Aristotle University of Thessaloniki
Department of Mathematics
Postgraduate Program in Statistics and
Mathematical Modelling

Statistical Design of Phase II and Phase III Clinical Trials

Evripidis Kapanidis

Supervisor:Assistant Professor Georgios Afendras

July 27, 2022

Statistical Design of Phase II and Phase III Clinical Trials

A thesis submitted for the degree of
Master of Science

Examination Committee:

Assistant Professor Georgios Afendras (Supervisor)

Research Personell-Lecturer Vassilis Karagiannis

Professor Georgios Tsaklidis

July 27, 2022

Declaration of Authorship

I, Kapanidis Evripidis, declare that this thesis titled, Statistical Design of Phase II and Phase III Clinical Trials, and the work presented in it are my own. I confirm that:

- Where any part of this thesis has previously been submitted for a degree or any other qualification at this university or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Abstract

In this thesis we will present statistical methodology for the design and analysis of phase II and phase III clinical trials. The statistical methods presented here are mainly found in the therapeutic area of oncology. Initially, we describe and present fundamental terminology commonly found in clinical trials, the general structure of clinical trials and the relative statistical considerations. Afterwards we study widely used phase II oncology designs. Particularly we present designs that terminate early for futility, designs that compare multiple treatments and inferential methodologies for phase II trial data (estimation of response rates). Furthermore, we present suitable designs for the special cases of heterogeneous populations (that is, when there are multiple subpopulations that have a significantly different response to a treatment) and non-existent historical data. Additionally, for phase III trials we introduce and describe categories of designs that ask different questions of clinical interest (superiority,non-inferiority and equivalence trials) and utilize different endpoints (quantitative,dichotomous and survival endpoints). First we present the main theory for survival endpoints. We then examine the sample size needed for each clinical question and endpoint to acquire proper power and also present methods to determine the duration of phase III clinical trials when survival data is used. Moreover, we examine the methodology and assumptions in group sequential trials as well as the different terminating boundaries that may be used for a group sequential design. Afterwards, we perform simulation studies for both Phase II and Phase III trials. Through the use of simulations we compare different designs for the same hypotheses and examine how the sample size and expected sample size changes. We also compare the distributions of different estimators for the response rate and perform power analyses for superiority, non-inferiority and equivalence phase III trials. Lastly, we analyze real survival data and use the results of this analysis to design a hypothetical future clinical trial.

Περίληψη

Στην παρούσα εργασία θα ασχοληθούμε με μεθόδους σχεδιασμού και ανάλυσης κλινικών δοκιμών φάσης II και φάσης III που συναντώνται στην ογκολογία. Στην αρχή, παρουσιάζουμε βασικές ορολογίες που χρησιμοποιούνται στις κλινι-κές δοκιμές, τη γενικευμένη δομή τους καθώς και τον ρόλο της στατιστικής για τον κατάλληλο σχεδιασμό και ανάλυση τους.Εν συνεχεία, παρουσιάζουμε τους πιο γνωστούς σχεδιασμούς για κλινικές δοκιμές φάσης II στην ογκολογία. Συγκεκριμένα θα παρουσιάσουμε σχεδιασμούς που τερματίζονται νωρίς όταν τα πρώτα αποτελέσματα απο την κλινική δοκιμή υποδεικνύουν ότι η πειραματική θεραπεία δεν έχει κάποιο όφελος, σχεδιασμούς που συγκρίνουν πολλά φάρμακα και μεθοδολογίες εκτίμησης των response rates κατόπιν της κλινικής δοκιμής. Παρουσιάζουμε επίσης σχεδιασμούς για ειδικές περιστάσεις, συγκεκριμένα σχεδιασμούς για ετερογενείς πληθυσμούς και για όταν συναντάται έλλειψη ιστορικών δεδομένων. Στο κεφάλαιο 3 θα παρουσιάσουμε και θα περιγράψουμε σχεδιασμούς που θέτουν διαφορετικές ερωτήσεις κλινικού ενδιαφέροντος (ανωτερότητας, μη-κατωτερότητας, ισοδυναμίας) και που χρησιμοποιούν διαφορετικούς τύπους δεδομένων (ποσοτικά, διχοτομικά και δεδομένα επιβίωσης). Πρώτα παρουσιάζεται η θεωρία που χρησιμοποιείται για δεδομένα επιβίωσης. Έπειτα μελετάμε το μέγεθος δείγματος που χρείαζεται για να έχουμε κατάλληλη ισχύ για κάθε κλινική ερώτηση και τύπο δεδομένου και παρουσιάζουμε μεθοδολογία βάσει της οποίας αποφασίζουμε τη διάρκεια της μελέτης όταν χρησιμοποιούνται δεδομένα επιβίωσης. Τέλος εξετάζουμε τους ακολουθιακούς σχεδιασμούς και τις παραδοχές τους, καθώς και τους διαφορετικούς κανόνες τερματισμού. Κατόπιν, πραγματοποιούμε μελέτες προσομοίωσης. Μέσω των προσομοιώσεων συγκρίνουμε διαφορετικούς σχεδιασμούς και προσεγγίσεις για τις ίδιες υποθέσεις και εξετάζουμε πως το απαιτούμενο δείγμα και το αναμενόμενο δείγμα μεταβάλλονται. Συγκρίνουμε επίσης τις κατανομές εκτιμητριών των response rates και πραγματοποιούμε ανάλυση ισχύος για δοκιμές ανωτερότητας, μή-κατωτερότητας και ισοδυναμίας. Στο τέλος αναλύουμε πραγματικά δεδομένα επιβίωσης και αξιοποιούμε τα αποτελέσματα της ανάλυσης για να σχεδιάσουμε κατάλληλα μια υποθετική μελλοντική κλινική δοκιμή.

Acknowledgements

I would like to thank Dr. Afendras for his precious advice and guidance throughout the preparation of this thesis. I am also grateful to my family for supporting me during my studies.

Contents

Chapter 1

Introduction

1.1 Definition of Clinical Trials

There is no universal definition for clinical trials. Different health agencies or organizations give similar definitions and essentially describe the same process, but ultimately there is some variability in each definition. A strict and detailed definition of a clinical trial is given by the United States' National Institutes of Health (NIH) Agency.

According to the United States' National Institutes of Health (NIH) Agency a clinical trial is defined as "a research study in which one or more human subjects are prospectively assigned to one or more interventions to evaluate the effects of those interventions on health related biomedical or behavioral outcomes" [1].

The term intervention is defined as a manipulation of the subject or subject's environment for the purpose of modifying one or more health-related biomedical or behavioral processes and/or endpoints [2]. A simplified definition, is that the term intervention applies to any activity undertaken with the objective of improving human health by preventing disease, by curing or reducing the severity or duration of an existing disease, or by restoring function lost through disease or injury [3]. There are two main types of interventions:

- Preventive interventions are those that prevent disease from occurring and thus reduce the incidence (new cases) of disease. Common examples are: vaccines, nutritional interventions and preventive drugs.
- Therapeutic interventions. Common examples are: surgical treatments, radiation treatments and treatment for the control of chronic diseases.

The term prospectively assigned refers to a pre-defined process that stipulates the assignment of each subject to a specific intervention or control group. A control group is the standard to which comparisons are made in a research study [4].

With these definitions, it is apparent that clinical trials are a subset of the whole spectrum of research studies and that in order to consider a research study as a clinical trial, it is essential that the participants are human subjects, that they are assigned to an intervention with a specific process during the study and that the final objective of the study is to assess the interventions' effectiveness based on biomedical or behavioral outcomes.

1.1.1 Subtypes of Clinical Trials

Clinical trials can be broken into different sub-categories/types. Types of clinical trials are (according to NIH):

- Prevention trials explore interventions to prevent a disease in people who have never had the disease or to prevent the disease from returning. Approaches may include medicines, vaccines, or lifestyle changes.
- Screening trials test new ways for detecting diseases or health conditions.
- Treatment trials test new treatments and possibly compare them with standard ones.
- Behavioral trials evaluate or compare ways to promote behavioral changes designed to improve health.
- Quality of life trials (or supportive care trials) explore and measure ways to improve the comfort and quality of life of people with conditions or illnesses.

