.

Additionally, the lag between the initial narrative synthesis conducted in January 2018 and now meant that two quantitative papers that compared single-arm and randomised trials have been published since (128, 129). Snyder *et al.* in 2019 assessed the quality of historical controls in single-arm trials by comparing docetaxel outcomes with those estimated from randomised trials in a non-small-cell lung cancer setting. The study found that there was significant heterogeneity in the historical controls chosen for single-arm trials from 2000-2017. The paper concluded that this demonstrated evidence that the use of historical controls may not be a valuable approach to replace randomised trials (128). The other paper presented by Tan AC *et al.* in 2022 identified nine pairs of early phase single-arm trials and late phase randomised trials which examined one of six treatments in lung cancer. Treatment estimates from the early phase and late phase in each pair were compared, namely overall response rate, overall survival and progression-free survival. It was found that early phase outcomes were consistent with larger randomised trials, and concluded that single-arm trials provide reliable estimates for subsequent randomised trials (129). Had these papers been included in the narrative synthesis, the same conclusions would have been made as neither paper considered all five key elements.

6.3 Future Research

6.3.1 Extension Using More Published Pairings

As previously mentioned, one of the more obvious avenues for further research is to expand upon the findings of Chapter 5. More pairings of published phase II-phase III trials can be identified in alternative databases to clinicaltrial.gov, and also outside of the context of cancer. More examples of published pairings can increase this breadth of investigation and contribute to the optimisation of development plans for other disease areas.

6.3.2 Extension to Include More Trial Designs

Another opportunity for further research is to adapt the existing simulation to allow common adaptations to the simplified single-arm and randomised trial designs. This could include creating simulations that allow for multiple stages, like Simon's two-stage design, and trials with multiple arms (72, 114-117, 119, 122). This could also the inclusion of both phase IIA and phase IIB trials within the same development plan with their differing aims (25). Allowing this flexibility also means that some of the previously dismissed published phase II-phase III pairings identified in the systematic literature review can be used to inform the next simulation studies. It will also give more insight into how each additional arm in a phase II trial could impact a development plan, or even how each additional stage effects the optimal phase II design choice. An expansion can also be made to reflect the non-linear nature of drug development plans. This could be done by allowing for multiple sources of treatment effects to influence the design of phase III trials, possibly using Bayesian methods by using the sources to inform priors for phase III sample size calculation.

6.3.3 Guidance

Based on the work presented in this thesis, three formal recommendations to phase II investigators can be made. These recommendations are for phase II-phase III oncology development plans that use binary phase II endpoints, and time-to-event phase III endpoints.

The most important recommendation that comes from this thesis is that investigators should place high importance on the selection of treatment effect used to inform the phase II sample size, ideally looking at multiple sources of research, particularly for the control arm.

It is reasonable to anticipate that there would be less information available for the experimental arm, especially if it is a new compound that is being investigated. However, as the control arm is often Treatment As Usual or Standard of Care in an oncology setting, there should be multiple reliable sources to provide an estimate for the response rate that is close to the truth. Selecting a value for the control arm that has come from a recent, large phase III or phase IV trial in a similar population of interest would reduce the risk of the trial being overpowered or underpowered. If overpowered, this could lead to an unnecessary waste of resources which can persist into the subsequent phase III trial. If underpowered, it decreases the ability for the phase II-phase III development plan to correctly conclude in favour of effective treatment.

The second recommendation is for when a phase II cancer trial is investigating a new drug that has not been tested in any setting aside from phase I trials. In this scenario, a randomised controlled trial should be used.

This is because there would be very little existing evidence for the treatment effect, and poor estimations of hypothesised treatment effect have a greater negative impact on a phase II-phase III development plan with single-arm phase II trials than those with randomised phase II trials. For example, if a phase II trial is designed based on a smaller treatment effect than the truth, then it may demand a larger sample size than needed. As the single-arm design does not re-estimate the control arm in a phase II setting, this error may persist through to the subsequent phase III trial. This may lead the subsequent phase III trial to demand a larger sample size than needed, potentially leading to a higher cost than a development plan with a randomised phase II trial. Similarly, if the phase II trial is designed with a treatment effect that is larger than the truth then the trial will be underpowered. However, in these scenarios, development plans with randomised phase II trials are more likely to make correct decisions on treatment.

The third recommendation is when a phase II cancer trial investigates an existing drug but is being tested in a similar disease area, or similar patient population. In this scenario, a single-arm trial should be used.

This is because when the phase II trial is designed with effect values that are similar to the truth, a development plan with a single-arm phase II trial requires a smaller overall development plan sample size, whilst having a similar ability to make correct decisions on treatment as a development plan with a randomised phase II trial.

Further research of different published phase II-phase III pairings needs to be conducted before further formal recommendations can be made, as detailed in section 6.3.1.

Further research could also allow guidance in form of a flow-chart to be created, which could guide design choice based on the information available when planning a phase II trial. For example, a hopeful phase II investigator could be asked how much previous evidence is already available to justify anticipated values for phase II sample size calculation. Based on the results of this thesis, a recommendation would be made to use a randomised phase II trial when there is little previous evidence. However, once more published pairings are identified and used in further simulations, investigators could find an example with similar anticipated parameters to theirs. Given each scenario, questions could be asked to help the phase II investigator choose a design based on the simulation results, such as “how many patients are available to recruit” or “are you certain the anticipated control response rate is within a 5%-point error margin?” Examples of trade-offs at each point could be described, i.e., “an equivalent randomised trial would require double the sample size, however, could increase likelihood of detecting true treatment effect by ~3%”. With anticipated clinical trial parameters within these tipping-zones, investigators can make an informed decision on design choice and their potential impact on the development plan.

Additionally, recommendations can be made to phase III investigators about how to account for, and anticipate, expected imperfect correspondence between phase II-phase III trials. Results of the published phase II-phase III pairings identified can be gathered to prove the existence of imperfect correspondence in specific treatment and disease scenarios. Phase III investigators can use this to gauge how much imperfect correspondence they might expect if their experimental drug shares similarities between the published pairings. If so, the level of anticipated imperfect correspondence should factor into the anticipated treatment effect when calculating phase III sample size.

6.3.4 Investigation of Phase II-Phase III Translation

There are many ways the methods developed throughout this thesis can contribute towards phase II-phase III translation of treatment effect. For example, the methods presented in Chapter 4 and Chapter 5 could assist in identifying common imperfect correspondence factors between phase II and phase III trials. This could be used to aid investigators who are using phase II treatment effect estimates to calculate phase III sample size.

Additionally, the methods developed throughout this thesis could help define a translation between phase II response rates and phase III survival outcomes. Specifically, the phase II trials from the five published phase II-phase III pairings for Chapter 5 required both binary response rates and time-to-event survival outcomes to be collected. Phase II response rates can be mapped to overall survival in each of the five settings, and if present, patterns can be identified. This can be further explored if further pairings are identified in alternative databases, or in other disease areas to cancer. From here, the simulation methods developed can be used to inject stochastic probability, which can inform the expected error margins of the translations.

Additionally, if further research was conducted to find more published phase II-phase III pairings, these could be used to describe more nuances that exist in the translation between these two outcomes at different phases. If translations can be identified, this can provide further insight on whether phase II response rates are a valid outcome to predict phase III survival. Further, if the translations are accurate and the error margins are small, then they can be used to improve the accuracy of predicting phase III survival using estimations of phase II response rates for sample size calculations. If translations are unreliable with large margins of error, then this may indicate that phase II response rates are not appropriate proxy for phase III survival, and current practice should change to collect phase II time-to-event outcomes instead. In either scenario, further investigation into the translation can produce more reliable treatment effect estimates for phase III trials. This could, over time, improve development plans, and therefore improve the overall success rates of phase III trials.

6.4 Conclusion

The aim of this thesis was to address the question: what is the impact of phase II trial design choice in cancer research? The motivation was to investigate whether choice of phase II design could improve the failure rate of 58% currently seen in phase III cancer clinical trials. Therefore, this thesis focused on the choice between the two most commonly used phase II cancer clinical trials: single-arm and randomised.

As a result, key findings from the simulation study conducted can be used to inform the future practice of phase II investigators. These will in turn improve the likelihood of detecting an effective treatment, and recruit patients more efficiently to gain the highest quality research. It was also able to identify other aspects of practice that can change to improve performance of development plans, such as justifying values used for anticipated response rates for phase II sample size calculation. Based on the research from this thesis, these changes to the practice of designing phase II trials can be made to improve the success rates of new drugs developed to treat cancer. This will prevent effective drugs slipping through the radar as readily, reduce the average time and cost it takes for a new cancer drug to obtain regulatory approval, and allow patients to get access to life-saving treatment sooner.

7 Bibliography

1. Siu LL, Even C, Mesía R, Remenar E, Daste A, Delord JP, et al. Safety and Efficacy of Durvalumab With or Without Tremelimumab in Patients With PD-L1-Low/Negative Recurrent or Metastatic HNSCC: The Phase 2 CONDOR Randomized Clinical Trial. JAMA Oncol. 2019;5(2):195-203.
2. Psyrri A, Fayette J, Harrington K, Gillison M, Ahn MJ, Takahashi S, et al. Durvalumab with or without tremelimumab versus the EXTREME regimen as first-line treatment for recurrent or metastatic squamous cell carcinoma of the head and neck: KESTREL, a randomized, open-label, phase III study. Ann Oncol. 2023;34(3):262-74.
3. Hodi FS, Chesney J, Pavlick AC, Robert C, Grossmann KF, McDermott DF, et al. Combined nivolumab and ipilimumab versus ipilimumab alone in patients with advanced melanoma: 2-year overall survival outcomes in a multicentre, randomised, controlled, phase 2 trial. The Lancet Oncology. 2016;17(11):1558-68.
4. Wolchok JD, Chiarion-Sileni V, Gonzalez R, Rutkowski P, Grob J-J, Cowey CL, et al. Overall Survival with Combined Nivolumab and Ipilimumab in Advanced Melanoma. New England Journal of Medicine. 2017;377(14):1345-56.
5. Bokemeyer C, Bondarenko I, Makhson A, Hartmann JT, Aparicio J, de Braud F, et al. Fluorouracil, leucovorin, and oxaliplatin with and without cetuximab in the first-line treatment of metastatic colorectal cancer. J Clin Oncol. 2009;27(5):663-71.
6. Qin S, Li J, Wang L, Xu J, Cheng Y, Bai Y, et al. Efficacy and Tolerability of First-Line Cetuximab Plus Leucovorin, Fluorouracil, and Oxaliplatin (FOLFOX-4) Versus FOLFOX-4 in Patients With RAS Wild-Type Metastatic Colorectal Cancer: The Open-Label, Randomized, Phase III TAILOR Trial. J Clin Oncol. 2018;36(30):3031-9.
7. Hironaka S, Sugimoto N, Yamaguchi K, Moriwaki T, Komatsu Y, Nishina T, et al. S-1 plus leucovorin versus S-1 plus leucovorin and oxaliplatin versus S-1 plus cisplatin in patients with advanced gastric cancer: a randomised, multicentre, open-label, phase 2 trial. Lancet Oncol. 2016;17(1):99-108.
8. Kang YK, Chin K, Chung HC, Kadowaki S, Oh SC, Nakayama N, et al. S-1 plus leucovorin and oxaliplatin versus S-1 plus cisplatin as first-line therapy in patients with advanced gastric cancer (SOLAR): a randomised, open-label, phase 3 trial. Lancet Oncol. 2020;21(8):1045-56.
9. Johnson DH, Fehrenbacher L, Novotny WF, Herbst RS, Nemunaitis JJ, Jablons DM, et al. Randomized phase II trial comparing bevacizumab plus carboplatin and paclitaxel with carboplatin and paclitaxel alone in previously untreated locally advanced or metastatic non-small-cell lung cancer. J Clin Oncol. 2004;22(11):2184-91.
10. Sandler A, Gray R, Perry MC, Brahmer J, Schiller JH, Dowlati A, et al. Paclitaxel-carboplatin alone or with bevacizumab for non-small-cell lung cancer. N Engl J Med. 2006;355(24):2542-50.
11. Kandi V, Vadakedath S. Clinical Trials and Clinical Research: A Comprehensive Review. Cureus Journal of Medical Science. 2023;15(2).
12. Iasonos A, O'Quigley J. Randomised Phase 1 clinical trials in oncology. Br J Cancer. 2021;125(7):920-6.
13. Umscheid CA, Margolis DJ, Grossman CE. Key concepts of clinical trials: a narrative review. Postgrad Med. 2011;123(5):194-204.
14. Sinha S, Vohora D. Drug discovery and development: An overview. Pharmaceutical medicine and translational clinical research. 2018:19-32.
15. Aban IB, George B. Statistical considerations for preclinical studies. Experimental Neurology. 2015;270:82-7.
16. Huang WL, du Sert NP, Vollert J, Rice ASC. General Principles of Preclinical Study Design. Good Research Practice in Non-Clinical Pharmacology and Biomedicine. 2020;257:55-69.
17. Mohs RC, Greig NH. Drug discovery and development: Role of basic biological research. Alzheimers Dement (N Y). 2017;3(4):651-7.