1.1.2 Organizational Structure of Clinical Trials

The design, conduct and analysis of clinical trials is a multidimensional task that involves a variety of scientific disciples, organizations and committees. The size of the research team depends on the funding and the nature of the research performed but common categories of the study staff are [5,6]:

- Principal Investigators: Responsible for all clinical research activities during the trial. A PI generally oversees every activity, from the overall design to the practical management of the clinical trial.

- Biostatisticians: The biostatisticians perform duties relevant to their expertise during the design, conduct and analysis of clinical trials. Examples of these are writing statistical analysis plans, reviewing the data collection methods and performing data cleaning. More detailed descriptions of particular roles and contributions of statisticians to clinical trial designs is presented below .

- Study/Clinical Trial Coordinator: Coordinates the daily clinical trial activities and plays a critical role in the conduct of the study.

- Clinical Trials Supply Management : Responsible for the accurate and timely supply of drugs to patient sites.

Additionally a clinical trial is financially supported by public or private organizations (pharmaceuticals, government organizations etc). Lastly, clinical trials are monitored by special committees. This happens to ensure ethical and safety requirements regarding the patients/participants. Something to note, is that as the definition of clinical trials differ in literature the names of the monitoring committees might as well differ in each country, even if they have the same function and duties. Also, different monitoring committees might be requested depending on the clinical trial [6]. Usual monitoring committees can be (quoting [21]):

- The ethics committee: They are constituted of medical and nonmedical members and are mandatory in all clinical trials in human subjects. The responsibility of an Ethics Committee is to ensure the protection of the rights, safety and overall condition of the participants in clinical trials.

- Data monitoring Committee: A Data Monitoring Committee is a group of independent experts external to a study assessing mainly the progress and safety data of a clinical study.

- Steering Committee: Usually these committees are appointed by the sponsor and comprise of investigators, (sometimes) clinical experts not directly involved in the clinical trial and staff from the sponsor. Among others a Steering Committee often takes responsibility for the scientific validity of the study protocol, assessment of study quality and conduct as well as for the scientific quality of the final study report.

1.1.3 Phases of Clinical Trials

Before testing of a potential drug on human beings, laboratory tests are performed on animals to assess if there are favorable results. If there are indeed positive results, the

therapy is approved to be applied to human beings.

In order for a trial to be approved, a research protocol must also be written. A research protocol is the detailed written plan for a clinical trial. It details every part and aspect of the study plan. In essence, trial protocols are documents that describe the general parameters and statistical considerations related to the organization of clinical trials [12]. Common points that are and should be mentioned in research protocols are described in [13]. A research protocol is in most cases approximately 50 pages.

Clinical trials advance through four phases (Phase I-IV) but in some cases there is an additional trial called phase 0 that is performed before the other four trials. The questions asked in each phase are different and as a result the methodology used is also different. The objectives of each phase are described below:

- Phase 0 Trials: After a drug is approved to be tested in humans a phase 0 trial might take place. A phase 0 trial is an exploratory trial, where the candidate therapy is used on a limited number of patients and the dosages administered do not possess any therapeutic effect. In this phase the goal is to assess if a mechanism of action that was observed in pre-clinical /laboratory testing can also be observed in human subjects. If there are multiple possible treatments that have been approved to be tested in humans, a Phase 0 trial can help select the most promising one. Phase 0 is important as a screening tool for scientists, contributing to the early elimination of therapies that have poor PK/PD properties (see the next paragraph, for the definition) and do not show the same promising results as they did in pre-clinical testing. They also contribute in recognizing essential PK/PD properties for promising therapies before moving to larger sample phases [10].

- Phase I Trials: In many cases, they are the first step of testing a new drug in human beings. Its main goal is to ensure that a treatment is safe for people to take and additionally measure how well the agent can be tolerated by the human organism[8]. For example, in a Phase I study the maximum dosage that can be given to a patient before they start experiencing heavy adverse effects is studied.

- Phase II trials: This phase is designed to assess whether the drug administered has any biological activity that is beneficial to the patient [16]. In essence, a Phase II trial decides if the drug has any therapeutic effect and whether to proceed to a larger trial that will confirm and quantify this effect [8]. Simultaneously, Phase I assessments are continued. Phase II trials play a significant role in assessing if it

is worth to invest more time as well as financial and physical resources to study the effect of a specific drug, as it is the phase where initial evidence of the drug being effective in humans should appear.

- Phase III trials: If a potential benefit regarding treatment effectiveness is revealed in a phase II study, a phase III study follows [10]. In this phase large amounts of data are collected under a long period of follow up to assess in a decisive way the therapeutic effect of the drug [8]. As a result, Phase III trials are the definitive assessment of the efficiency of a new drug. If a Phase III trial is successful, the drug will acquire marketing authorization (i.e.it will become available to the pubic).

- Phase IV trials: After a drug gains market authorization a phase IV trial may begin. The purpose of a phase IV trial is to continue to monitor the patients' safety and the occurrence of the drug related side effects, identifying problems that have not been recognized in the previous phases [8]. Phase IV trials, can help recognize side effects that previous trials did not (due to their more limited duration) and also identify subgroups of people that may be more susceptible to observed side effects due to a number of factors that were not taken into account in previous trials. During Phase IV trials, rare side effects of drugs have been discovered and in some cases drugs have even been withdrawn from the market [8].

1.2 The Statistical Viewpoint of Clinical Trials

Even though the methodology used for clinical trials varies for each trial there is some common terminology used in almost every clinical trial. Specifically quoting [8]:

- RANDOMIZATION: A method in which study participants are assigned to a treatment group or, as mentioned before, to a specific intervention or control group. This way, the decision about which treatment someone will receive is based on chance. Randomization is the best way of ensuring that the results of trials are not biased. For example, in a non-randomized setting a physician might assign young and healthy participants to a new treatment, and others to a standard treatment. A possible result would be that the new treatment is more effective to the standard one but this result would not be trustworthy because there would be selection bias.

- BLINDING: Blinding means that whoever is receiving or assessing the effects of treatment does not know which treatment the person has received. This helps

to prevent bias. If a patient knows that he is receiving a treatment instead of a placebo than he believes he is doing better and if researchers know who is receiving the treatment then they might overestimate the patient's response, thus creating bias in both sides. There are specific categories of blinding:

  - A Double Blinding clinical trial is a trial design in which neither the participating individuals nor the study staff knows which participants are receiving the experimental drug and which are receiving a placebo (or another therapy).
  - A Single Blinding clinical trial is a trial design in which only the participating individuals do not know which treatment they are receiving.
  - A Triple Blinding clinical trial is a trial design where not only the investigators and the participants do not know what treatment they are given, but also the staff responsible for the data analysis is not aware. Generally triple blinding clinical trials are not common and the most common blinding is double blinding.

- ENDPOINT: The primary objective of a trial is to address a scientific question by collecting appropriate data. The selection of the primary endpoint is made to address the primary objective of the trial. In essence an endpoint is the overall outcome that the protocol is designed to evaluate [9]. The primary end-point should be clinically relevant, interpretable and affordable to measure.

  Endpoints can generally be categorized by their scale of measurement. The three most common types of endpoints in clinical trials are continuous endpoints (e.g., pain on a visual analogue scale), categorical (including binary, e.g., response vs. no response) endpoints, and event-time endpoints (e.g., time to death). An endpoint that is a binary categorical variable is alco known as a dichotomous endpoint.

  In clinical studies where endpoints are complex to assess and/or include subjective components or the study cannot be blinded, an Endpoint Adjudication Committee, consisting of clinical experts in a specific clinical area, might be set up to assess possible endpoints [6].

- ARM: A group or subgroup of participants in a clinical trial that receives a specific intervention/treatment, or no intervention, according to the trial's protocol.

For example three arms may mean three different groups receiving different interventions.

1.2.1 General Methodology in Phases of Clinical Trials

Each phase of a clinical trial presents a different objective. Thus, different methodology is used in each phase:

- Phase 0 Trials: In order to assess if a mechanism of action will appear in humans as it did in animals, pharmacokinetics (PK) and pharmacodynamics (PD) analysis is performed. Pharmacokinetics, describes how the body affects a specific xenobiotic/chemical after administration through the mechanisms of absorption and distribution, as well as the metabolic changes of the substance in the body whereas pharmacodynamics describes the biochemical and physiologic effects of the drug (especially pharmaceutical drugs) [15].

- Phase I Trials: Methodology in Phase I trials, includes mathematical modelling of the pharmacokinetics and pharmacodynamics of the drug administered. Sample size in Phase I trials is small, usually 20 to 80 healthy subjects are enrolled. The subjects are grouped depending on their date of enrollment and are treated in cohorts. For safety reasons, the first cohort is given the lowest specified dose of the treatment. Dose escalation or de-escalation is based on whether for a specific dose side effects are observed. When dose escalation reaches a point where intolerable side effects appear than the drug is said to have reached its theoretical Maximum Tolerated Dose. Correct specification of the MTD is of major importance for the following clinical trials, because the MTD will be studied in the next phases. Assuming a larger dosage of MTD will expose subjects to toxic-doses and damage their organisms while assuming a smaller dosage might lead to no therapeutic effect for patients (whereas correct specification would make the drug effective). It is noteworthy, that during this phase there is no research related to the therapeutic efficiency of the drug [8].

- Phase II trials: The Methodology in phase II clinical trials, might contain randomizing patients to different treatment arms. For example, if a new treatment is to be tested patients may be randomized to two arms, one where patients are given the new treatment and one where patients are given the standard one. Phase II might as well be single armed assessing only if there is a statistically significant response/improvement of patients' status after administering the drug. Pharmacokinetic analysis might be performed on the participants [14]. Sample

size in Phase II trials is larger, than that of Phase I trials. Typical endpoints in Phase II clinical trials are mostly dichotomous, characterizing the response of a patient to the treatment (e.g. defining a tumor decreasing by 30 percent as a response to treatment and perform hypothesis testing to assess if the percentage of responses is significantly improved using the treatment) [8].

- Phase III trials: The standard methodology is designing a double-blind, randomized placebo-controlled study. In many cases, Phase III trials are not only two-armed but multi-armed and the new treatment is compared ,commonly through hypothesis testing, to multiple standard ones. Actually, placebo controlled arms are mainly used when no other available treatments exist [8]. The typical endpoint in Phase III is usually a time-to-event measurement. Phase III trials have a larger sample than phase I and II trials (hundreds to thousands of participants) [10].

- Phase IV trials, are mostly observational and longitudinal studies. Common Phase IV trials include cross-sectional studies, case control studies and cohort studies [11] . All these studies are observational without interventions and even though there is a difference of methodology between them, their ultimate goal is to assess risk factors for someone developing the observed side effects or assess if the drug is a significant risk factor for a side effect/heath problem not observed in previous trials.

1.2.2 Statistical Issues in Phase II and III Designs

The contribution of statistical methodology to the success of a clinical trial is essential and that is recognized by The International Council for Harmonization (ICH) E9 statistical principles for clinical trials guideline [24, page 4]. Multiple clinical trial design issues that should be addressed by the biostatistician including:

- The clinical trial design configuration: Dependent on factors such as, the clinical question that is asked, participants might be randomized to different treatment groups via a randomization process (parallel group design), or they might be randomized to a sequence of two or more treatments (crossover design) or they might be randomized to be given one, none or two simultaneous treatments (factorial designs). Strong knowledge of statistics is essential to apply the correct statistical model as well as recognize potential problems that might occur in each configuration[24].

- Determining the sample size: It is essential that the sample size in a clinical trial is large enough to provide the answers to the questions asked. Specifically, in a case where a significant difference exists between the control and experimental group the probability that the null hypothesis will be rejected (power of the test) must be large. As a result, Knowledge of statistical inference and asymptotic statistics are important for the derivation of power formulas for test statistics[24]. Additionally, in many clinical trials multiple endpoints are tested at different stages. Thus, due to multiple testing, methods to avoid type 1 error are of interest. Considerations for the sample size and the power and type I error are also present for single arm designs as we will see.

At this point it is also worth mentioning that statistical methodology in clinical trials is connected to other principles of clinical trials, for example ethical principles. Particularly according to the Belmont report [16], during a clinical trial it is the investigators duty to avoid side effects. During some clinical trials it is possible that adverse side effects might appear. Additionally it is also ethical to terminate a trial when during its course it seems that the experimental treatment does not have any efficacy. Thus it is the statistician's duty to design a clinical trial that will terminate early, if negative side effects appear or no efficacy is observed, and simultaneously derive statistical methodology in order to preserve/satisfy important statistical considerations in such trials (e.g. power).

For phase II/III trials there is a number of different trial designs, that contain different methodology. Statistically, we may group the methodology based on two classifications: Frequentist/Bayesian Designs and Adaptive/Non-Adaptive Designs.

One basic difference between Frequentist and Bayesian statistics is the definition of the prior distribution. In Bayesian statistical inference the prior distribution represents the information/distribution about an unknown population parameter (whereas in frequentist statistics the unknown parameter is considered as a fixed value) [23]. Based on this definition the prior distribution represents our ideas, beliefs, and past experiences about the parameter before we collect and use the data. Then, using the Bayes formula we can compute the distribution of the posterior, i.e. the distribution of the unknown parameter after collecting information (data) from the population.

In a clinical trial, bayesian statistics can contribute to its design because in this case there are two sources of information. One source it the data acquired from the participants (sample data, as in the case of frequentist methodology). The second additional

source is the prior distribution, representing the additional (external) information that is available. Particularly, in a clinical trial design setting the prior distribution is usually based on data from previous trials. As a result It is essential that appropriate prior information is carefully selected and incorporated into the analysis correctly. It is recommended that as many sources of good prior information as possible should be identified. On the other hand, frequentist statistics are more straightforward, because parameters are considered fixed during the analysis of data and there is no need to research for prior information in such detail as a Bayesian framework demands. Additionally, the mathematical models used are less complex in the frequentist framework [17].

A non-adaptive design refers to the most traditional way of running a clinical trial which entails [18]:

- Designing the trial.
- Conducting the trial as prescribed by the design.
- After acquiring the data, perform statistical analysis based on a predefined analysis plan.

Thus, a non-adaptive design is straightforward but inflexible, as changes that may be desirable during the study are not feasible. An adaptive design is defined as a design that allows modifications to the trial and/or statistical procedures of the trial after its initiation without undermining its validity and integrity [18]. Consequently, adaptive designs provide a solution to the inflexibility of traditional/classical trials. It is noteworthy, that flexibility here does not mean that the trial can be modified any time at will. The modification and adaptations have to be pre-planned and should be based on data collected from the study itself [19]. Analysis of data that is collected up to a certain point during a clinical trial is called interim analysis. In other words, the adaptations are based on the results of an interim data analysis. Interim analyses of the accumulating study data are performed at pre-planned timepoints within the study [19].

Yet, problems with adaptive designs might arise. For example an issue with adaptive designs is the additional complexity that they cause to the conduct of a clinical trial .To provide an advantage over a non-adaptive design, interim analyses must be conducted quickly and to a high standard. This involves having an effective infrastructure within the trial team that may require considerable investment of resources [20].

Another problem is that many adaptative procedures rely on bayesian framworks and many researchers still consider the Bayesian statistical methods as non-standard [19].

Regarding the above considerations and phase II oncology trials, significant frequentist contributions have been made by Gehan, Fleming, Simon and Jung. Gehan was one of the first to design Clinical trials including an interim analysis that allowed the trial to terminate early if only a small number of patients responded positively at the time of the interim analysis. Fleming and Therneau continued in this direction, until Simon in 1989 proposed a two stage design that became a very popular and standard design in oncology trials. Simon also proposed a great number of other innovative Phase II designs, that are also used in today's Phase II trials. Jung also proposed methods to estimate response rates in clinical trials, as well as designs for cases such as heterogeneous populations and randomized comparisons.

Regarding Phase III trials, many oncology trials are using time to event data to assess survivability of patients. Significant contributions were made by Kaplan and Meier (Kaplan and Meier Survival function estimator), Mantel and Haenszel (Log-Rank test) David Cox (Cox Proportional hazards model) ,Therneau and Grambsch (hypothesis and graphical tests of proportionality), Schoenfeld (Schoenfeld Residuals, asymptotic distribution of Log-rank test under the alternative hypothesis) and Tsiatis (consistency of partial MLE estimates, markov property of group sequential statistics) in terms of analyzing survival data and drawing inference, as well as asymptotic results. Jenisson and Turnbull, have also made significant contributions regarding group sequential designs.

1.3 Structure of this Thesis

In chapter 2 we will present a number of frequentist Phase II designs, both signle and multi armed. We will present popular designs, such as Simon's two stage designs and the Simon's Pick the winner designs. We will also provide proofs and mathematical justification for the formulas used in these designs. We will additionally present Jung's designs for heterogeneous designs and randomized Phase II experiments. Lastly, we will provide a proof for Jung's UMVUE of the response rate in Phase II clinical trials.