18. Hughes JP, Rees S, Kalindjian SB, Philpott KL. Principles of early drug discovery. Br J Pharmacol. 2011;162(6):1239-49.
19. Brodniewicz T, Grynkiewicz G. Preclinical drug development. Acta Pol Pharm. 2010;67(6):578-85.
20. Kummar S, Rubinstein L, Kinders R, Parchment RE, Gutierrez ME, Murgo AJ, et al. Phase 0 clinical trials: conceptions and misconceptions. Cancer J. 2008;14(3):133-7.
21. Murgo AJ, Kummar S, Rubinstein L, Gutierrez M, Collins J, Kinders R, et al. Designing phase 0 cancer clinical trials. Clin Cancer Res. 2008;14(12):3675-82.
22. Le Tourneau C, Lee JJ, Siu LL. Dose escalation methods in phase I cancer clinical trials. J Natl Cancer Inst. 2009;101(10):708-20.
23. Wong KM, Capasso A, Eckhardt SG. The changing landscape of phase I trials in oncology. Nature Reviews Clinical Oncology. 2016;13(2):106-17.
24. Van Norman GA. Phase II Trials in Drug Development and Adaptive Trial Design. JACC Basic Transl Sci. 2019;4(3):428-37.
25. Brown S, Brown J, Gregory W, Twelves C. A Practical Guide to Designing Phase II Trials in Oncology Introduction. Practical Guide to Designing Phase Ii Trials in Oncology. 2014:1-11.
26. Lavine KJ, Mann DL. Rethinking Phase II Clinical Trial Design in Heart Failure. Clin Investig (Lond). 2013;3(1):57-68.
27. Cummings JL. Optimizing phase II of drug development for disease-modifying compounds. Alzheimers Dement. 2008;4(1 Suppl 1):S15-20.
28. Buyse M. Phase III design: principles. Chin Clin Oncol. 2016;5(1):10.
29. Suvarna V. Phase IV of Drug Development. Perspect Clin Res. 2010;1(2):57-60.
30. Henry BM, Lippi G, Nasser A, Ostrowski P. Characteristics of Phase IV Clinical Trials in Oncology: An Analysis Using the ClinicalTrials.gov Registry Data. Curr Oncol. 2023;30(6):5932-45.
31. Subbiah V. The next generation of evidence-based medicine. Nat Med. 2023;29(1):49-58.
32. Comprehensive Toxicology, Vol 3: Toxicology Testing and Evaluation, 2nd Edition. Comprehensive Toxicology, Vol 3: Toxicology Testing and Evaluation, 2nd Edition. 2010:1-259.
33. Palmer S, Raftery J. Economic Notes: opportunity cost. BMJ. 1999;318(7197):1551-2.
34. GOV.UK. Public sector expenditure on medical research in the United Kingdom (UK) from 2013/14 to 2018/19 (in million GBP). Statista. Statista Inc2019.
35. Thomas DW. Clinical development success rates 2006–2015. BIO Industry Anal. 2016;1:16.
36. Nagai H, Kim YH. Cancer prevention from the perspective of global cancer burden patterns. J Thorac Dis. 2017;9(3):448-51.
37. Ma X, Yu H. Global burden of cancer. Yale J Biol Med. 2006;79(3-4):85-94.
38. Mattiuzzi C, Lippi G. Current Cancer Epidemiology. Journal of Epidemiology and Global Health. 2019;9(4):217-22.
39. Zhou EW, Jackson MJ, Ledley FD. Spending on Phased Clinical Development of Approved Drugs by the US National Institutes of Health Compared With Industry. JAMA Health Forum. 2023;4(7):e231921.
40. 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.
41. Evans SR. Clinical trial structures. Journal of experimental stroke & translational medicine. 2010;3(1):8.
42. Collignon O, Schritz A, Spezia R, Senn SJ. Implementing Historical Controls in Oncology Trials. Oncologist. 2021;26(5):E859-E62.
43. Lee JJ, Feng L. Randomized phase II designs in cancer clinical trials: current status and future directions. J Clin Oncol. 2005;23(19):4450-7.

44. Grossman SA, Schreck KC, Ballman K, Alexander B. Point/counterpoint: randomized versus single-arm phase II clinical trials for patients with newly diagnosed glioblastoma. Neuro Oncol. 2017;19(4):469-74.
45. Thall PF, Simon R. Practical Bayesian guidelines for phase IIB clinical trials. Biometrics. 1994;50(2):337-49.
46. Shih WJ, Zhao YQ, Xie T. Modified Simon's Two-Stage Design for Phase IIA Clinical Trials in Oncology-Dynamic Monitoring and More Flexibility. Stat Biopharm Res. 2023;15(4):838-44.
47. 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.
48. Kahan BC, Rehal S, Cro S. Risk of selection bias in randomised trials. Trials. 2015;16:405.
49. Berger VW, Bour LJ, Carter K, Chipman JJ, Everett CC, Heussen N, et al. A roadmap to using randomization in clinical trials. BMC Med Res Methodol. 2021;21(1):168.
50. Berger VW. Selection Bias and Covariate Imbalances in Randomized Clinical Trials Preface. Selection Bias and Covariate Imbalances in Randomized Clinical Trials. 2005:Ix-+.
51. Altman DG, Bland JM. Statistics notes - Treatment allocation in controlled trials: why randomise? British Medical Journal. 1999;318(7192):1209-.
52. Kang M, Ragan BG, Park JH. Issues in outcomes research: an overview of randomization techniques for clinical trials. J Athl Train. 2008;43(2):215-21.
53. Sacks H, Chalmers TC, Smith H, Jr. Randomized versus historical controls for clinical trials. Am J Med. 1982;72(2):233-40.
54. Monzon JG, Hay AE, McDonald GT, Pater JL, Meyer RM, Chen E, et al. Correlation of single arm versus randomised phase 2 oncology trial characteristics with phase 3 outcome. Eur J Cancer. 2015;51(17):2501-7.
55. Simmons Z. Can We Eliminate Placebo in Als Clinical Trials? Muscle & Nerve. 2009;39(6):861-5.
56. 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. Clin Trials. 2006;3(4):335-48.
57. De Ridder F. Predicting the outcome of phase III trials using phase II data: a case study of clinical trial simulation in late stage drug development. Basic Clin Pharmacol Toxicol. 2005;96(3):235-41.
58. Popay J, Roberts H, Sowden A, Petticrew M, Arai L, Rodgers M, et al. Guidance on the conduct of narrative synthesis in systematic reviews. A product from the ESRC methods programme Version. 2006;1:b92.
59. Redman MW, Goldman BH, LeBlanc M, Schott A, Baker LH. Modeling the Relationship between Progression-Free Survival and Overall Survival: The Phase II/III Trial. Clinical Cancer Research. 2013;19(10):2646-56.
60. Hunsberger S, Zhao Y, Simon R. A comparison of phase II study strategies. Clin Cancer Res. 2009;15(19):5950-5.
61. Moroz V, Wilson JS, Kearns P, Wheatley K. Comparison of anticipated and actual control group outcomes in randomised trials in paediatric oncology provides evidence that historically controlled studies are biased in favour of the novel treatment. Trials. 2014;15.
62. Pond GR, Abbasi S. Quantitative evaluation of single-arm versus randomized phase II cancer clinical trials. Clin Trials. 2011;8(3):260-9.
63. Sambucini V. Comparison of single-arm vs. randomized phase II clinical trials: a Bayesian approach. J Biopharm Stat. 2015;25(3):474-89.
64. 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. J Clin Oncol. 2010;28(11):1936-41.
65. Sharma MR, Karrison TG, Jin Y, Bies RR, Maitland ML, Stadler WM, Ratain MJ. Resampling phase III data to assess phase II trial designs and endpoints. Clin Cancer Res. 2012;18(8):2309-15.

66. Maitland ML, Hudoba C, Snider KL, Ratain MJ. Analysis of the yield of phase II combination therapy trials in medical oncology. Clin Cancer Res. 2010;16(21):5296-302.
67. Grayling MJ, Mander AP. Do single-arm trials have a role in drug development plans incorporating randomised trials? Pharmaceutical Statistics. 2016;15(2):143-51.
68. Digumarti R, Bapsy PP, Suresh AV, Bhattacharyya GS, Dasappa L, Shan JS, Gerber DE. Bavituximab plus paclitaxel and carboplatin for the treatment of advanced non-small-cell lung cancer. Lung Cancer. 2014;86(2):231-6.
69. Huff CA, Matsui W, Smith BD, Jones RJ. The paradox of response and survival in cancer therapeutics. Blood. 2006;107(2):431-4.
70. Ou FS, Tang J, An MW, Mandrekar SJ. Modeling tumor measurement data to predict overall survival (OS) in cancer clinical trials. Contemporary Clinical Trials Communications. 2021;23.
71. Jaki T, André V, Su TL, Whitehead J. Designing exploratory cancer trials using change in tumour size as primary endpoint. Statistics in Medicine. 2013;32(15):2544-54.
72. Simon R. Optimal two-stage designs for phase II clinical trials. Control Clin Trials. 1989;10(1):1-10.
73. Jung SH. Randomized phase II trials with a prospective control. Stat Med. 2008;27(4):568-83.
74. Escudier B, Eisen T, Stadler WM, Szczylik C, Oudard S, Siebels M, et al. Sorafenib in advanced clear-cell renal-cell carcinoma. N Engl J Med. 2007;356(2):125-34.
75. Escudier B, Choueiri TK, Oudard S, Szczylik C, Negrier S, Ravaud A, et al. Prognostic factors of metastatic renal cell carcinoma after failure of immunotherapy: new paradigm from a large phase III trial with shark cartilage extract AE 941. J Urol. 2007;178(5):1901-5.
76. Sanoff HK, Sargent DJ, Campbell ME, Morton RF, Fuchs CS, Ramanathan RK, et al. Five-year data and prognostic factor analysis of oxaliplatin and irinotecan combinations for advanced colorectal cancer: N9741. J Clin Oncol. 2008;26(35):5721-7.
77. Saltz LB, Cox JV, Blanke C, Rosen LS, Fehrenbacher L, Moore MJ, et al. Irinotecan plus fluorouracil and leucovorin for metastatic colorectal cancer. Irinotecan Study Group. N Engl J Med. 2000;343(13):905-14.
78. Haslam A, Olivier T, Prasad V. Design, power, and alpha levels in randomized phase II oncology trials. ESMO Open. 2023;8(1):100779.
79. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Statistics in Medicine. 2019;38(11):2074-102.
80. Rahman RA, Mariam NBG, Mistry H, Aruketty S, Church M, Adamson-Raieste A, et al. Differential response rates in early-phase cancer clinical trials (EPCCT). Journal of Clinical Oncology. 2021;39(15).
81. Xu W, Huang SH, Su J, Gudi S, O'Sullivan B. Statistical fundamentals on cancer research for clinicians: Working with your statisticians. Clin Transl Radiat Oncol. 2021;27:75-84.
82. Follmann D, Proschan M. Two Stage Designs for Phase III Clinical Trials. medRxiv. 2020.
83. Khan I, Sarker SJ, 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. Br J Cancer. 2012;107(11):1801-9.
84. A'Hern RP. Sample size tables for exact single-stage phase II designs. Stat Med. 2001;20(6):859-66.
85. Fleming TR. One-sample multiple testing procedure for phase II clinical trials. Biometrics. 1982;38(1):143-51.
86. Perneger TV. How to use likelihood ratios to interpret evidence from randomized trials. Journal of Clinical Epidemiology. 2021;136:235-42.
87. Wason JM, Dentamaro A, Eisen TG. The power of phase II end-points for different possible mechanisms of action of an experimental treatment. Eur J Cancer. 2015;51(8):984-92.
88. Steyn H. On extensions of the binomial distributions of Bernoulli and Poisson. South African Statistical Journal. 1975;9(2):163-72.

89. Zichi C, Paratore C, Gargiulo P, Mariniello A, Reale ML, Audisio M, et al. Adoption of multiple primary endpoints in phase III trials of systemic treatments in patients with advanced solid tumours. A systematic review. Eur J Cancer. 2021;149:49-60.

90. Liang F, Wu Z, Mo M, Zhou C, Shen J, Wang Z, Zheng Y. Comparison of treatment effect from randomised controlled phase II trials and subsequent phase III trials using identical regimens in the same treatment setting. Eur J Cancer. 2019;121:19-28.

91. Hanin L. Paradoxical Effects of Tumor Shrinkage on Long-Term Survival of Cancer Patients. Frontiers in Applied Mathematics and Statistics. 2020;6.

92. Martínez JC, Geskus RB, Kim KM, Melis GG. Using the geometric average hazard ratio in sample size calculation for time-to-event data with composite endpoints. Bmc Medical Research Methodology. 2021;21(1).

93. Torri V, Simon R, Russek-Cohen E, Midthune D, Friedman M. Statistical model to determine the relationship of response and survival in patients with advanced ovarian cancer treated with chemotherapy. J Natl Cancer Inst. 1992;84(6):407-14.

94. Rubinstein LV, Korn EL, Freidlin B, Hunsberger S, Ivy SP, Smith MA. Design issues of randomized phase II trials and a proposal for phase II screening trials. J Clin Oncol. 2005;23(28):7199-206.

95. Dhani N, Tu D, Sargent DJ, Seymour L, Moore MJ. Alternate endpoints for screening phase II studies. Clin Cancer Res. 2009;15(6):1873-82.

96. Suzuki C, Blomqvist L, Sundin A, Jacobsson H, Bystrom P, Berglund A, et al. The initial change in tumor size predicts response and survival in patients with metastatic colorectal cancer treated with combination chemotherapy. Ann Oncol. 2012;23(4):948-54.

97. Mandrekar SJ, An MW, Meyers J, Grothey A, Bogaerts J, Sargent DJ. Evaluation of Alternate Categorical Tumor Metrics and Cut Points for Response Categorization Using the RECIST 1.1 Data Warehouse. Journal of Clinical Oncology. 2014;32(8):841-+.

98. Tuma RS. Sometimes size doesn't matter: reevaluating RECIST and tumor response rate endpoints. J Natl Cancer Inst. 2006;98(18):1272-4.

99. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71.

100. Karrison TG, Maitland ML, Stadler WM, Ratain MJ. Design of phase II cancer trials using a continuous endpoint of change in tumor size: application to a study of sorafenib and erlotinib in non small-cell lung cancer. J Natl Cancer Inst. 2007;99(19):1455-61.

101. Piessevaux H, Buyse M, Schlichting M, Van Cutsem E, Bokemeyer C, Heeger S, Tejpar S. Use of Early Tumor Shrinkage to Predict Long-Term Outcome in Metastatic Colorectal Cancer Treated With Cetuximab. Journal of Clinical Oncology. 2013;31(30):3764-+.

102. An MW, Dong X, Meyers J, Han Y, Grothey A, Bogaerts J, et al. Evaluating Continuous Tumor Measurement-Based Metrics as Phase II Endpoints for Predicting Overall Survival. Jnci-Journal of the National Cancer Institute. 2015;107(11).

103. Schünemann H BJ, Guyatt G, Oxman A, editors. GRADE handbook for grading quality of evidence and strength of recommendations - Chapter 5, section 5.2.4.1. The GRADE Working Group, 2013.Updated October 2013. Available from: guidelinedevelopment.org/handbook.

104. Wei YH, Royston P. Reconstructing time-to-event data from published Kaplan-Meier curves. Stata J. 2017;17(4):786-801.

105. Royston P, Parmar MK. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Stat Med. 2002;21(15):2175-97.

106. Crowther MJ. merlin—A unified modeling framework for data analysis and methods development in Stata. The Stata Journal. 2020;20(4):763-84.

107. Royston P, Choodari-Oskooei B, Parmar MKB, Rogers JK. Combined test versus logrank/Cox test in 50 randomised trials. Trials. 2019;20.

108. Crowther MJ. SURVSIM: Stata module to simulate complex survival data. Statistical Software Components. Boston College Department of Economics2011///Aug.

109. Team RC. R: A Language and Environment for Statistical Computing. In: Computing RFfS, editor. Vienna, Austria2021.

110. Savage P, Cox B, Linden K, Coburn J, Shahmohammadi M, Menown I. Advances in Clinical Cardiology 2021: A Summary of Key Clinical Trials. Adv Ther. 2022;39(6):2398-437.