In chapter 3 we will present superiority,non-inferiority and equivalence phase III trials for quantitative, time to event and dichotomous outcomes. We study the basic theory used in survival analysis (Kaplan-Meier, Cox proportional hazards, Log-rank test, Diagnostic Residuals in Cox's Model). In the last section We also study group

sequential designs for different endpoints and terminating boundaries.

In chapter 4 we perform simulations for Simon's two stage design under different power and type I error constraints as well as different null and alternative hypotheses.We also perform simulations for the pick the winner design. We compare the distributions of the UMVUE and the MLE of the response rates in phase II trials. Last but not least, for phase III trials we perform power analyses for different types of designs and compute and compare the terminating boundaries for different group sequential designs.We will also perform survival data analysis on real hisotrical data. The purpose of this analysis is to demonstrate how to analyze survival data and how to design a future clinical trial in the same field.

Chapter 2

Phase II Trials

2.1 Single Arm Designs

As it was mentioned in chapter 1, the purpose of a phase II clinical trial is to assess the short term efficacy of an intervention. Thus, a phase II trial is expected to be carried out fast and the clinical response should be measurable within a short term of time following treatment. The clinical response in phase II trials is in most cases dichotomous and is expressed as a proportion (e.g. percent of patients that experienced tumor decrease when given an experimental drug).

It is also the case that most phase II trials are single arm designs, i.e.there is not a placebo arm and all participants receive the experimental intervention. In a single arm design, where a new experimental drug is to be tested, the hypothesis is formulated as follows:

$$H_0 : p = p_0$$

$$H_1 : p = p_1$$

The responses p_0, p_1 are not defined arbitrarily based on what the investigator considers suitable. The null hypothesis proportion is actually a historical control. Historical control refers to the practice of using data from past studies to estimate potential response to placebo or the standard treatment among patients in an ongoing study [23]. For example, collecting historical data from the placebo groups of randomized controlled trials to assess a specific cancer type can give an estimate of the true proportion of patients' responses when they receive the dummy drug. At that point, it is evident that to create an accurate and precise historical control, historical data from studies that are as similar as possible to the patients being enrolled in the study of

interest is needed. If proper methodology is not used historical controls can become a significant source of bias [23].

Based on the previous definitions in simpler terms ,p_0 can be considered as the largest response proportion that, if true, clearly implies that the treatment does not warrant further study. For the alternative hypothesis, [p_1 can be considered as the smallest response proportion that, if true, clearly implies that the treatment does warrant further study. Both values can be defined based on historical data, when it exists.

When historical data exists and a single arm design is feasible, there are many advantages from designing such a trial. The sample size needed is obviously less of what a trial would require in the case of a two-arm randomized phase II trial. With less participants needed to draw inference for a drug, results are faster and the financial burdens of running the clinical trial are mitigated. In the next sub-section the most widely used single arm designs in phase II clinical trials are presented.

2.1.1 Simon's Two Stage Design

One of the most widely used designs for phase II designs is Simon's design (1989) [24].This design includes two stages. In The first stage, a fixed number of participants is enrolled and if the number of responses is low the trial is stopped for futility. If the number of responses is satisfactory the trial proceeds in stage 2 where additional participants are enrolled. After the trial is finished, if the number of responses from both stages combined is sufficient,the trial is considered a success. If this is not the case, the intervention will not proceed to a phase 3 trial. As previously noted the hypothesis that is tested is:

$$H_0 : p = p_0$$
$$H_1 : p = p_1$$

Additionally, the notation of the number of responses observed in stage 1 and stage 2 during the trial is

$$Y_1 \sim \mathcal{B}(n_1, p)$$
$$Y_2 \sim \mathcal{B}(n_2, p)$$

respectively. For specific stage 1 and stage 2 thresholds (r_1, r) H_0, is rejected when

$$\{Y_1 > r_1\} \cap \{Y_1 + Y_2 > r\}$$

Based,on this rejection region one could ask why the rejection region is not:

$$\{\{Y_1 > r_1\} \cap \{Y_1 + Y_2 > r\}\} \cup \{Y_1 > r\}$$

That is, terminate the trial early for efficacy during the first stage when the number of responses is higher than the threshold of the second stage (which would suggest that the drug is very effective).Simon mentions that there is no reason to terminate a trial early when the drug seems to work·it is actually more ethical to continue the trial and enroll more patients since they will probably benefit from the treatment.Thus, even if the last event holds the null hypothesis is not rejected. In [8] it is also mentioned that the last event is highly unlikely to occur. In the case of actually observing it, the intervention is deemed effective from the first stage of the trial. Because of the above justifications, Simon's two stage design generally does not include the event $\{Y_1 > r\}$.

Using the above region as a test to not reject or reject the null hypothesis, it remains to calibrate its power and its type 1 error to a specified desirable level. That is:

$$P(\{Y_1 > r_1\} \cap \{Y_1 + Y_2 > r\}|H_0) \leq \alpha$$
$$P(\{Y_1 > r_1\} \cap \{Y_1 + Y_2 > r\}|H_1) \geq 1 - \beta$$

Since the number of responses is binomial distributed for any hypothesis it can be computed:

$$\begin{aligned} P(\{Y_1 > r_1\} \cap \{Y_1 + Y_2 > r\}|H_i) &= \sum_{y_1=r_1}^{n_1} \sum_{y_2=r-y_1}^{n_2} f(y_1, y_2|H_i) \\ &= \sum_{y_1=r_1}^{n_1} \sum_{y_2=r-y_1}^{n_2} f(y_1|H_i) f(y_2|H_i) \\ &= \sum_{y_1=1}^{n_1} \sum_{y_2=r-y_1}^{n_2} [\binom{n_1}{y_1} (p_i)^{y_1} (1-p_i)^{n_1-y_1} \times \\ &\times \binom{n_2}{y_2} (p_i)^{y_2} (1-p_i)^{n_2-y_2}] \end{aligned}$$

Under any of the two hypotheses it is evident that the type 1 error and the power of the test depends on a quadruple of parameters. These are (n_1, r_1, n, r). Since there are multiple combinations that satisfy the constraints, Simon proposed two additional criteria for selecting the parameters:

- **The minimax design:** The design that is selected minimizes the sample size

$$n = n_1 + n_2$$

- **Optimal design:** Under the null hypothesis define the total sample size variable N:

$$N = \begin{cases} n_1 & \text{, does not proceed to 2nd stage} \\ n_1 + n_2 & \text{, proceeds to 2nd stage} \end{cases}$$

Under the null hypothesis the expected value of this variable is,

$$\begin{aligned} E_{H_0}(N) &= n_1 P(Y_1 \leq r_1) + (n_1 + n_2) P(Y_1 > r_1) \\ &= n_1 P(Y_1 \leq r_1) + n_1 P(Y_1 > r_1) + n_2 P(Y_1 > r_1) \\ &= n_1 + n_2 P(Y_1 > r_1) \end{aligned}$$

The quarduple (n_1, r_1, n, r) that is selected is the one that minimizes the expected sample size. After specifying the parameters using one of the two criteria the trial will have an interim analysis when the participants reach the (n_1). The responses (y_1) will be compared with the first threshold (r_1). If the responses are fewer the trial will stop. If the response are greater, the trial will continue to accrue patients until they reach (n_2). Then, $y_1 + y_2$ will be compared to r. If the cumulative responses are greater we will accept the alternative hypothesis.

Notice also that in order to assess algorithmically wich quarduples satisfy the error constraints, one needs to specify the maximum sample size from the beggining, that is, specify the maximum possible total number N of patients that could be accrued during the clinical trial. If a small N is selected then the error constraint equations might not be satisfied.