111. Dake MD, Ansel GM, Jaff MR, Ohki T, Saxon RR, Smouse HB, et al. Sustained Safety and Effectiveness of Paclitaxel-Eluting Stents for Femoropopliteal Lesions. Journal of the American College of Cardiology. 2013;61(24):2417-27.

112. Groeneweg S, Peeters RP, Moran C, Stoupa A, Auriol F, Tonduti D, et al. Effectiveness and safety of the tri-iodothyronine analogue Triac in children and adults with MCT8 deficiency: an international, single-arm, open-label, phase 2 trial. The Lancet Diabetes & Endocrinology. 2019;7(9):695-706.

113. Heinze G, Boulesteix AL, Kammer M, Morris TP, White IR, Simulation Panel of the Si. Phases of methodological research in biostatistics-Building the evidence base for new methods. Biom J. 2023:e2200222.

114. Grayling MJ, Mander AP. Two-Stage Single-Arm Trials Are Rarely Analyzed Effectively or Reported Adequately. JCO Precision Oncology. 2021(5):1813-20.

115. Ivanova A, Paul B, Marchenko O, Song G, Patel N, Moschos SJ. Nine-year change in statistical design, profile, and success rates of phase II oncology trials. Journal of biopharmaceutical statistics. 2016;26(1):141-9.

116. Torres-Saavedra PA, Winter KA. An Overview of Phase 2 Clinical Trial Designs. Int J Radiat Oncol Biol Phys. 2022;112(1):22-9.

117. Logan BR. Optimal two-stage randomized phase II clinical trials. Clinical Trials. 2005;2(1):5-12.

118. Porcher R, Desseaux K. What inference for two-stage phase II trials? Bmc Medical Research Methodology. 2012;12.

119. Chen TT. Optimal three-stage designs for phase II cancer clinical trials. Stat Med. 1997;16(23):2701-11.

120. 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.

121. Wang M, Dignam JJ, Zhang QE, DeGroot JF, Mehta MP, Hunsberger S. Integrated phase II/III clinical trials in oncology: a case study. Clin Trials. 2012;9(6):741-7.

122. Millen GC, Yap C. Adaptive trial designs: what are multiarm, multistage trials? Archives of Disease in Childhood-Education and Practice Edition. 2020;105(6):376-8.

123. Chen L, Pan J, Wu Y, Wang J, Chen F, Zhao J, Chen P. Bayesian two-stage design for phase II oncology trials with binary endpoint. Stat Med. 2022;41(12):2291-301.

124. Abou-Alfa GK, Lau G, Kudo M, Chan SL, Kelley RK, Furuse J, et al. Tremelimumab plus durvalumab in unresectable hepatocellular carcinoma. NEJM Evidence. 2022;1(8):EVIDoa2100070.

125. Johnson M, Cho BC, Luft A, Alatorre-Alexander J, Geater S, Laktionov K, et al. PL02. 01 Durvalumab±tremelimumab+ chemotherapy as first-line treatment for mNSCLC: results from the phase 3 POSEIDON study. Journal of Thoracic Oncology. 2021;16(10):S844.

126. Chabner B. Approval of New Agents after Phase II Trials. American Society of Clinical Oncology Educational Book. 2012(32):e1-e3.

127. DeLoughery EP, Prasad V. The US Food and Drug Administration's use of regular approval for cancer drugs based on single-arm studies: implications for subsequent evidence generation. Ann Oncol. 2018;29(3):527-9.

128. Snyders K, Cho D, Hong JH, Lord S, Asher R, Marschner I, Lee CK. Benchmarking single-arm studies against historical controls from non-small cell lung cancer trials ? an empirical analysis of bias. Acta Oncologica. 2020;59(1):90-5.

129. Tan AC, Tan SH, Zhou SQ, Peters S, Curigliano G, Tan DSW. Efficacy of targeted therapies for oncogene-driven lung cancer in early single-arm versus late phase randomized clinical trials: A comparative analysis. Cancer Treatment Reviews. 2022;104.

8 Appendix

8.1 Equations

Fleming sample size

$$N = \left(\frac{\left[Z_{1-\beta}\{P_1(1-P_1)\}^{\frac{1}{2}} + Z_{1-\alpha}\{P_0(1-P_0)\}^{\frac{1}{2}} \right]}{(P_1 - P_0)} \right)^2$$

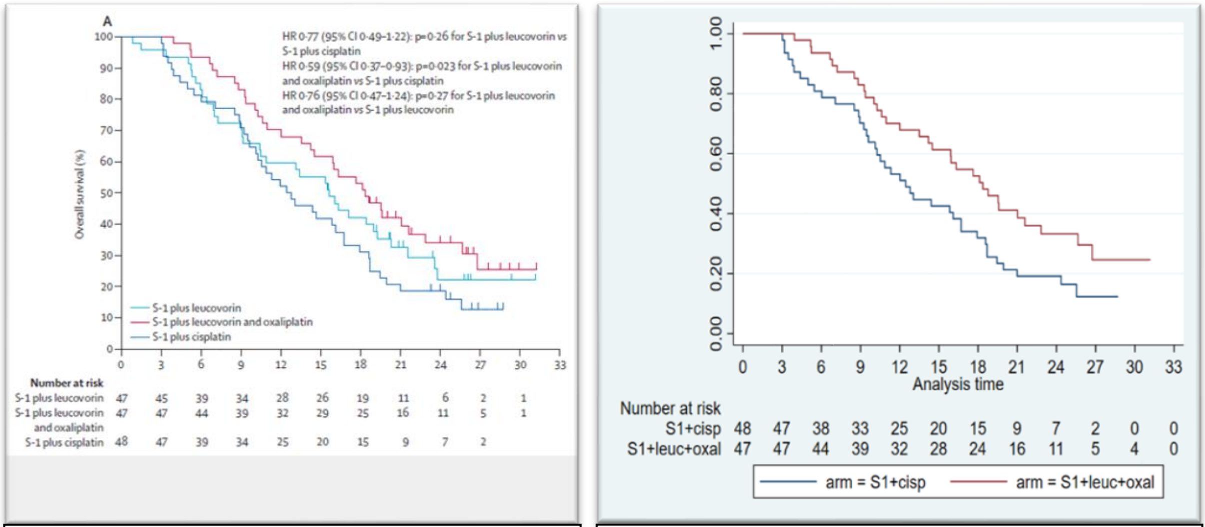
(85)

8.2 Comparing Hazard Ratios From Papers to Digitized Datasets

8.2.1 Pairing #3

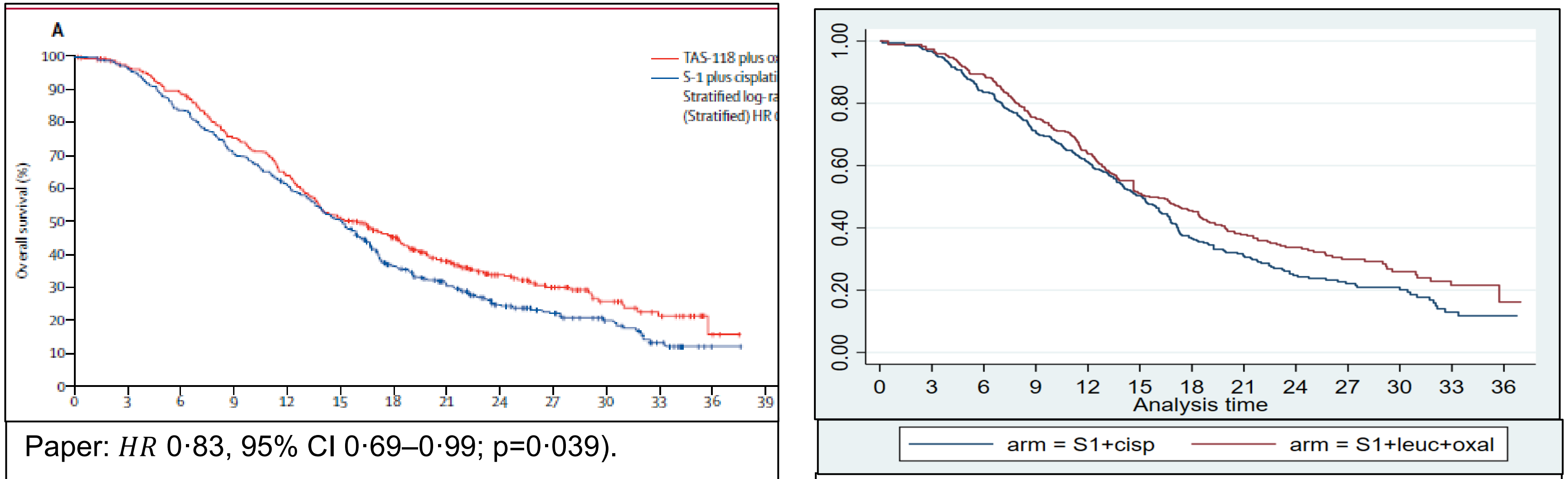
Figure 29 and Figure 30 compare the published Kaplan-Meier curves and respective hazard ratios extracted from pairing #3 with the Kaplan-Meier curves and respective hazard ratios obtained from the associated digitized dataset.

Figure 29 compares the survival outcomes from the phase II setting, and Figure 30 compares the survival outcomes from the phase III setting.



Paper: HR 0·59 (95% CI 0·37–0·93): p=0·023 (for S-1 plus leucovorin and	Digitized: HR 0·59 (95% CI 0·37–0·94): p=0·028

Figure 29 - Comparing Kaplan-Meier curves to those from the published phase II paper of pairing #3, and those I generated from the associated digitized dataset.



Paper: HR 0·83, 95% CI 0·69–0·99; p=0·039).

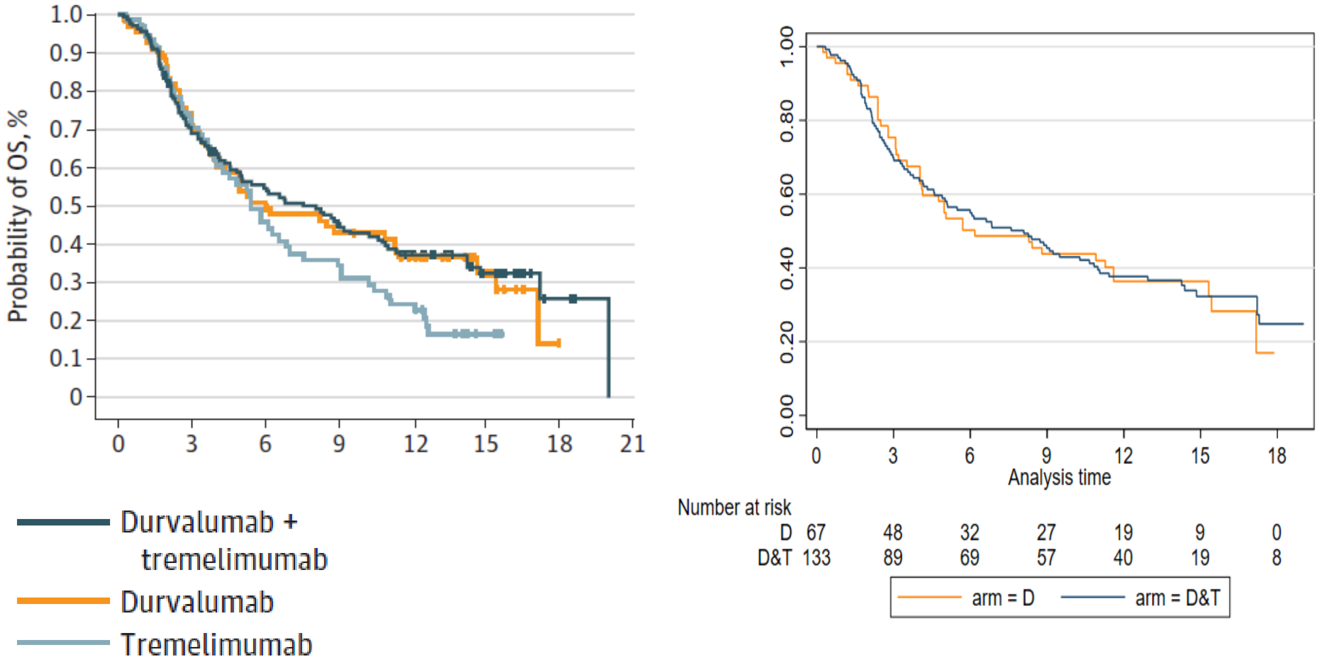
Digitized: HR 0·82, 95% CI 0·69–0·98; p=0·027).

Figure 30 - Comparing Kaplan-Meier curves to those from the published phase III paper of pairing #3, and those I generated from the associated digitized dataset.

8.2.2 Pairing #21

Figure 31 and Figure 32 compare the published Kaplan-Meier curves and respective hazard ratios extracted from pairing #21 with the Kaplan-Meier curves and respective hazard ratios obtained from the associated digitized dataset.

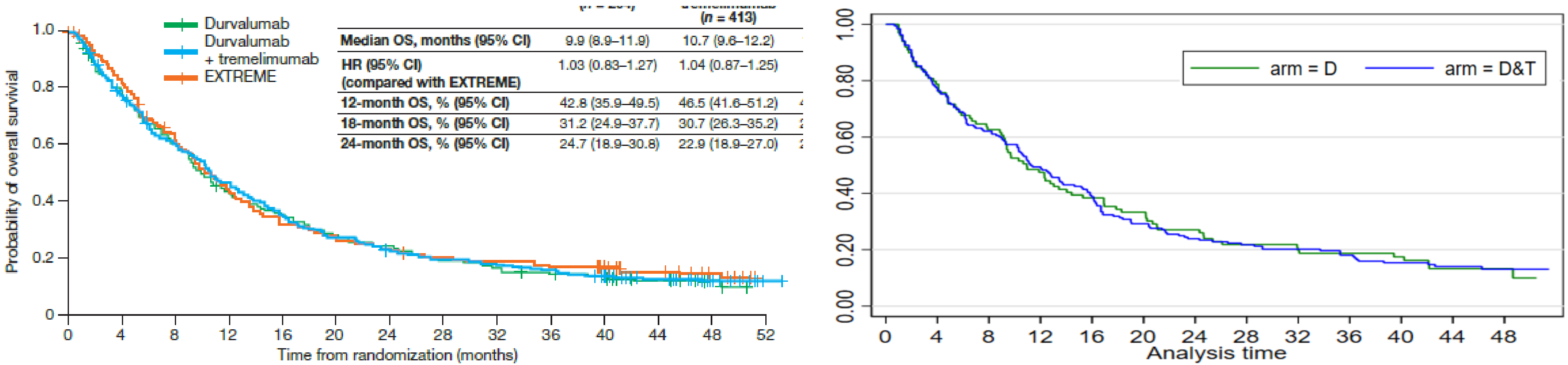
Figure 31 compares the survival outcomes from the phase II setting, and Figure 32 compares the survival outcomes from the phase III setting.



Paper: HR 0·99, 95% CI 0·69–1.43; p=0·89) (durvalumab + tremelimumab vs durvalumab

Digitized: HR 0·96, 95% CI 0·67–1.39; p=0·85).

Figure 31 - Comparing Kaplan-Meier curves to those from the published phase II paper of pairing #21, and those I generated from the associated digitized dataset.



		(n = 413)
Median OS, months (95% CI)	9.9 (8.9–11.9)	10.7 (9.6–12.2)
HR (95% CI) **(compared with EXTREME)**	1.03 (0.83–1.27)	1.04 (0.87–1.25)
12-month OS, % (95% CI)	42.8 (35.9–49.5)	46.5 (41.6–51.2)
18-month OS, % (95% CI)	31.2 (24.9–37.7)	30.7 (26.3–35.2)
24-month OS, % (95% CI)	24.7 (18.9–30.8)	22.9 (18.9–27.0)

Paper: N/A

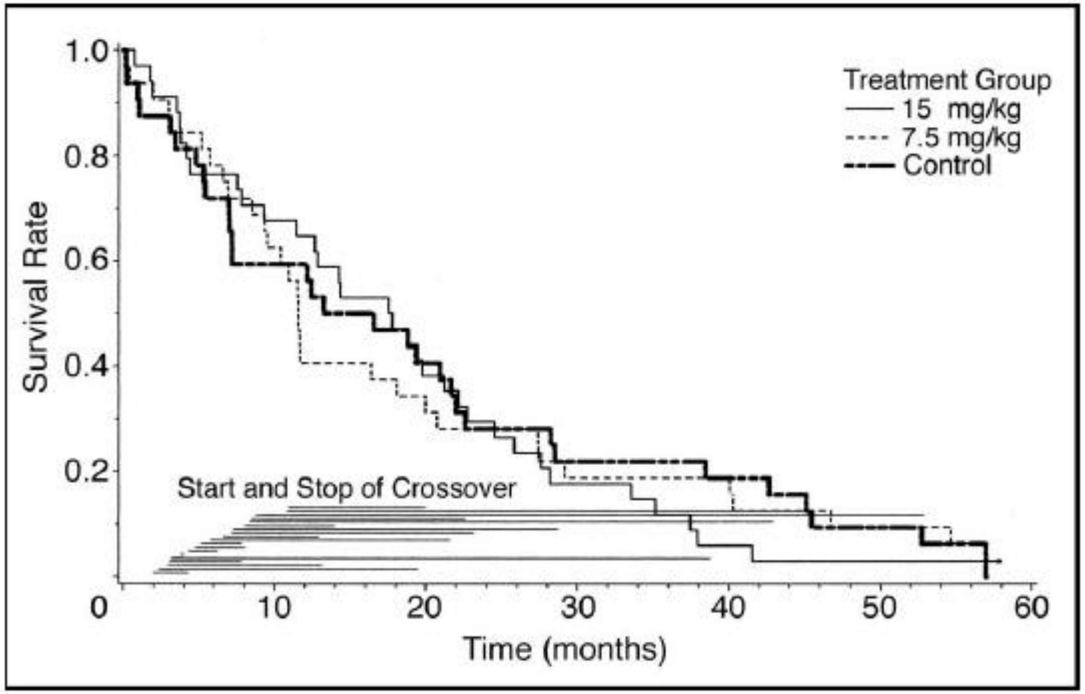
Digitized: HR: 1.00, 95% CI 0.77-1.3, p=0.98

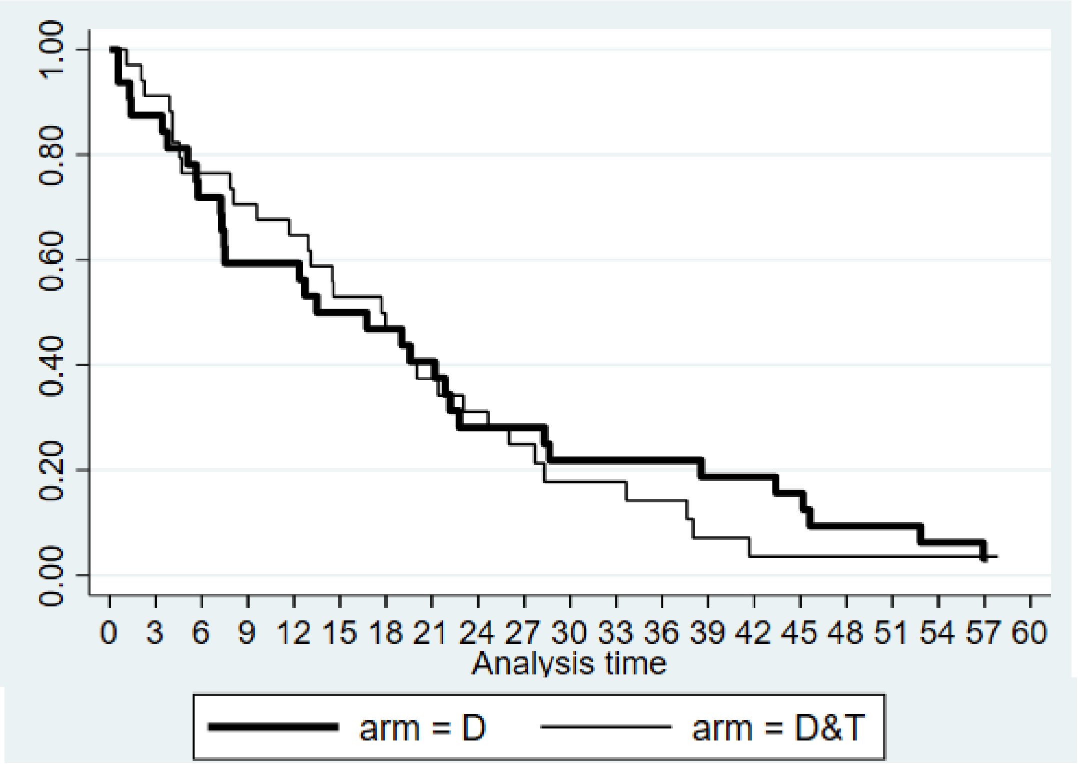
Figure 32 - Comparing Kaplan-Meier curves to those from the published phase III paper of pairing #21, and those I generated from the associated digitized dataset.

8.2.3 Pairing #22

Figure 33 and Figure 34 compare the published Kaplan-Meier curves and respective hazard ratios extracted from pairing #22 with the Kaplan-Meier curves and respective hazard ratios obtained from the associated digitized dataset.

Figure 33 compares the survival outcomes from the phase II setting, and Figure 34 compares the survival outcomes from the phase III setting.

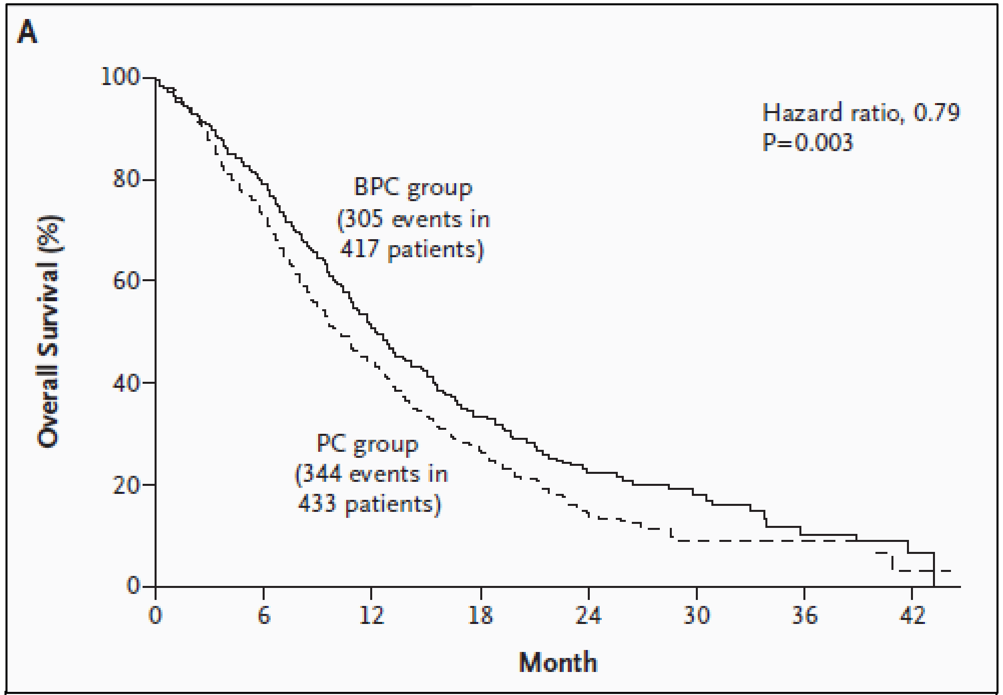




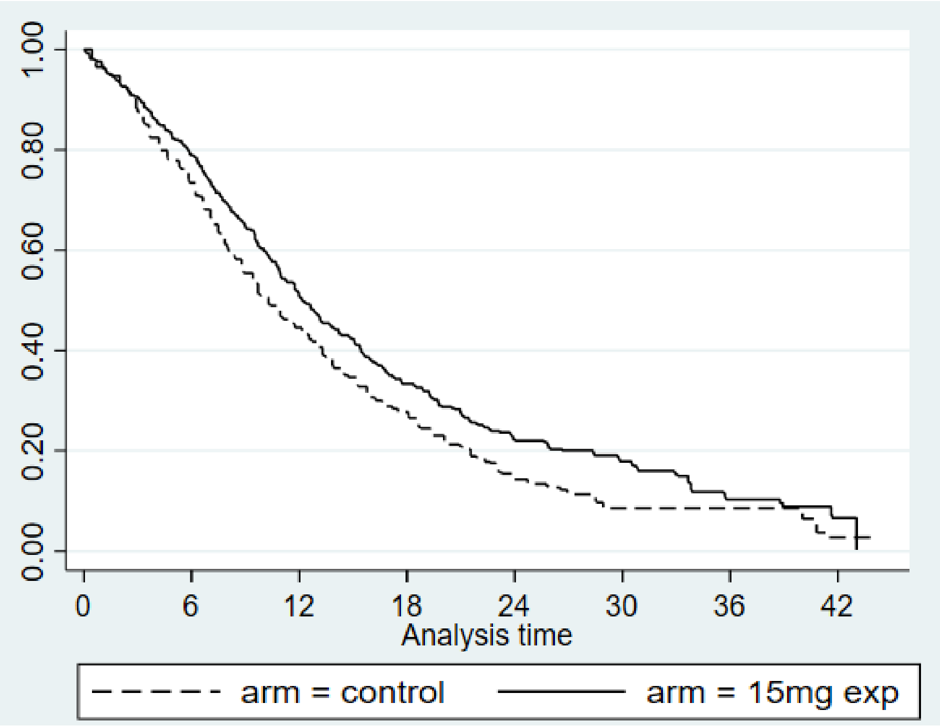
Paper: N/A

Digitized: HR: 1.1, 95% CI 0.66-1.83, p=0.714

Figure 33 - Comparing Kaplan-Meier curves to those from the published phase II paper of pairing #22, and those I generated from the associated digitized dataset.



Paper: HR 0.79, 95% CI 0.67-0.92, p=0.003



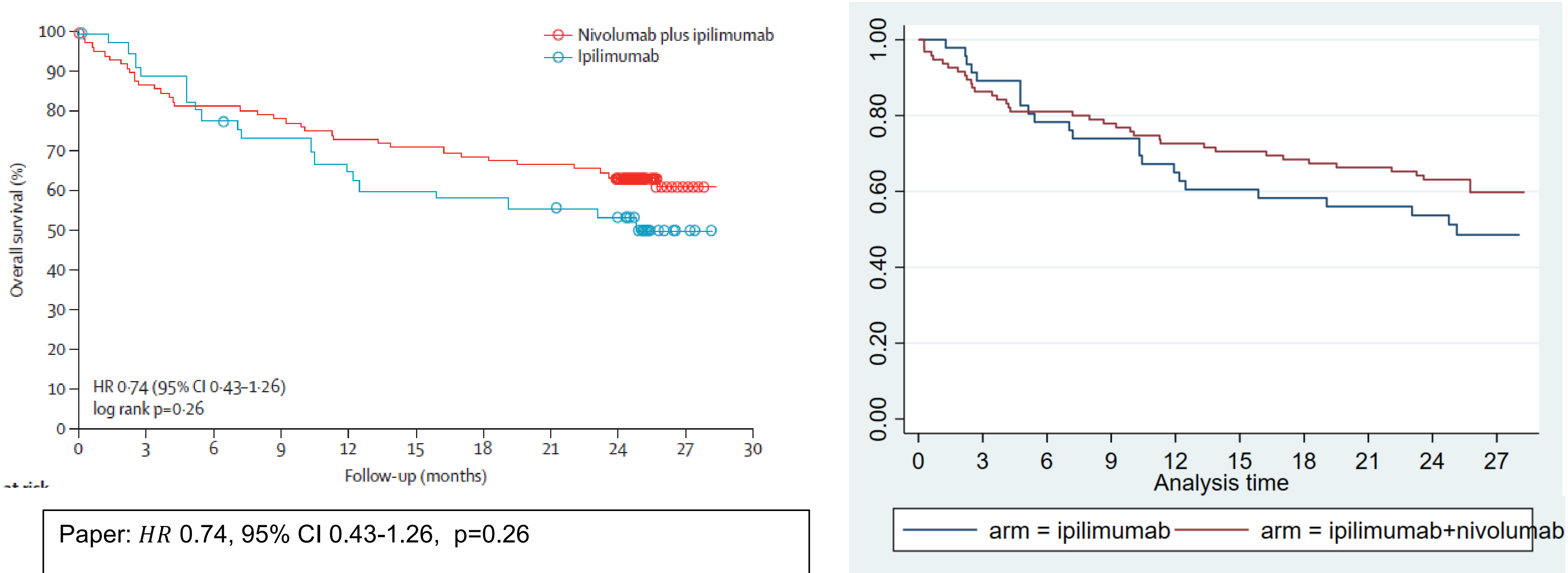
Digitized: HR 0.84, 95% CI 0.73-0.96, p=0.01

Figure 34 - Comparing Kaplan-Meier curves to those from the published phase III paper of pairing #22, and those I generated from the associated digitized dataset.