Additionally, if one differentiates the probability of rejecting the null hypothesis with respect to p a positive derivative will occur [51]. This essentially means that in terms of sample size and boundary values Simon's design is also proper for the hypotheses

$$H_0 : p \leq p_0, H_1 : p \geq p_1$$

2.1.2 Heterogeneous Designs

In the previous paragraph Simon's two stage design was presented for a population. This entails the admission that the patients of a phase II trial population have the same response. However this is usually not the case since there are many subpopulations

that have different responses.For example, in recent years, technological advancements have given scientists the ability to recognize cancer subtypes based on genomic characteristics in patients. Based on these characteristics different responses to a specific therapy have been noticed [25]. A problem that can arise in such a case is that if there is a significant difference between subpopulations in their responses, during a clinical trial patients from the subpopulation with the smaller responses might participate more frequently than those that respond better to therapy. Without adapting the type 1 error based on the two sample sizes observed wrong results might arise. Additionally, in a case where the effect of a drug in the general population is assessed taking out of the study specific subpopulations (that is enriching other subpopulations) is not achievable. Thus, methodologies for designs that:

- Take into account The overall effect of a drug on a population based on responses from its subpopulations.
- Adapt the power and the type 1 error of the test applied based on the accrual of patients belonging to two different subgroups.

For these purposes we present Jung's heterogeneous design[26].Supposing that there are two subpopulations a,b with different responses and known prevalence c for subpopulation a it can be assumed that $p_{a0}, p_{a1}, p_{b0}, p_{b1}$ are known based on historical data. With these data the overall response rates under the null and alternative hypotheses are respectively.

$$H_0 : p = cp_{a0} + (1-c)p_{b0}$$

$$H_1 : p = cp_{a1} + (1-c)p_{a0}$$

Similarly to Simon's design the heterogeneous version has two stages one for early stopping and one to assess the results of the final data. The null hypothesis wil be rejected when:

$$\{Y_{11} + Y_{12} > r_1\} \cap \{Y_{11} + Y_{12} + Y_{21} + Y_{22} > r\}$$

Where Y_{ij} is the number of responses of subpopulation i in stage j and r, r_1 are the thresholds as defined previously. In this case it also defined m_{ij}, the number of patients of subpopulation i in stage j. To satisfy the second consideration for stratified designs (2) the conditional type I error and the conditional power is computed as:

$$a(m_1, m_2) = P(\{Y_{11} + Y_{12} > r_1\} \cap \{Y_{11} + Y_{12} + Y_{21} + Y_{22} > r\} | p_{a0}, p_{b0})$$

$$b(m_1, m_2) = P(\{Y_{11} + Y_{12} > r_1\} \cap \{Y_{11} + Y_{12} + Y_{21} + Y_{22} > r\} | p_{a1}, p_{b1})$$

Since the responses are binomial random variables:

$$
\begin{aligned}
& a(m_1, m_2) = \\
=& P(Y_{11} + Y_{12} > r_1, Y_{11} + Y_{12} + Y_{21} + Y_{22} > r | p_{a0}, p_{b0}) \\
=& \sum_{y_{12}=0}^{m_{12}} \sum_{y_{21}=0}^{m_{21}} P(Y_{11} > r_1 - Y_{12}, Y_{22} > r - Y_{11} + Y_{12} + Y_{21}, Y_{12} = y_{12}, Y_{21} = y_{21} | p_{a0}, p_{b0}) \\
=& \sum_{y_{12}=0}^{m_{12}} \sum_{y_{21}=0}^{m_{21}} [P(Y_{11} > r_1 - Y_{12}, Y_{22} > r - Y_{11} + y_{12} + y_{21} | Y_{12} = y_{12}, Y_{21} = y_{21}, p_{a0}, p_{b0}) f(y_{12}|p_{a0}) \times \\
& \times f(y_{21}|p_{b0})] \\
=& \sum_{y_{12}=0}^{m_{12}} \sum_{y_{21}=0}^{m_{21}} \sum_{y_{11}=r_1-y_{12}}^{m_{11}} P(Y_{11} = y_{11}, Y_{22} > r - y_{11} + y_{12} + y_{21} | y_{12}, y_{21}, p_{a0}, p_{b0}) f(y_{12}|p_{a0}) f(y_{21}|p_{b0}) \\
=& \sum_{y_{12}=0}^{m_{12}} \sum_{y_{21}=0}^{m_{21}} \sum_{y_{11}=r_1-y_{12}}^{m_{11}} P(Y_{22} > r - y_{11} + y_{12} + y_{21} | y_{11}, y_{12}, y_{21}, p_{a0}, p_{b0}) f(y_{11}|p_{a0}) f(y_{12}|p_{a0}) f(y_{21}|p_{b0}) \\
=& \sum_{y_{12}=0}^{m_{12}} \sum_{y_{21}=0}^{m_{21}} \sum_{y_{11}=r_1-y_{12}}^{m_{11}} \sum_{y_{22}=r-y_{11}+y_{12}+y_{21}}^{m_{22}} f(y_{11}|p_{a0}) f(y_{12}|p_{a0}) f(y_{21}|p_{b0}) f(y_{22}|p_{b0})
\end{aligned}
\tag{2.1}
$$

Similarly, the conditional power will be:

$$
\begin{aligned}
& 1 - b(m_1, m_2) = \\
=& \sum_{y_{12}=0}^{m_{12}} \sum_{y_{21}=0}^{m_{21}} \sum_{y_{11}=r_1-y_{12}}^{m_{11}} \sum_{y_{22}=r-y_{11}+y_{12}+y_{21}}^{m_{22}} f(y_{11}|p_{a1}) f(y_{12}|p_{a1}) f(y_{21}|p_{b1}) f(y_{22}|p_{b1})
\end{aligned}
\tag{2.2}
$$

For (2.1),(2.2) the parameters $(r_1, r, n_1, n_2))$ need to be calibrated such as the desirable constraints for the errors are satisfied. Notice that the constraints should be satisfied for any $m_{k1} < n_1, m_{k2} < n_2$ that will arise during the trial. In order to reduce the computations it is advised that n_1, n_2 are fixed based on a Simon's two stage design for the same $p_0.p_1$. Additionally, it is advised to fix r_1 as

$$[m_{11}p_{01} + m_{21}p_{02}].$$

In the above expression, [x] denote the interger part of the contained quantity. This is derived as follows. Under the null hypothesis the expected number of early

termination of the trial in stage 1 is $m_{11}p_{01} + m_{21}p_{02}$. It makes sense that under the null observed number of responses would be equal or less than the expected number. Fixing those parameters the only parameter that remains to be calibrated is r. Algorithmically, the specific design can be expressed in the steps below:

- Specify $(p_{01}, p_{02}, p_{a1}, p_{a2})$ and the error constraints $(a*, 1 - b*)$.
- Step 1:Specify the prevalence c_1 for subpopulation 1.
- Step 2: For $p_0 = c_1p_{01} + (1 - c_1)p_{02}$ and $p_a = c_1p_{a1} + (1 - c_1)p_{a2}$ apply a Simon's two stage design for for testing:

$$H_0 : p = p_0$$
$$H_1 : p = p_1$$

  That satisfies the error constraints. We use (n_1, n_2) from Simon's two stage to design to fix the number of participants in the heterogeneous designs.

After these computations the trial is conducted. During the stages:

- In stage 1: observe n_1 patients and observe (m_{11}, Y_{11}, Y_{21}). Calculate $r_1 = [m_{11}p_{01} + m_{12}p_{02}]$ based on the observed m_{11}. Reject the experimental therapy if $Y_{11} + Y_{21}$ is smaller than or equal to r_1. Otherwise, we proceed to stage 2.
- In stage 2: Treat n_2 patients and observe (m_{21}, Y_{21}, Y_{22}). Choose the largest integer r satisfying $a(m_{11}, m_{21}) < a$ conditioning on (m_{11}, m_{21}). Accept the therapy if $Y = Y_{11} + Y_{12} + Y_{21} + Y_{22}$ is larger than r.
- Calculate the conditional power from (2.2) with the previously chosen r.

In designing a two-stage phase II trial with stratified analysis, we should include the description of the whole procedure described above as well as the design parameter values in the study protocol [26].

2.1.3 Designs for Early Termination

In the previous sections single arm designs that include the possibility of early termination for futility where presented. One can also design trials that will terminate early for efficiency, that is, the evidence of efficacy are very strong and there is no need to proceed to a second stage. As noted there is rarely a reason to terminate a phase II trial early for efficacy. However, following Jung[27] we present this design here because it has some interesting theoretical aspects which will also be used in the next section. A two stage design that terminates early for efficiency will have the following structure:

- Stage 1:
  - If $X_1 \leq a_1$, reject the experimental therapy and stop the trial.
  - If $X_1 \geq b_1$, accept the experimental therapy and stop the trial.
  - If $a_1 < X_1 < b_1$, continue to stage 2.
- Stage 2:
  - Treat n_1 patients and observe the number of responders X_2.
  - If $X_1 + X_2 \leq a$ reject the experimental therapy.
  - If $X_1 > a$ accept the experimental therapy.