8.2.4 Pairing #23

Figure 35 and Figure 36 compare the published Kaplan-Meier curves and respective hazard ratios extracted from pairing #23 with the Kaplan-Meier curves and respective hazard ratios obtained from the associated digitized dataset.

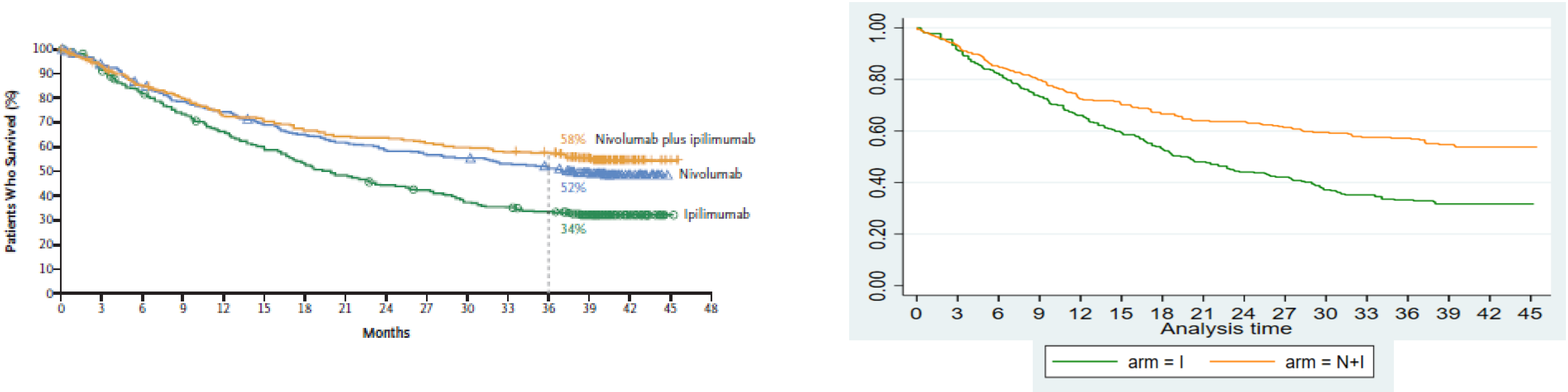
Figure 35 compares the survival outcomes from the phase II setting, and Figure 36 compares the survival outcomes from the phase III setting.



Paper: HR 0.74, 95% CI 0.43-1.26, p=0.26

Digitized: HR 0.74, 95% CI 0.44-1.24, p=0.26

Figure 35 - Comparing Kaplan-Meier curves to those from the published phase II paper of pairing #23, and those I generated from the associated digitized dataset.



Paper: HR 0.55, 98% CI 0.42-0.72, p<0.001

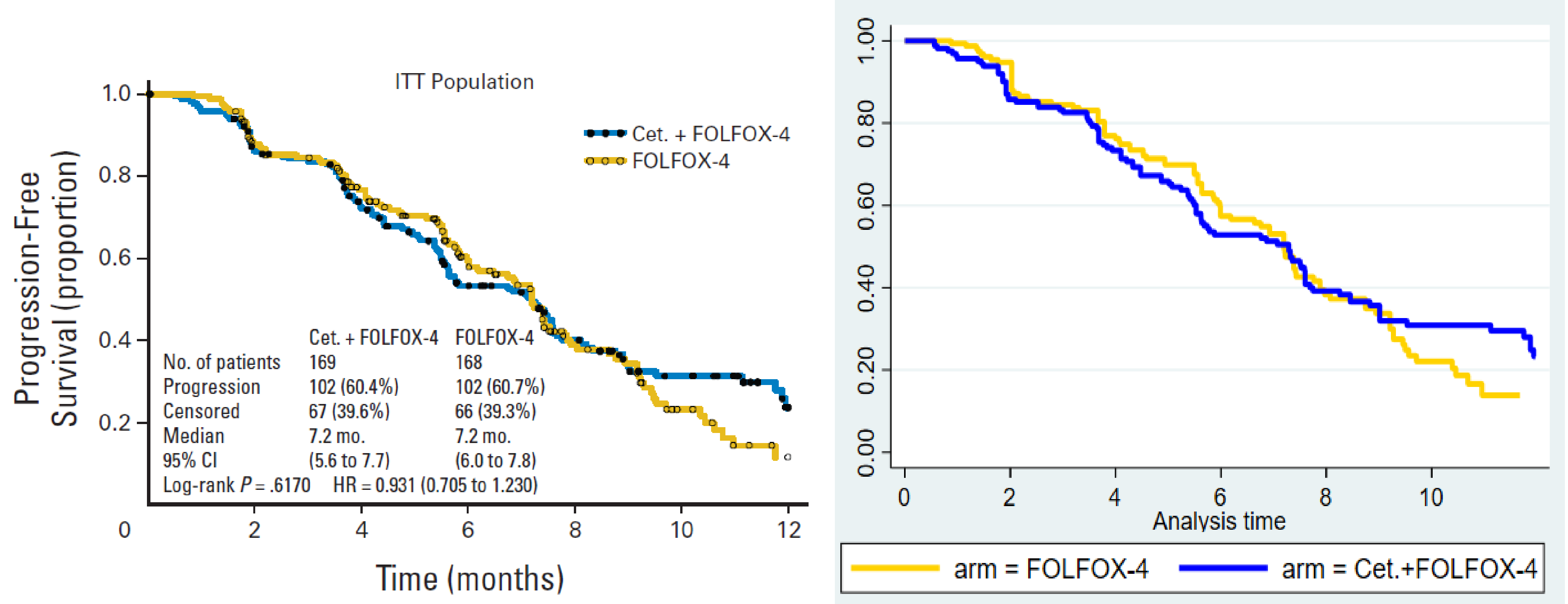
Digitized: HR 0.57, 98% CI 0.44-0.73, p<0.0001

Figure 36 - Comparing Kaplan-Meier curves to those from the published phase III paper of pairing #23, and those I generated from the associated digitized dataset.

8.2.5 Pairing #24

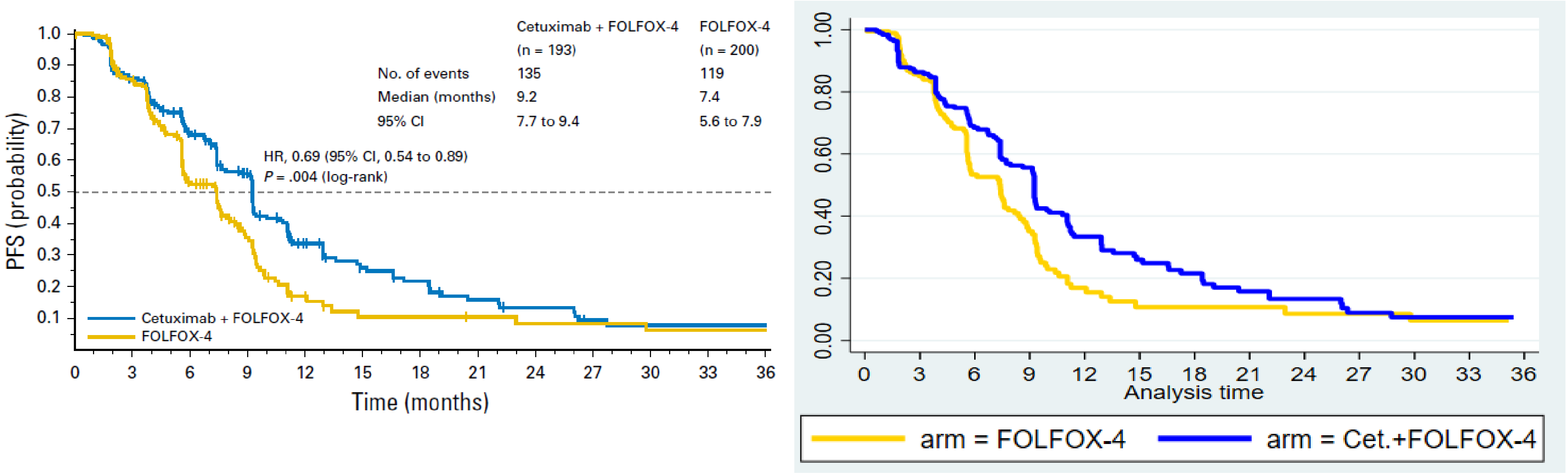
Figure 37 and Figure 38 compare the published Kaplan-Meier curves and respective hazard ratios extracted from pairing #24 with the Kaplan-Meier curves and respective hazard ratios obtained from the associated digitized dataset.

Figure 37 compares the survival outcomes from the phase II setting, and Figure 38 compares the survival outcomes from the phase III setting.



Paper: HR 0.93, 95% CI 0.705-1.23, p=0.617	Digitized: HR 0.93, 95% CI 0.7-1.23, p=0.59

Figure 37 - Comparing Kaplan-Meier curves to those from the published phase II paper of pairing #24, and those I generated from the associated digitized dataset.



Paper: HR 0.69, 95% CI 0.54-0.89, p=0.004

Digitized: HR 0.7, 95% CI 0.55-0.9, p=0.006

Figure 38 - Comparing Kaplan-Meier curves to those from the published phase III paper of pairing #24, and those I generated from the associated digitized dataset.

8.3 Simulated Phase II Relative Risk Versus Simulated Phase III Sample Size

The following graphs (Figure 39, Figure 40, Figure 41, Figure 42, Figure 43) demonstrate how the results of the simulated phase II trials impact the subsequent simulated phase III sample size. Specifically, for simulated development plans with randomised phase II trials. As demonstrated, if a large treatment effect (simulated phase II relative risk) is found in the simulated phase II trial, it is assumed this will be seen in the subsequent trial also, and therefore requires a smaller phase III sample size for a given one-sided α=0.025 and $1-\beta$=0.9.

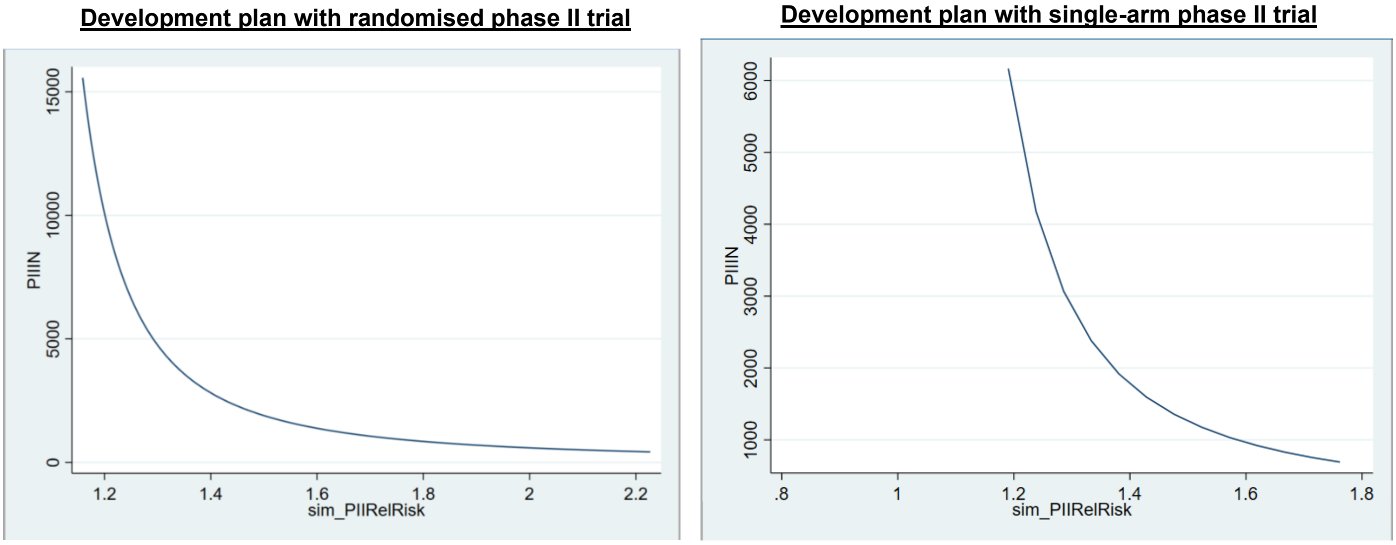
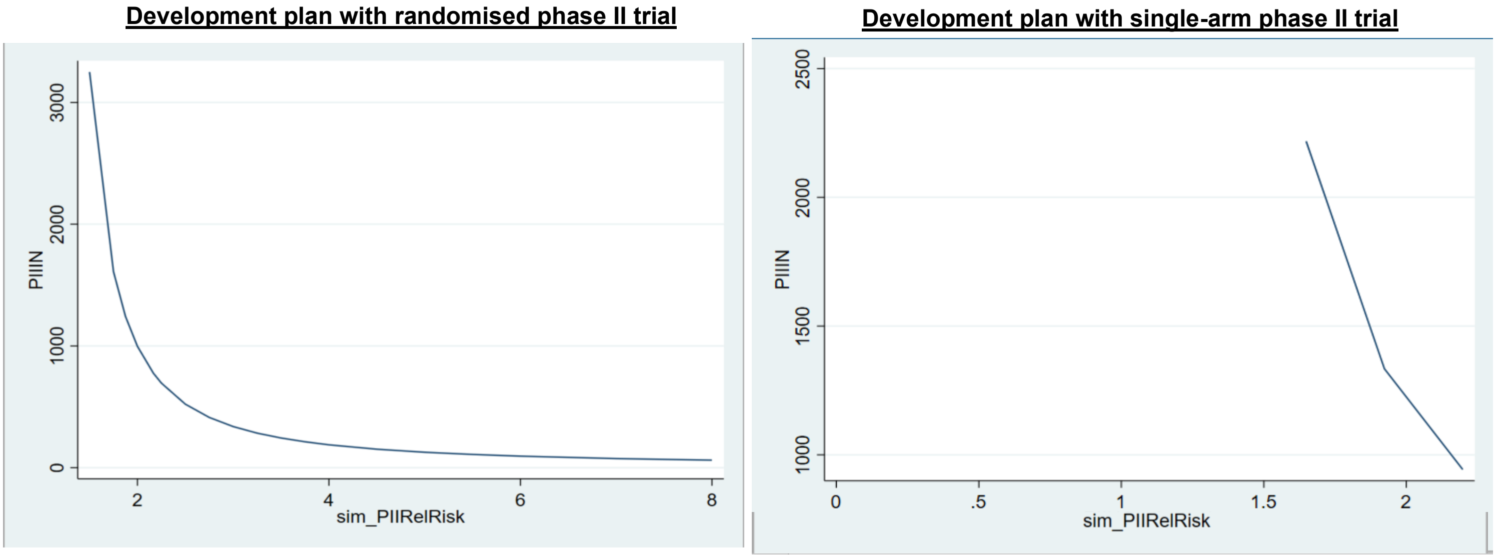