Defining the null and alternative hypothesis as:

$$H_0 : p = p_0$$
$$H_1 : p = p_1$$

The probability of not rejecting the null, which we will denote as $R(p)$ will be given as:

$$\begin{aligned} R(p) = & \\ & = P(\{X_1 \leq a_1\} \cup \{X_1 + X_2 \leq a, a_1 < X_1 < b_1\}|p) \\ & = P(\{X_1 \leq a_1\} \cup \{X_2 \leq a - X_1, a_1 < X_1 < b_1\}|p) \\ & = F_{x_1}(a_1; n_1, p) + \sum_{x_1=a_1+1}^{b_1-1} \sum_{x_2=0}^{a-x_1} f(x_1, x_2|p) = F(a_1; n_1, p) + \sum_{x_1=a_1+1}^{b_1-1} f(x_1|p) \sum_{x_2=0}^{a-x_1} f(x_2|p) \\ & = F_{x_1}(a_1; n_1, p) + \sum_{x_1=a_1+1}^{b_1-1} f(x_1|p) F_{x_2}(a - x_1; n_2, p) \end{aligned} \tag{2.3}$$

Here we denote as $F(x; n, p)$ the binomial cdf with parameters n and p.

Under H_0 the type I error constraint is of the form $R_(p) \geq 1 - a$ (notice that under H_0 we have computed the probability of correctly not rejecting it). Under the alternative hypothesis a constraint is given as: $R_(p) \leq b$ where b is the type II error.

Similar to Simon's two stage design optimality criteria should be used to restrict the set of parameters (a_1, b_1, n_1, n, a) parameters that satisfy the error constraints. In this case define again the r.v. of the sample size of the clinical trial:

$$N = \begin{cases} n_1 & \text{, does not proceed to 2nd stage} \\ n_1 + n_2 & \text{, proceeds to 2nd stage} \end{cases}$$

The probability of early termination (P.E.T.),i.e. stop the trial at the first stage either for futility of efficacy, in the case of this design is:

$$P.E.T.(p) = P(X_1 \leq a_1, X_1 > b_1) = P(X_1 \leq a_1) + 1 - P(X_1 < b_1) \tag{2.4}$$

Thus,

$$E_p(N) = n_1 P.E.T.(p) + (n_1 + n_2)(1 - P.E.T.(p))$$

The two optimality criteria are similar to Simon's two stage design:

- The minimax design minimizes the maximum number of patients $n_1 + n_2$
- The optimal design minimizes the average of H_0 expected sample size for $p = p_0$ and $p = p_1$ which is given as:

$$E(N) = \frac{E_{p_0}(N) + E_{p_1}(N)}{2}$$

Before proceeding we make two important notes:

1) If we set $M = 2$ and arbitrarily $b_1 = c > n_1$ then this design will end up being Simon's two Stage design. Notice how $1 - R(p)$ will be equal to:

$$P(X_1 > a_1, X_1 + X_2 > a)$$

because,

$$P(X_1 > c) = 0$$

and

$$\begin{aligned} &P(a_1 < X_1 < c) = P(X_1 < c) - P(X_1 \leq a_1) = 1 - P(X_1 \leq a_1) \\ &= P(X_1 > a_1) \end{aligned} \tag{2.5}$$

Which means that we can replace the event $\{a_1 < X_1 < c\}$ with $\{X_1 > a_1\}$ yielding the rejection region of Simon's two stage design.Thus with proper conventions this design entails Simon's two stage design.

2) This design as wel as Simon's can be generalised to K stages:

For example,

- stage 1: Similar to the two stage design.
- stages until K-1:

$$\begin{aligned} &X_1 + ... + X_k \leq a_{k-1}, \text{do not reject the null} \\ &X_1 + ... + X_k \geq b_{k-1}, \text{reject the null} \\ &a_{k-1} < X_1 + ... + X_k \leq b_{k-1}, \text{proceed to next stage} \end{aligned} \tag{2.6}$$

- stage K:

$$\begin{aligned} &X_1 + ... + X_K \leq a_K, \text{do not reject the null} \\ &X_1 + ... + X_K > a_K, \text{ reject the null} \end{aligned} \tag{2.7}$$

These two notes will prove usefull in the next section.

2.1.4 UMVU Estimator of Reponse Rate

Based on the two previous notes the most general form of phase II design is a K-stage design with thresholds for early futility and efficacy. Finding an UMVUE in this case covers all the other cases (including Simon's two stage design). It is noteworthy that in a multistage phase II trial the number of stages M is a random variable too. Defining $S_M = X_1 + X_2 + ... + X_M$ as the sum of all responses. During a K-stage phase II trial a sample of the form $(X_1, ...X_M, M)$ will be aqcuired. The distribution of this sample has the response rate as its parameter because,using bayes' rule:

$$P(X_1,...X_M,M) = P(\{X_1+....+X_M \le a_m \cup X_1+....+X_k \ge a_m\} \cap \\ \cap \{a_k < X_1+....+X_k < b_k, k=1,....,m-1\}) \times P(X_1,...X_M|M) \tag{2.8}$$

This probability is consisted only from independent binomial random variables that have the sample probability but different sample sizes. Thus the only parameter to be estimated is the response rate.

Presenting Jung's and Kim's method [28] first it will be shown that (M, S_M) is a sufficient and complete statistic for $(X_1,...X_M,M)$. For sufficieny instead of using the factorization theorem as Jung and Kim did in their paper [27] we will use the traditional definition of sufficiency. Thus we will show that for any fixed $M=m$ observing S_M the distribution of X does not depend on its parameter or:

$$P_p(\mathbf{X}_m|M=m,S_M=s) = g(x_1 \dots x_m)$$

Expanding this conditional distribution will give:

$$P_p(X_1,...X_m|M=m,S_M=s) = \frac{P_p(X_1,...,X_m,M=m,S_M=s)}{P(M=m.S_M=s)} \\ = \frac{P_p(M=m,S_M=s|X_1,...X_m)P(X_1,...X_m)}{P(M=m.S_M=s)} \tag{2.9}$$

but obviously,

$$P_p(M=m,S_M=s|X_1,...X_m)=1,$$

so what remains is:

$$\frac{P(X_1,...X_m)}{P(M=m.S_M=s)}$$

To complete the proof of sufficieny one needs to compute the joint distribution of (M, S_M):

$$P(M_m,S_M=s) = P(a_k+1 \le s_k \le b_k-1, S_m=s, 1 \le k \le m-1) \\ = \sum_{\mathcal{P}_1} P(X_1=x_1,....,X_m=x_m) \tag{2.10}$$

where $\mathcal{P}_1$ is defined as the set:

$$\{(x_1, ...x_m) : a_k{+}1 \leq S_k \leq b_k{-}1, 1 \leq k \leq m{-}1, a_{m-1} \leq s \leq a_m, \text{or}, b_m \leq s \leq n_m{+}b_{m-1}{-}1\}$$

In simpler terms $\mathcal{P}_1$ is the set that contains all the possible combinations of x_i's such that the trial was terminated at stage M.

Continuing with (2.10):

$$\sum_{\mathcal{P}_1} P(X_1 = x_1), ..., P(X_m = x_m) = p^s p^{\sum_{i=1}^m n_i - s} \sum_{\mathcal{P}_1} \binom{n_1}{x_1} ... \binom{n_1}{x_m} = c_{m,s} p^s p^{\sum_{i=1}^m n_i - s} \tag{2.11}$$

where:

$$c_{m,s} = \sum_{\mathcal{P}_1} \binom{n_1}{x_1} ... \binom{n_1}{x_m}$$

Similarly, in the numerator:

$$\begin{aligned} P(X_1 = x_1, ..., X_m = x_m) &= P(X_1 = x_1) ... P(X_m = x_m) \\ &= \binom{n_1}{x_1} p^{x_1} (1-p)^{n_{x_1}} ..... \binom{n_1}{x_1} p^{x_2} (1-p)^{n_{x_2}} \\ &= p^{x_1 + ... + x_m} (1-p)^{n_1 + ... + n_m - x_1 - ... - x_m} \binom{n_1}{x_1} ... \binom{n_m}{x_m} \end{aligned} \tag{2.12}$$

Replacing the terms in (2.9) with (2.11),(2.12):

$$\frac{p^{x_1 + ... + x_m} (1-p)^{n_1 + ... + n_m - x_1 - ... - x_m} \binom{n_1}{x_1} ... \binom{n_m}{x_m}}{p^s (1-p)^{\sum_{i=1}^m n_i - s} \sum_{\mathcal{P}_1} \binom{n_1}{x_1} ... \binom{n_1}{x_m}} \tag{2.13}$$

However, due to the initial condition that $S_m = s$, $x_1, ..., x_m$ are such that $\sum_{i=1}^m x_i = s$. Thus it follows immediatelly that all the terms containing p are eliminated and sufficiency is proved.