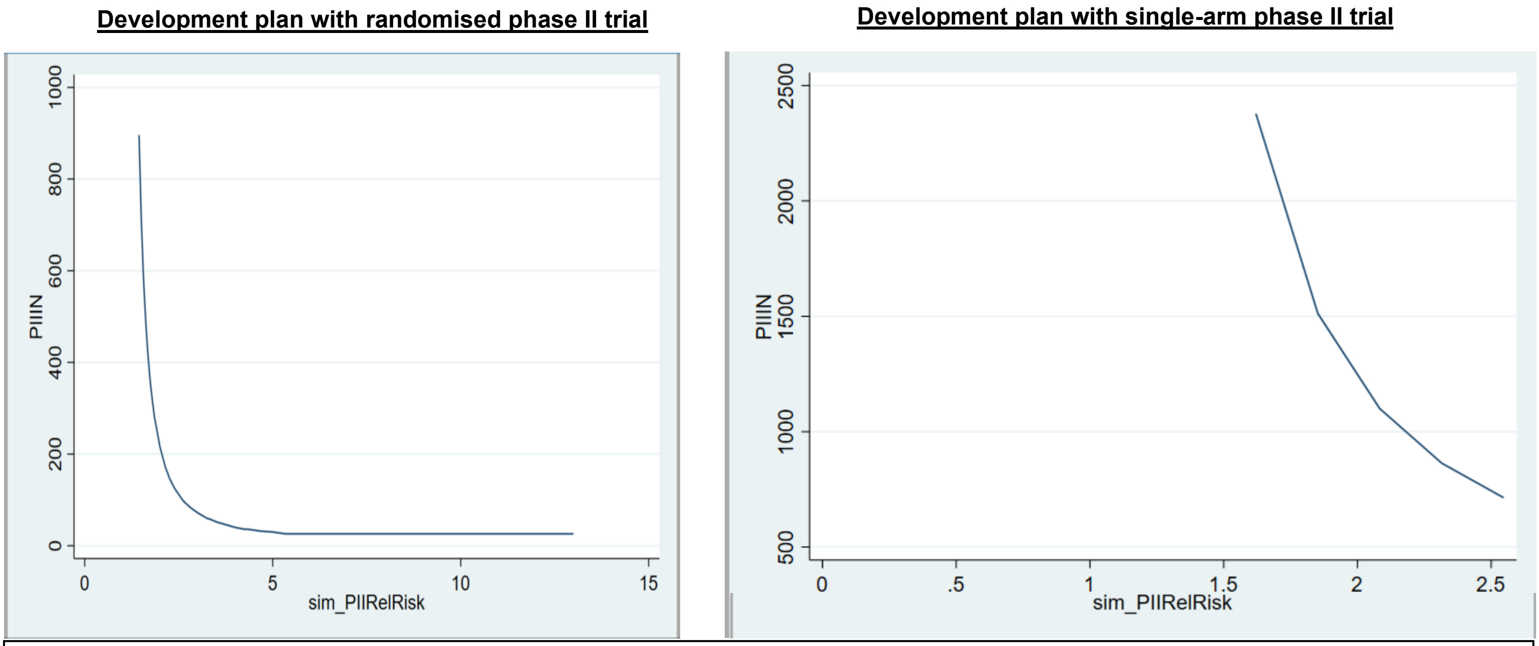
Figure 39 - Simulated development plans in the setting of pairing #3– relationship between phase II simulated relative risk and phase III sample size



Development plan with randomised phase II trial

Development plan with single-arm phase II trial

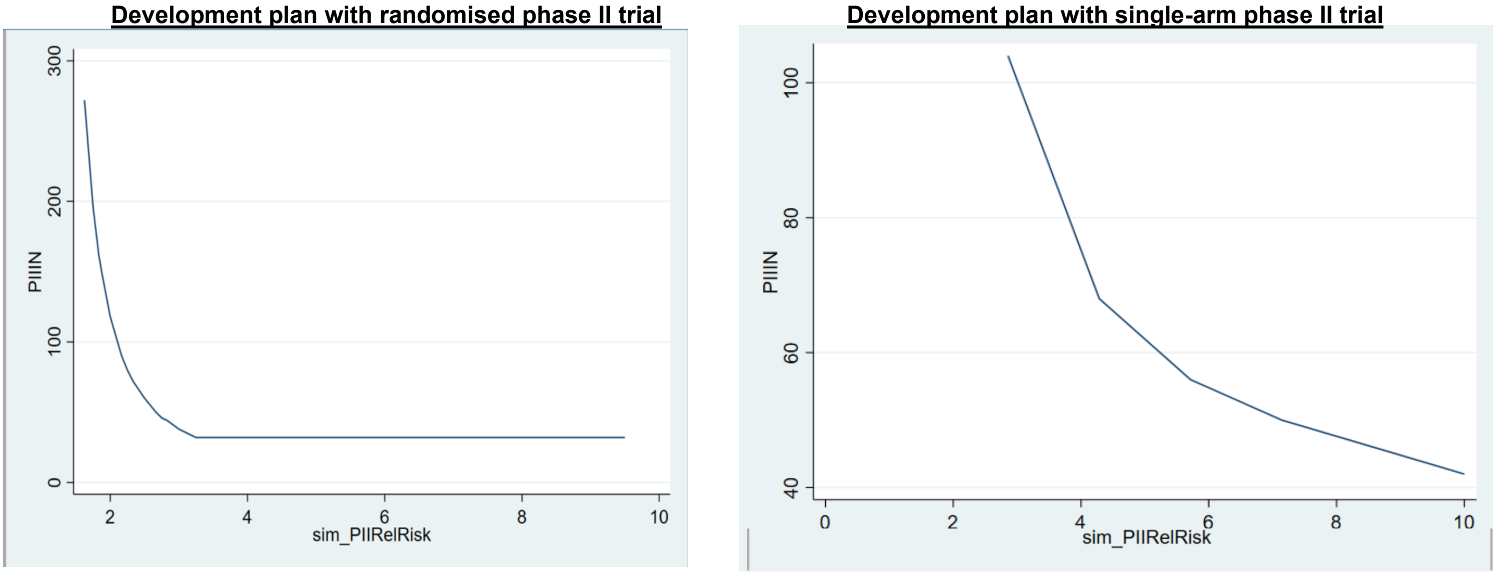
Figure 40 - Simulated development plans in the setting of pairing #21– relationship between phase II simulated relative risk and phase III sample size



Development plan with randomised phase II trial

Development plan with single-arm phase II trial

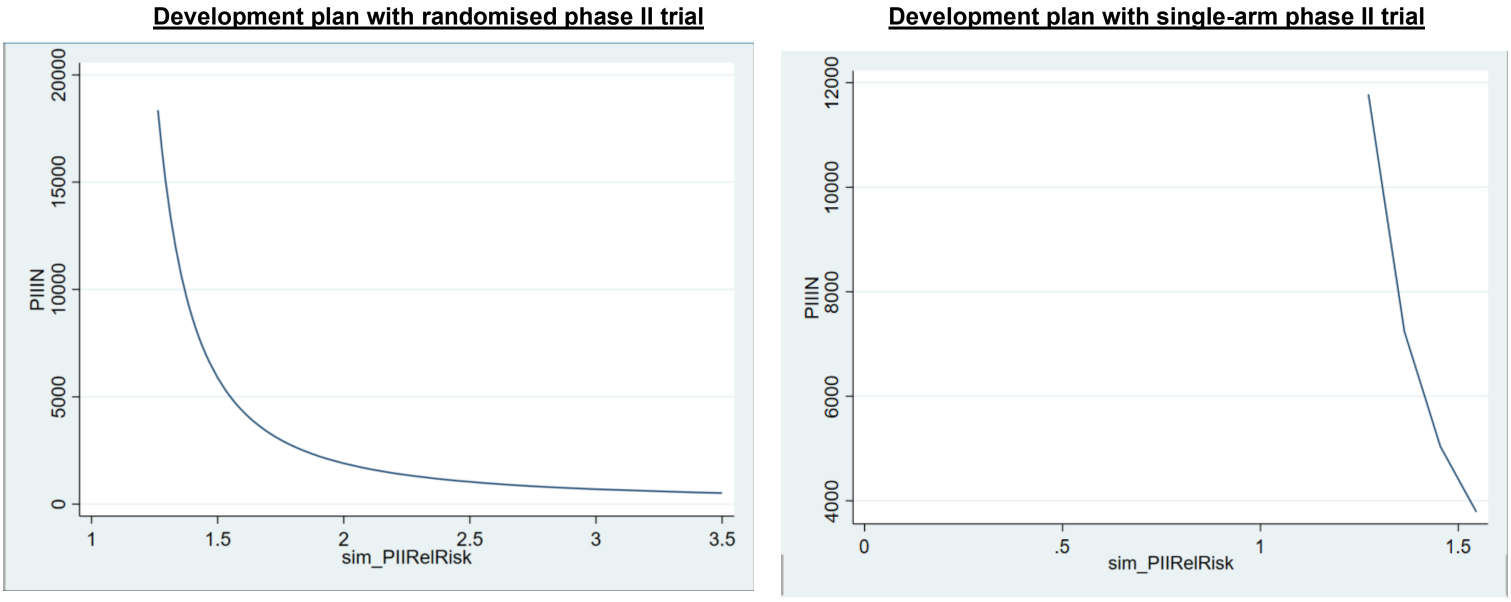
Figure 41 - Simulated development plans in the setting of pairing #22– relationship between phase II simulated relative risk and phase III sample size



Development plan with randomised phase II trial

Development plan with single-arm phase II trial

Figure 42 - Simulated development plans in the setting of pairing #23 – relationship between phase II simulated relative risk and phase III sample size



Development plan with randomised phase II trial

Development plan with single-arm phase II trial

Figure 43 - Simulated development plans in the setting of pairing #24– relationship between phase II simulated relative risk and phase III sample size

8.4 Example code for Chapter 3 simulation study

This is example code from Study 3 presented in Chapter 3. Specifically, this is for a development plan with a randomised phase II trial under the alternate hypothesis. Here, HP_0=0.1, HP_1=0.2 and there is 1%-point of historical control error, where P_0=0.11. 10000 simulations were run, and the proportion of times the development plan successfully concluded in favour of treatment was collected.

```
Clear
//all the variables that are collected for the phase II trial
tempname post1
postfile `post1' i PIIexp_Resp PIIexp_NonResp PIIctr_NonResp
PIIctr_Resp PIIn_arm PIIp PIIdir_favour_exp PIIp_less_alpha
PIItrial_success PIIP0RR PIIP1RR PIIISS PIIIexp_Resp
PIIIexp_NonResp PIIIctr_NonResp PIIIctr_Resp PIIIp
PIIIdir_favour_exp PIIIp_less_alpha PIIItrial_success PIIIP0RR
PIIIP1RR using postfile_PIIPIIIsuccess, replace

//all the variables that are collected for the phase III trial
tempname post2
postfile `post2' i PIIexp_Resp PIIexp_NonResp PIIctr_NonResp
PIIctr_Resp PIIn_arm PIIp PIIdir_favour_exp PIIp_less_alpha
PIItrial_success PIIP0RR PIIP1RR using postfile_PIIPIIIfail,
replace

clear

//phase II sample size calculation
power twoproportions 0.1 0.2, alpha(0.3) beta(0.2)
local N1=`r(N1)'

//phase II trial
set seed 476270
quietly{
     forvalues i=1/10000{
     clear
     set obs `N1'
     gen int id = _n
     //control arm truth
     gen byte outcome0 = rbinomial(1,0.11)
     //experimental arm truth
     gen byte outcome1 = rbinomial(1,0.2)

     reshape long outcome, i(id) j(group)

     //counting number of responders and non-responders in
each arm
     count if outcome==1 & group==1
     local PIIexp_Resp=r(N)
     count if outcome==0 & group==1
     local PIIexp_NonResp=r(N)
     count if outcome==0 & group==0
     local PIIctr_NonResp=r(N)
     count if outcome==1 & group==0
     local PIIctr_Resp=r(N)

     count if group==0
     local PIIn_arm=r(N)

     //likelihood ratio test on proportion of responders in
each arm
     tab outcome group, lrchi2
     return list
     local PIIp=r(p_lr)/2

     //conditions for the phase II trial to continue to a
phase III trial
```

```
     local PIIdir_favour_exp=(`PIIexp_Resp'>`PIIctr_Resp')
     local PIIp_less_alpha=(`PIIp'<0.15)
     local PIItrial_success=(`PIIdir_favour_exp'==1 &
`PIIp_less_alpha'==1)

     //Stata does not allow proportions to equal 0, therefore,
for phase II P0 response rate estimation purposes
     if `PIIctr_Resp'==0{
          local PIIP0RR=0.0001
     }
     else{
          local PIIP0RR=round((`PIIctr_Resp'/`PIIn_arm'),0.01)
     }
     if `PIIexp_Resp'==0{
          local PIIP1RR=0.0001
     }
     else{
          local PIIP1RR=round((`PIIexp_Resp'/`PIIn_arm'),0.01)
     }

     //PIII trial
     if `PIItrial_success'==1 {
          clear
          //phase III sample size calculation based on phase
II estimates
          power twoproportions (`PIIP0RR') (`PIIP1RR'),
test(lrchi2) power(0.9) alpha(0.025) onesided
          return list

          local PIIISS=r(N1)
          set obs `PIIISS'
          gen int id = _n
          //control arm truth
          gen byte outcome0 = rbinomial(1,0.11)
          //experimental arm truth
          gen byte outcome1 = rbinomial(1,0.2)

          reshape long outcome, i(id) j(group)
          tab outcome group, lrchi2

          //counting number of responders and non-responders
in each arm
          count if outcome==1 & group==1
          local PIIIexp_Resp=r(N)
          count if outcome==0 & group==1
          local PIIIexp_NonResp=r(N)
          count if outcome==0 & group==0
          local PIIIctr_NonResp=r(N)
          count if outcome==1 & group==0
          local PIIIctr_Resp=r(N)

          //likelihood ratio test on proportion of responders
     in each arm
          tab outcome group, lrchi2

          return list
          local PIIIp=r(p_lr)/2

          //conditions for phase III trial to conclude in
favour of treatment
          local
PIIIdir_favour_exp=(`PIIIexp_Resp'>`PIIIctr_Resp')
          local PIIIp_less_alpha=(`PIIIp'<0.025)
          local PIIItrial_success=(`PIIIdir_favour_exp'==1 &
`PIIIp_less_alpha'==1)
```

```
          local PIIIP0RR=
round((`PIIIctr_Resp'/`PIIISS'),0.01)
          local PIIIP1RR=
round((`PIIIexp_Resp'/`PIIISS'),0.01)

(`PIIexp_NonResp') (`PIIctr_NonResp') (`PIIctr_Resp')
(`PIIn_arm') (`PIIp') (`PIIdir_favour_exp')
(`PIIp_less_alpha') (`PIItrial_success') (`PIIP0RR')
(`PIIP1RR') (`PIIISS') (`PIIIexp_Resp') (`PIIIexp_NonResp')
(`PIIIctr_NonResp') (`PIIIctr_Resp') (`PIIIp')
(`PIIIdir_favour_exp') (`PIIIp_less_alpha')
(`PIIItrial_success') (`PIIIP0RR') (`PIIIP1RR')
          }

     if `PIItrial_success'==0 {
          post `post2' (`i') (`PIIexp_Resp')
(`PIIexp_NonResp') (`PIIctr_NonResp') (`PIIctr_Resp')
(`PIIn_arm') (`PIIp') (`PIIdir_favour_exp')
(`PIIp_less_alpha') (`PIItrial_success') (`PIIP0RR')
(`PIIP1RR')
          }
     }
}

postclose `post1'
postclose `post2'
use postfile_PIIPIIIsuccess, clear PIIPIIIfPIIPIII
append using postfile_PIIPIIIfail

sort i
//calculate the proportion of times the development plans
concluded in favour of treatment at each stage
tab PIItrial_success
tab PIIItrial_success
```

8.5 Example code for Chapter 4 simulation study

This is example code for simulation study in Chapter 4 with imperfect correspondence. The code ran a phase II trial with binary outcomes P_0=0.4 and P_1=0.7, which represent proportion of survivors at 12-months. If the phase II trial was successful, it moved onto the phase III trial which translated P_0 and P_1 into time-to-event outcomes via the exponential curve. Phase III sample size was based on these translated values; however the true hazard ratio had shifted from phase II to phase III towards the null through an imperfect correspondence factor of 1.5.