Having derived the joint distribution of (M, S) note that for a specific value of $M = m$ (that is, the trial is terminated at stage m) the values $S_M = s$ can take are

$$\mathcal{R}_m = \{(m, s) : a_{m-1} + 1 \leq s \leq a_m \text{ or } b_m \leq s \leq n_m + b_{m-1} - 1\}$$

Where $a_0 = -1, b_0 = 0$. Thus the support of the joint distribution is:

$$\mathcal{R} = \cup_{m=1}^{K} \mathcal{R}_m$$

For the completeness of (M, S) setting $h(p) = E_p(g(M, S))$ it must be shown that if $h(p) = 0$ for any $p \in (0, 1)$ then $g(m, s) = 0$ for all values (m,s) in the support of (M,S). Based on the support of the joint distribution, it is straightforward that:

$$E_p(g(M,S)) = \sum_{m=1}^{K} \sum_{s=a_{m-1}+1}^{a_m} g(m,s) f(m,s|p) + \sum_{m=1}^{K} \sum_{s=b_m}^{s=n_m+b_{m-1}-1} g(m,s) f(m,s|p)$$

and expanding the joint distribution we have that

$$\begin{aligned} E_p(g(M,S)) = &\sum_{m=1}^{K} \sum_{s=a_{m-1}+1}^{a_m} g(m,s) c_{m,s} p^s (1-p)^{\sum_{i=1}^{m} n_i - s} + \\ &+ \sum_{m=1}^{K} \sum_{s=b_m}^{s=n_m+b_{m-1}-1} g(m,s) c_{m,s} p^s (1-p)^{\sum_{i=1}^{m} n_i - s} \end{aligned} \tag{2.14}$$

Now, it can be noted that this expected is actually a polynomial of p with finite order but it is zero for every value of p. This implies the fact that the coefficients of p must be zero and since $c_{m,s}$ cannot be zero the result follows.

An unbiased estimator of the response rate is $\hat{p} = \frac{x_1}{n_1}$,which is the estimator of the response rate using data only from the first stage. Additionally, (M, S_M) is a sufficient and complete statistic and as a result the Rao-Blackwell theorem can be used to acquire an UMVU estimator for the response rate $p^* = E(\hat{p}|M, S_M) = \frac{E(X_1|M,S_M)}{n_1}$.

The conditional distribution of X_1 given (M, S) is

$$P(X_1 = x_1|M = m, S = s) = \frac{P(X_1 = x_1, M = m, S = s)}{P(M = m, S = s)}$$
$$= \frac{P(M = m, \sum_{i=2}^{m} x_i = s - x_1|X_1)P(X_1 = x_1)}{P(M = m, S = s)} \quad (2.15)$$
$$= \frac{\sum_{\mathcal{P}_2} P(X_2, ..., X_m)P(X_1 = x_1)}{P(M = m, S = s)}$$

where

$$\mathcal{P}_2 = \{(x_2, ...x_m) : x_2 + ... + x_m = s - x_1, a_k + 1 \leq s_k \leq b_k - 1, k = 2, ..., m - 1\}$$

Thus,

$$p^* = \frac{\sum_{x_1}^{n_1} x_1 P(X_1|M, S_M)}{n_1}$$

2.2 Multiple Arm Designs

2.2.1 Simon's Randomized Designs

Simon (1989) mentioned that there is high variability in the response rates observed throughout different Phase II clinical trials for the same experimental therapy. He pointed out, that this variability is attributed to[29]:

- Patient Selection
- Response Criteria
- Inter-Observer variability in response assessment
- Dosage modification and protocol compliance
- Reporting Procedures
- Sample Size

Patient selection creates variability because each patient has specific characteristics that contribute in a different way to treatment efficacy. Response criteria is also an important factor since there are not universal criteria among institutions and groups.

Subjectivity in response assessment is also a factor as well as measurement errors for the response. Different dosages, reflect the treatments' aggressiveness, and consequently with a higher dose there might greater or fewer responses. Different guidelines and protocols followed during the Phase II trials (for example, exclusions made for the calculation of response rates) as well as different sample sizes can also contribute to variations in response rates.

In such cases randomized trials might help reduce this variability. For example, in a randomized phase II trial the protocol followed (e.g. assessment of response,eligibility criteria) is the same for each arm and thus difference in responses are not dependent on such design parameters but true response rates. As noted many differences in response rates stem from different dosages administered to participants or different methods of administration. With a randomized phase II trial different arms that administer the same intervention but with different methods can be designed in order to evaluate what is response-wise the best method to administer the drug. Additionally, many new different agents can be assessed simultaneously and ranked in selecting therapies for patients.

It is interesting to note that Simon's paper contains two types of randomized phase II designs and both of them do not contain a placebo arm. In one design there is a control arm where participants are given a standard treatment and in the other design there is no control arm at all (only experimental interventions are given). Also, the first design is not particularly useful because it entails the possibility that both responses are low and this implies that the results are inconclusive (because if the response rates in the standard treatment are also low than patient selection is probably inappropriate). Thus only the second type of design is described broadly. It is also logical to ask why would someone design a trial where treatments that may be ineffective are compared to each other. As Simon notes, Phase II trials do not have the sole goal of assessing early efficacy. Other goals are to assess the degree of anti-tumor activity, the extent of tumour shrinkage, the propotions of who respond (response rates) and the durability of response. Under a randomized phase II these parameters can be estimated with the same methodology and also be compared to each other. Such data can be used to rank available agents in selecting therapy for patients and in developing plans for introducing the investigational agent with the best phase II results in front-line combinations.

Regarding methodology of such a designs, the aspects of error control are replaced by statistical selection theory.

From K experimental arms the probability that the most effective is selected is computed under different sample sizes. More specifically, suppose that $p_{(1)}, ..., p_{(K)}$ are the ranked true response rates of the experimental drugs. The objective is to design a trial that will pick the treatment with $p_{(K)}$ true response rate as the most effective (which is the correct choice). This is not the same as the power of a test. A proposal would be to select the intervention with the most responses during the trial as the most effective. However, it is probable that ties in response numbers will occur between different arms. In this case one of the tied treatments will be chosen randomly. Even though Simon presents the final formula and its parameteres for the probability of selecting the most effective intervention following the previously mentioned framework, the derivation formula is not proved in a strict probabilistic manner. Also the derivation of the formula is not simple, since during the design of the trial the number of ties that will observed is not known and is considered a random variable. A detailed proof is given below:

We consider that at each arm equal numbers of patients are enrolled. Additionally, we suppose that every other treatment except from the most effective one has response rate p. We suppose that the most effective treatment has response rate $p + D$ where D is decided based on clinical criteria.Also, in each arm equal numbers of patients are accrued we will denote the number in each arm as n_1.

The probability that following the previous framework (that is, select the treatment with the most responses during the trial and if ties occur choose one of the response-tied treatments randomly) will give the most effective (M.E.) intervention is:

$$P(\{X_{(M.E.)} > X_{(K-1)}\} \cup \{X_{(M.E.)} = X_{(K-t)}, \forall 1 \leq t \leq J, X_{(K-t)} < X_{(K-J)}, t > J\} \cap \\ \cap \{\text{choose the M.E. treatments from } J{+}1 \text{ ties}\}) \tag{2.16}$$

where $J = 1, ..., K-1$ is the variable that denotes the number of ties between the largest response numbers. Also one can notice that

$$\begin{aligned} \{J = r\} = \{&X_{(M.E.)} = X_{(K-t)}, \forall 1 \leq t \leq r\, X_{(K-t)} < X_{(K-r)}, \forall t > r\} \cup \\ &\cup \{X_{(K-1)} = X_{(K-1-t)}, \forall 1 \leq t \leq r, X_{(K-t)} < X_{(K-1-r)}, \\ &, X_{(M.E.)} < X_{(K-1-r)} \forall t > r\} \end{aligned} \tag{2.17}$$

where $r < K-1$ since in the case of $r = K-1$ all the response numbers are equal.