This development plan was simulated 10000 times, and the proportion of times the development plan correctly concluded in favour of treatment at phase II and phase III was collected, in addition to the proportion of times the development plan failed to conclude in favour of treatment at phase II and phase III. The X^2-statistics were collected for each of these four outcomes and plotted in a graph.

```
**phase II – P0=0.4, P1=0.7. Sample size one-sided alpha=0.15,
beta=0.8. In this scenario, P0 and P1 are proportion of
survivors at 12-months. Hazard ratio = 0.39**
cd "N:\Old Home Drives\PhD\SIM one-arm vs rct\post
upgrade\Chapter 4\stata results\Study 1, no HC error, no SS
adj\test"
//variables collected in the phase II trial
tempname post
postfile `post' i exp_Resp_DP9P2H1 exp_NonResp_DP9P2H1
ctr_Resp_DP9P2H1 ctr_NonResp_DP9P2H1 PIItstat p_DP9P2H1 using
postfile_DP9P2H1, replace

//phase II sample size calculation
power twoproportions 0.4 0.7, a(0.3) power(0.8)
return list
local PIIN = r(N1)

//simulating phase II trial
set seed 464615
```

```
quietly {
forvalues i=1/10000{

clear
set obs `PIIN'

gen int id = _n
//control arm truth
gen byte outcome0 = rbinomial(1,0.4)
//experimental arm truth
gen byte outcome1 = rbinomial(1,0.7)
reshape long outcome, i(id) j(group)

tab outcome group, column

//counting the number of survivors in each group
count if outcome==1 & group==1
local exp_Resp_DP9P2H1=r(N)
count if outcome==0 & group==1
local exp_NonResp_DP9P2H1=r(N)
count if outcome==0 & group==0
local ctr_NonResp_DP9P2H1=r(N)
count if outcome==1 & group==0
local ctr_Resp_DP9P2H1=r(N)

//performing a likelihood ratio test on the proportion of
survivors in each group
tab outcome group, lrchi2
return list
local PIItstat= r(chi2_lr)
local p_DP9P2H1=r(p_lr)/2
post `post' (`i') (`exp_Resp_DP9P2H1') (`exp_NonResp_DP9P2H1')
(`ctr_Resp_DP9P2H1') (`ctr_NonResp_DP9P2H1') (`PIItstat')
(`p_DP9P2H1')

}
 }

postclose `post'
use postfile_DP9P2H1,clear

//conditions for the phase II trial to proceed to a phase III
trial
gen dir_favour_exp_DP9P2H1=1 if
exp_Resp_DP9P2H1>ctr_Resp_DP9P2H1
gen p_less_alpha_DP9P2H1=1 if p_DP9P2H1<0.15
gen trial_success_DP9P2H1=1 if dir_favour_exp_DP9P2H1==1 &
p_less_alpha_DP9P2H1==1
replace trial_success_DP9P2H1=0 if trial_success_DP9P2H1==.

quietly tab trial_success_DP9P2H1

save "N:\Old Home Drives\PhD\SIM one-arm vs rct\post
upgrade\Chapter 4\stata results\Study 1, no HC error, no SS
adj\test\SC3FABRCTDP9binTTEP2H110kIC.300322.dta", replace

***Phase III trial – Sample size is one-sided alpha=0.025,
beta=0.9. Proportion of survivors at 12-months P0=0.4 and
P1=0.7 are translated into hazard rates using the exponential
curve. Phase III hazard ratio is phase II hazard ratio with
imperfect correspondence factor 1.5, so phase III HR =
0.39*1.5=0.585*
```

```
clear
cd "N:\Old Home Drives\PhD\SIM one-arm vs rct\post
upgrade\Chapter 4\stata results\Study 1, no HC error, no SS
adj\test"

//variables collected in the phase III trial
tempname post
postfile `post' i PIIP0 PIIP1 t_PII t_PIII lambda0 surv0_PIII
lambda1 surv1_PIII hr lnhr PIIIN PIIIg MC_hr MC_lnhr
MC_lambda1 MC_PIIP1 PIIItstat PIIIp PIIIhr using
postfile_DP9P3H1, replace

set seed 464615
//phase III trial
quietly {
     forvalues i=1/10000{
          clear

          local PIIP0=0.4
          local PIIP1=0.7
          local t_PII=12
          local t_PIII=60

          ///calculations to find median0 and median1, when 12
months survival0= 0.4, and 12 months survival1=0.5
          //probability of surviving time, t, is S(t)=e^(-
(lambda*t))
          //therefore, lambda=[-ln(S(t))/t]

          local lambda0=(-ln(`PIIP0')/`t_PII')
          display `lambda0'
          local surv0_PIII = exp(1)^(-(`lambda0'*`t_PIII'))
          display `surv0_PIII'
          local lambda1=(-ln(`PIIP1')/`t_PII')
          display `lambda1'
          local surv1_PIII = exp(1)^(-(`lambda1' *`t_PIII'))
          display `surv1_PIII'

          local hr= `lambda1'/`lambda0'
          local lnhr= ln(`hr')
          display `lambda0'
          display `hr'
          display `lnhr'
          //when P0=0.4 and P1=0.7, HR is 0.39

          //sample sizes based on PII estimates of P0 and P1
          stpower logrank `surv0_PIII' `surv1_PIII',beta(0.1)
          return list
          matrix list r(N)
          matrix M = r(N)
          local PIIIN=M[1,1]
          display `PIIIN'
          local PIIIg=`PIIIN'/2

          //imperfect correspondence between pII and PIII
experimental arm
          local MC_hr=(`hr'*1.5)
          local MC_lnhr=ln(`MC_hr')
                    //MC_PIIP1=S(t)=e^(-(MC_lambda1*t))
                    //as HR=lamda1/lamda0,
MC_lambda1=MC_HR*lambda0
                    //Therefore, MC_PIIP1=e^(-
(MC_HR*lambda0*t))
          display `MC_hr'
```

```
     local MC_lambda1=(`lambda0'*`MC_hr')
     display `MC_lambda1'
     local MC_PIIP1 = exp(1)^(-(`MC_lambda1' *`t_PII'))

     //set number of observations as the phase III 
required sample size
     set obs `PIIIN'

     gen treatment =_n
     gen treatment1=_n
     replace treatment=0 if treatment1<(`PIIIg'+1)
     replace treatment=1 if treatment1>`PIIIg'
     drop treatment1
     //simulate survival data using values calculated 
with the exponential curve
     survsim stime event, dist(exp) lambdas(`lambda0') 
cov(treatment `MC_lnhr' ) maxtime(60)

     stset stime, f(event)

     //COMPARE LOGRANK RESULTS WITH EXPONENTIAL RESULT
     sts test treatment, logrank
     local PIIItstat=r(chi2)
     local PIIIp=chi2tail(r(df), r(chi2))

     //cox test to compare survival curves between 
treatments
     stcox treatment
     matrix list r(table)
     matrix T = r(table)
     display T[1,1]
     local PIIIhr=T[1,1]
```

```
         post `post' (`i') (`PIIP0') (`PIIP1') (`t_PII')
(`t_PIII') (`lambda0') (`surv0_PIII') (`lambda1')
(`surv1_PIII') (`hr') (`lnhr') (`PIIIN') (`PIIIg') (`MC_hr')
(`MC_lnhr') (`MC_lambda1') (`MC_PIIP1') (`PIIItstat')
(`PIIIp') (`PIIIhr')
     }
}

postclose `post'
use postfile_DP9P3H1,clear

//conditions for phase III trial to conclude in favour of
treatment
gen dir_favour_exp_DP9P3H1=1 if `PIIIhr'<1
gen p_less_alpha_DP9P3H1=1 if PIIIp<0.05
gen trial_success_DP9P3H1=1 if dir_favour_exp_DP9P3H1==1 &
p_less_alpha_DP9P3H1==1
replace trial_success_DP9P3H1=0 if trial_success_DP9P3H1==.

tab trial_success_DP9P3H1

save "N:\Old Home Drives\PhD\SIM one-arm vs rct\post
upgrade\Chapter 4\stata results\Study 1, no HC error, no SS
adj\test\SC3FABRCTDP9binTTEP3H110kIC.300322.dta", replace

**MERGE DATA SETS FROM PHASE II AND PHASE III**
use "N:\Old Home Drives\PhD\SIM one-arm vs rct\post
upgrade\Chapter 4\stata results\Study 1, no HC error, no SS
adj\test\SC3FABRCTDP9binTTEP2H110kIC.300322.dta", clear
merge 1:1 i using "N:\Old Home Drives\PhD\SIM one-arm vs
rct\post upgrade\Chapter 4\stata results\Study 1, no HC error,
no SS adj\test\SC3FABRCTDP9binTTEP3H110kIC.300322.dta"
drop _merge
```

```
save "N:\Old Home Drives\PhD\SIM one-arm vs rct\post
upgrade\Chapter 4\stata results\Study 1, no HC error, no SS
adj\test\SC3FABRCTDP9binTTE10kIC.300322.dta", replace

//tabulate trial success between phase II and phase III
tab trial_success_DP9P2H1 trial_success_DP9P3H1
//assess correlation between phase II and phase III test-
statistics
corr PIItstat PIIItstat

//drop any test-statistics that concluded in the wrong
direction (control>experiment)
drop if dir_favour_exp_DP9P2H1!=1 & p_less_alpha_DP9P2H1==1
drop if dir_favour_exp_DP9P3H1!=1 & p_less_alpha_DP9P3H1==1

//count how many test-statistics fall within each quadrant
count if PIItstat<1.074 & PIIItstat<3.84
count if PIItstat<=1.074 & PIIItstat>=3.84
count if PIItstat>=1.074 & PIIItstat<3.84
count if PIItstat>=1.074 & PIIItstat>=3.84

graph twoway scatter PIItstat PIIItstat, title("Phase II chi-2
value against Phase III chi-2 value P0=40%, P1=70%." "(Phase
II a=0.15 1-b=0.8, Phase III a=0.05 1-b=0.9)", size(med))
ylabel(0 5 10 15 20 25 30 35 40 45 50) ytitle("PII chi2 test-
stat", size(small)) yline(1.07) xlabel(0 5 10 15 20 25 30 35
40 45 50) xtitle("PIII chi2 test-stat", size(small))
xline(3.84) aspectratio(1)
```

8.6 Example code for user-written A'hern package in R

Below is example code from the user-written A'hern package in R. In this instance, the two-sided alpha is 0.15, a power of 0.8, and 0.1 and 0.2 representing hypothesised values of P_0 and P_1 respectively.

```
# The possible sample size vector N needs to be selected in
such a fashion that it covers the possible range of values
that include the true minima. My example here does with a
finite range and makes the plot easier to visualize.

rm(list = ls())

N <- 10:100

Alpha <- 0.15
Pow <- 0.80
p0 <- 0.10
p1 <- 0.20

# Required number of events, given a vector of sample sizes
(N)
# to be considered at the null proportion, for the given Alpha
CritVal <- qbinom(p = 1 - Alpha, size = N, prob = p0)

# Get Beta (Type II error) for each N at the alternate
hypothesis
# proportion
Beta <- pbinom(CritVal, N, p1)

# Get the Power
Power <- 1 - Beta
```

```
# Find the smallest sample size yielding at least the required
power
SampSize <- min(which(Power > Pow))

# Get and print the required number of events to reject the
null
# given the sample size 9required
(Res <- paste(CritVal[SampSize] + 1, "out of", N[SampSize]))
head(Res)

output <- data.frame(N=N, power=Power)
head(output)

power_N50 <- subset(output,N==50)
power_N50

power_N74 <- subset(output,N==74)
power_N74

power_N100 <- subset(output,N==100)
power_N100



# Plot it all
plot(N, Power, type = "b", las = 1)

title(paste("One Sided Sample Size and Critical Value for H0
=", p0,
            "versus HA = ", p1, "\n",
            "For Power = ", Pow),
      cex.main = 0.95)

points(N[SampSize], Power[SampSize], col = "red", pch = 19)
```

```
text(N[SampSize], Power[SampSize], col = "red",
     label = Res, pos = 3)

abline(h = Pow, lty = "dashed")
```

8.7 Example code for Chapter 5 simulation study

This is example code for one of the simulation studies in Chapter 5, namely the recreation of pairing #3 with a development plan with a randomised phase II trial. *Data generating mechanisms* are provided through data extraction of the published pairing. The digitized dataset from the published phase II trial is used to provide a time-to-event control arm value for the simulated phase III sample size. The digitized dataset from the published phase III trial is used to sample outcomes for the simulated phase III trial. The simulated development plan is simulated 10000 times, and the proportion of times it successfully concludes in favour of treatment is collected.

```
clear

cd "N:\Old Home Drives\PhD\SIM one-arm vs rct\post
upgrade\Chapter 5\PII and PIII pairings\others  220822\pairing
#3\stata results"

//variables collected for phase II trial
tempname post1
postfile `post1' i PIISS PIIgroup truthP0 truthP1
exp_Resp_RDP3RCTPII exp_NonResp_RDP3RCTPII
ctr_NonResp_RDP3RCTPII ctr_Resp_RDP3RCTPII PIItstat
p_RDP3RCTPII PIIexpSS PIIexpRR PIIctrSS PIIctrRR
truth_PIIRelRisk sim_PIIRelRisk PIIdir_favour_exp
PIIp_less_alpha clinic_mean trial_success_RDP3RCTPII PIIIobs
PIIIP0 PIIIP1 t_PII t_PIII lambda0 surv0_PIII lambda1 surv1_PIII
hr lnhr adapt_surv1_PIII raw_PIIIN raw_PIIIN0 raw_PIIIN1 PIIIN
PIIIN0 PIIIN1 PIIItstat PIIIp PIIIhr PIIIdir_favour_exp
PIIIp_less_alpha PIIItrial_success using
postfile_PIIPIIIsuccess, replace

//variables collected for phase III trial
tempname post2
```

```
postfile `post2' i PIISS PIIgroup truthP0 truthP1
exp_Resp_RDP3RCTPII exp_NonResp_RDP3RCTPII
ctr_NonResp_RDP3RCTPII ctr_Resp_RDP3RCTPII PIItstat
p_RDP3RCTPII PIIexpSS PIIexpRR PIIctrSS PIIctrRR
truth_PIIRelRisk sim_PIIRelRisk PIIdir_favour_exp
PIIp_less_alpha clinic_mean trial_success_RDP3RCTPII using
postfile_PIIPIIIfail, replace

//phase II trial
set seed 211222
quietly {
     forvalues i=1/10000{
          clear

          //phase II sample size calculation with one-sided
     alpha=0.15, and power=0.8. Hypothesised values of P0 and
     P1 extracted from published phase II trial sample size
     calculation
          power twoproportions 0.5 0.65, a(0.3) b(0.2)
          return list

          local PIISS=r(N)
          local PIIgroup=(`PIISS'/2)
          set obs `PIIgroup'

          //control arm truth (control arm estimate from
published phase II trial)
          local truthP0 = 0.46
          //experimental arm truth (experimental arm estimate
from published phase II trial)
          local truthP1 = 0.66

          gen int id = _n
          gen byte outcome0 = rbinomial(1,`truthP0')
          gen byte outcome1 = rbinomial(1,`truthP1')
          reshape long outcome, i(id) j(group)
          tab outcome group, column

          //counting the number of responders and non-
     responders in each group
          count if outcome==1 & group==1
          local exp_Resp_RDP3RCTPII=r(N)
          count if outcome==0 & group==1
          local exp_NonResp_RDP3RCTPII=r(N)
          count if outcome==0 & group==0
          local ctr_NonResp_RDP3RCTPII=r(N)
          count if outcome==1 & group==0
          local ctr_Resp_RDP3RCTPII=r(N)

          //likelihood ratio test on the proportion of
responders in each arm
          tab outcome group, lrchi2
          return list
          local PIItstat= r(chi2_lr)
          local p_RDP3RCTPII=r(p_lr)/2

          //calculating experimental arm response rate
          local PIIexpSS =
(`exp_Resp_RDP3RCTPII'+`exp_NonResp_RDP3RCTPII')
          if `exp_Resp_RDP3RCTPII'>0{
               local PIIexpRR =
(`exp_Resp_RDP3RCTPII'/`PIIexpSS')
               }
          else if `exp_Resp_RDP3RCTPII'==0{
               local PIIexpRR = (1/`PIIexpSS')
               }
          //calculating control arm response rate
          local PIIctrSS =
(`ctr_Resp_RDP3RCTPII'+`ctr_NonResp_RDP3RCTPII')
          if `ctr_Resp_RDP3RCTPII'>0{
               local PIIctrRR =
(`ctr_Resp_RDP3RCTPII'/`PIIctrSS')
               }
          else if `ctr_Resp_RDP3RCTPII'==0{
               local PIIctrRR = (1/`PIIctrSS')
               }

          //calculating the true simulated phase II relative
risk
          local truth_PIIRelRisk = (`truthP1'/`truthP0')
          //calculating the simulated phase II relative risk
          local sim_PIIRelRisk = (`PIIexpRR'/`PIIctrRR')

          //conditions for phase II trial to continue to a
     phase III trial
          local
PIIdir_favour_exp=(`exp_Resp_RDP3RCTPII'>`ctr_Resp_RDP3RCTPII'
)
          local PIIp_less_alpha=(`p_RDP3RCTPII'<0.15)
          local clinic_mean=(`sim_PIIRelRisk'>=1.25)
          local
trial_success_RDP3RCTPII=(`PIIdir_favour_exp'==1 &
`PIIp_less_alpha')
          display `trial_success_RDP3RCTPII'

***Phase III***
          if `trial_success_RDP3RCTPII'==1 {
               clear
               cd "N:\Old Home Drives\PhD\SIM one-arm vs
rct\post upgrade\Chapter 5\PII and PIII pairings\others
220822\OS curves for digitize it\pairing #3"

               //import published phase III digitized dataset
               use "Pairing #3 PIII OS digitized dataset
          130922.dta"

               //count the number of observations in each arm
of the published phase III digitized dataset
               count
               local PIIIobs=r(N)

               count if arm==0
               local PIIIP0obs=r(N)
               count if arm==1
               local PIIIP1obs=r(N)

               clear
               cd "N:\Old Home Drives\PhD\SIM one-arm vs
rct\post upgrade\Chapter 5\PII and PIII pairings\others
220822\OS curves for digitize it\pairing #3"

               //import the published phase II digitized
dataset
               use "Pairing #3 PII OS digitized dataset
          130922.dta"
               //collect the proportion of survivors at the
          end of the published phase II trial
               stset t_ipd, failure(event_ipd)
               sts list, by(arm) risktable(27)
               return list
```

```
                local PIIP0=0.1204
                local PIIP1=0.2462
                local t_PII=27
                local t_PIII=36

                //use proportion of survivors at the end of the
published phase II trial to extrapolate hypothesised values
for proportion of survivors at the end of the phase III trial.
This extrapolation is conducted using the exponential curve,
and results are used for simulated phase III sample size
calculation
                local lambda0=(-ln(`PIIP0')/`t_PII')
                display `lambda0'
                local surv0_PIII = exp(1)^(-
(`lambda0'*`t_PIII'))
                display `surv0_PIII'

                local lambda1=(-ln(`PIIP1')/`t_PII')
                display `lambda1'
                local surv1_PIII = exp(1)^(-(`lambda1'
*`t_PIII'))
                display `surv1_PIII'

                local hr= `lambda1'/`lambda0'
                local lnhr= ln(`hr')
                display `lambda0'
                display `hr'
                display `lnhr'

                //once the survival proportion of the control
arm is obtained, the survival proportion of the experimental
arm is obtain by multiplying the control arm hazard rate by
simulated phase II relative risk
```

```
               local
adapt_surv1_PIII=(`surv0_PIII'*`sim_PIIRelRisk')

               //phase III sample size calculation
               stpower logrank `surv0_PIII'
`adapt_surv1_PIII',beta(0.1)

               return list
               matrix list r(N)
               matrix M = r(N)
               local PIIIN=M[1,1]
               local PIIIN0=M[1,2]
               local PIIIN1=M[1,3]
               display `PIIIN'
               display `PIIIN0'
               display `PIIIN1'

               //if simulated phase III sample size is less
than the number of observations in the published phase III
digitized dataset, simulated phase III trial outcomes are
sampled from the published phase III digitized dataset
               if (`PIIIN0'<=`PIIIP0obs') &
(`PIIIN1'<=`PIIIP1obs') {
                    cd "N:\Old Home Drives\PhD\SIM one-arm vs
rct\post upgrade\Chapter 5\PII and PIII pairings\others
220822\OS curves for digitize it\pairing #3"
                    use "Pairing #3 PIII OS digitized dataset
130922.dta"
                    bsample, strata(arm)
                    tabulate arm
                    stset t_ipd, f(event_ipd)

                    sts test arm, logrank
                    return list
                    local PIIItstat=r(chi2)
                    local PIIIp=chi2tail(r(df), r(chi2))
                    display `PIIIp'

                    stcox arm
                    matrix list r(table)
                    matrix T = r(table)
                    display T[1,1]
                    local PIIIhr=T[1,1]
                    }


               //if simulated phase III sample size is more
than the number of observations in the published phase III
digitized dataset, more survival data needs to be generated
                    else if
               ((`PIIIN0'>`PIIIP0obs')|(`PIIIN1'>`PIIIP1obs'))
               {
                    clear
                    cd "N:\Old Home Drives\PhD\SIM one-arm vs
rct\post upgrade\Chapter 5\PII and PIII pairings\others
220822\OS curves for digitize it\pairing #3"
                    use "Pairing #3 PIII OS digitized dataset
130922.dta"

                    generate ID=_n
                    drop trisk nrisk
                    rename t_ipd t_0
                    rename event_ipd event_0

                    //using merlin to fit Royston-Parmer model
               on published phase III digitized dataset
```

```
                    merlin (t_0 arm arm#rcs(t_0, df(4) log),
family(rp, failure(event_0) df(4)) timevar(t_0))
                    estimates store m1

                    //calculating the number of survival data
blocks or “chunks” need to be generated
                    if ceil((`PIIIN'-`PIIIobs')/`PIIIobs')>1 {
                         local merlinChunk=ceil((`PIIIN'-
`PIIIobs')/`PIIIobs')
                         }
                    //in the original dataset, obs=681 but
P0obs=334 and P1obs=347
                    //in the rare circumstance where PIIIN is
between 668 and 681, then merlinchunk is technically
calculated as 0 as PIIIN<obs. merlinchunk==0 is dropped later
anyway, therefore two chunks need to get generated to garentee
enough sample in each arm
                    else if ceil((`PIIIN'-
`PIIIobs')/`PIIIobs')<=1 {
                         local merlinChunk=2
                         }
                    display `merlinChunk'

                    forvalues j=1/`merlinChunk'{
                              survsim t_`j' event_`j',
maxtime(`t_PIII') model(m1)
                              recast float t_`j', force
                         }

                    reshape long t_ event_, i(ID)
j(merlinChunk)

                    sort merlinChunk
                    drop if merlinChunk==0

                    save "final pairing #3 phase III both arms
reshaped 161223II.dta", replace

                    //sample outcomes from original published
phase III dataset
                    use "Pairing #3 PIII OS digitized dataset
130922.dta"
                    bsample, strata(arm)
                    tabulate arm

                    //append sampled outcomes to generated
survival data
                    gen merlinChunk=0
                    append using "final pairing #3 phase III
both arms reshaped 161223II.dta"
                    count
                    drop if _n >`PIIIN'

                    if t_==. {
                         drop t_
                         rename t_ipd t_
                         }
                    if event_==.{
                         drop event_
                         rename event_ipd event_
                         }

                    stset t_, f(event_)
                    return list
```

```
                     //perform cox-test on sampled outcomes
                from simulated phase III trial

                     sts test arm, logrank
                     return list
                     local PIIItstat=r(chi2)
                     local PIIIp=chi2tail(r(df), r(chi2))
                     display `PIIIp'

                     stcox arm
                     matrix list r(table)
                     matrix T = r(table)
                     display T[1,1]
                     local PIIIhr=T[1,1]
                     display `PIIIhr'
                     }
                //conditions for simulated phase III trial to
conclude in favour of treatment
                local PIIIdir_favour_exp=(`PIIIhr'<1)
                local PIIIp_less_alpha=(`PIIIp'<0.025)
                local
PIIItrial_success=(`PIIIdir_favour_exp'==1 &
`PIIIp_less_alpha'==1)

                cd "N:\Old Home Drives\PhD\SIM one-arm vs
rct\post upgrade\Chapter 5\PII and PIII pairings\others
220822\pairing #3\stata results"



                post `post1' (`i') (`PIISS') (`PIIgroup')
(`truthP0') (`truthP1') (`exp_Resp_RDP3RCTPII')
(`exp_NonResp_RDP3RCTPII') (`ctr_NonResp_RDP3RCTPII')
(`ctr_Resp_RDP3RCTPII') (`PIItstat') (`p_RDP3RCTPII')
```

```
(`PIIexpSS') (`PIIexpRR') (`PIIctrSS') (`PIIctrRR')
(`truth_PIIRelRisk') (`sim_PIIRelRisk') (`PIIdir_favour_exp')
(`PIIp_less_alpha') (`clinic_mean')
(`trial_success_RDP3RCTPII') (`PIIIobs') (`PIIP0') (`PIIP1')
(`t_PII') (`t_PIII') (`lambda0') (`surv0_PIII') (`lambda1')
(`surv1_PIII') (`hr') (`lnhr') (`adapt_surv1_PIII')
(`raw_PIIIN') (`raw_PIIIN0') (`raw_PIIIN1') (`PIIIN')
(`PIIIN0') (`PIIIN1') (`PIIItstat') (`PIIIp') (`PIIIhr')
(`PIIIdir_favour_exp') (`PIIIp_less_alpha')
(`PIIItrial_success')

          }

          else if `trial_success_RDP3RCTPII'==0 {

             cd "N:\Old Home Drives\PhD\SIM one-arm vs
rct\post upgrade\Chapter 5\PII and PIII pairings\others
220822\pairing #3\stata results"
              post `post2' (`i') (`PIISS') (`PIIgroup')
(`truthP0') (`truthP1') (`exp_Resp_RDP3RCTPII')
(`exp_NonResp_RDP3RCTPII') (`ctr_NonResp_RDP3RCTPII')
(`ctr_Resp_RDP3RCTPII') (`PIItstat') (`p_RDP3RCTPII')
(`PIIexpSS') (`PIIexpRR') (`PIIctrSS') (`PIIctrRR')
(`truth_PIIRelRisk') (`sim_PIIRelRisk') (`PIIdir_favour_exp')
(`PIIp_less_alpha') (`clinic_mean')
(`trial_success_RDP3RCTPII')

          }

     }

}
```

```
postclose `post1'
postclose `post2'


use postfile_PIIPIIIsuccess, clear
append using postfile_PIIPIIIfail

//obtain values for simulated phase III sample sizes
sum PIIIN

//obtain proportion of times development plans concluded in
favour of treatment
sort i
tab trial_success_RDP3RCTPII
tab PIIItrial_success


tab trial_success_RDP3RCTPII PIIItrial_success, column row
```