Additionally we consider that the random variables that are not equal are strictly smaller in the above events, e.g. the second event implies $X_{M.E.}, X_{(K-r-1)} < X_{(K-1)}$. We denote the event of choosing the M.E. treatment from $J+1$ treatments as a random variable Z_{J+1} that takes the value 1 when the M.E. treatment is chosen and zero otherwise. It is obvious that the two events do not have a common intersection so the probability can break into two probabilities:

$$\begin{aligned}
&P(\{X_{(M.E.)} > X_{(K-1)}\})+ \\
&+ P(\{X_{(M.E.)} = X_{(K-r)}, \forall 1 \leq r \leq J, X_{(K-t)} < X_{(K-J)}, \forall t > J\} \cap \{Z_{J+1} = 1\}) \\
&= \sum_{i=1}^{n_1} P(\{X_{(M.E.)} > X_{(K-1)}, X_{(K-1)} = i\})+ \\
&+ \sum_{r=1}^{K-1} P(\{X_{(M.E.)} = X_{(K-t)}, \forall 1 \leq t \leq J, X_{(K-t)} < X_{(K-J)}, \forall t > J\} \cap \{Z_{J+1} = 1\} \cap \{J = r\}\})
\end{aligned} \tag{2.18}$$

Note that J is actually a random variable during the design stage, since it is not known how many ties will occur. Now we compute:

$$\begin{aligned}
&P(X_{(K-t)} = c, \forall 1 \leq t \leq r, X_{(K-t)} < c, t > r) = \\
&= P(X_{(K-1)} = c, ..., X_{(K-r)} = c, X_{(k-r-1)} < c, ..., X_{(1)} < c) \\
&= \binom{K-1}{r} P(X_{K-1} = c) \times P(X_{K-r} = c) \times P(X_{k-r-1} < c) \times \\
&\times P(X_1 < c) \\
&= \binom{K-1}{r} [P(X_{K-1} = c)]^r [P(X_1 < c)]^{K-r-1}
\end{aligned} \tag{2.19}$$

Additionally, given the number of ties r and the fact that M.E. is tied the distribution of Z_{r+1} is :

$$\frac{1}{r+1} \tag{2.20}$$

Continuing with the computations, (2.18) expands as below:

$$\sum_{i=1}^{n_1} P(X_{(M.E.)} > X_{(K-1)} | X_{(K-1)} = i) P(X_{(K-1)} = i) +$$
$$+ \sum_{r=1}^{K-1} P(\{X_{(M.E.)} = X_{(K-t)}, \forall 1 \leq t \leq J, X_{(K-t)} < X_{(K-J)}, t > J\}, Z_{J+1} = 1, J = r)$$
$$= \sum_{i=1}^{n_1} (1 - P(X_{(M.E.)} > i | X_{(K-1)} = i)) P(X_{(K-1)} = i) +$$
$$+ \sum_{r=1}^{K-1} [P(Z_{J+1} = 1, \{X_{(M.E.)} = X_{(K-t)}, \forall 1 \leq t \leq J, X_{(K-t)} < X_{(K-J)}, \forall t > J\}, J = r) \tag{2.21}$$

Considering that $X_{(M.E.)} \sim \mathcal{B}(n, p + D)$ and based on the fact that:

$$X_{(K-1)} = \max\{X_1, ..., X_{K-1}\}$$

implies $F(x_{(K-1)}) = [F(x)]^{K-1}$ the first sum of (2.21) continues as:

$$\sum_{i=1}^{n_1} (1 - \mathcal{B}(i, n, p + D))[(P(X_{(K-1)} \leq i + 1))^{K-1} - (P(X_{(K-1)} \leq i))^{K-1}] +$$
$$+ \sum_{r=1}^{K-1} P(Z_{J+1} = 1, \{X_{(M.E.)} = X_{(K-t)}, \forall 1 \leq t \leq J, X_{(K-t)} < X_{(K-J)}, \forall t > J\}, J = r) \tag{2.22}$$

Now, for the second part of the equation one needs to remember the formula for $\{J = r\}$ and see its intersection with the other events.Additionally taking one more summation for all the values of $X_{(K-t)}$ yields:

$$\sum_{i=1}^{n_1} (1 - \mathcal{B}(i, n, p + D))[(P(X_{(K-1)} \leq i + 1))^{K-1} - (P(X_{(K-1)} \leq i))^{K-1}] +$$
$$+ \sum_{r=1}^{K-1} \sum_{c=1}^{n_1} P(\{X_{(M.E.)} = X_{(K-t)}, t \leq r\}, \{X_{(K-t)} = c,\ t \leq r, X_{(K-t)} < c, t > r\}, Z_{r+1} = 1) \tag{2.23}$$

And continuing by using Bayes theorem we finally get:

$$\sum_{i=1}^{n_1}(1-\mathcal{B}(i,n,p+D))[(P(X_{(K-1)}\leq i+1))^{K-1}-(P(X_{(K-1)}\leq i))^{K-1}]+$$
$$+\sum_{r=1}^{K-1}\sum_{c=1}^{n_1}[P(Z_{r+1}=1|\{X_{(M.E.)}=X_{(K-t)},t\leq r\},\{X_{(K-t)}=c,t\leq r,X_{(K-t)}<c,t>r\})\times$$
$$\times P(\{X_{(M.E.)}=X_{(K-t)},t\leq r\}|X_{(K-t)}=c,t\leq r,X_{(K-t)}<c,t>r)\times$$
$$\times P(X_{(K-t)}=c,\forall 1\leq t\leq r,X_{(K-t)}<c,t>r)] \tag{2.24}$$

This was the formula originally proposed by Simon [29]. For this formula we calibrate the values of n_1 based on specified values of K, D, p.

2.2.2 Jung's Randomized Phase II Designs

From the previous paragraphs it is evident that single arm designs are applied when historical data exists. Aditionally, in the previous section a randomized design was presented but historical data was aso used. However, as mentioned in previous paragraphs, there are rare cases where historical data does not exist. In such a randomized phase II trial participants are randomized to an arm where they are given the experimental treatment and to a placebo group. Note that the term randomization refers to trials that may use historical controls or not. For the case of a randomized phase II trial without historical controls we present Jung's Randomized phase II design[31]. The hypotheses tested are essentially:

$$H_0 : p_x = p_y$$
$$H_1 : p_x > p_y$$

Where p_x, p_y are the proportions of response in the experimental and placebo groups. The statistical test is based on Fisher's exact test.

Fisher's exact test tests the hypotheses

$$H_0 : p_x = p_y$$
$$H_1 : p_x \neq p_y$$

and is based on the conditional probability of X given the total number of responders $Z = X + Y$ that is,

$$
\begin{aligned}
f(x|z) = P(X = x|Z = z) =& \frac{P(X = x, Z = z)}{P(Z = z)} \\
=& \frac{P(X + Y = z|X = x)P(X = x)}{P(Z = z)} \\
=& \frac{P(Y = z - x|X = x)P(X = x)}{P(Z = z)}
\end{aligned} \tag{2.25}
$$

with

$$
\begin{aligned}
P(Z = z) &= \sum_{x=m_-}^{m_+} f(x, z) = \sum_{m_-}^{m_+} f(z|x)f(x) \\
&= \sum_{m_-}^{m_+} \binom{n}{x}\binom{n}{z-x} p_y^{z-x}(1 - p_y)^{n-z+x} p_x^x(1 - p_x)^{n-x}
\end{aligned} \tag{2.26}
$$

where $m_- = max\{0, z - n\}$ and $m_+ = min\{n, z\}$.

Under $H_0 : p_x = p_y$ Expanding (2.25) in the numerator and replacing the denominator with (2.26) the distribution of X conditional on Z is:

$$
P(X = x|Z = z) = \frac{\binom{n}{x}\binom{n}{z-x}}{\sum_{m_-}^{m_+}\binom{n}{x}\binom{n}{z-x}} \tag{2.27}
$$

Even though the hypothesis is two sided, the null hypothesis in practical applications is rejected using a one-sided test [49] (thus we actually reject in favor of $H_1 : p_x > p_y$).Specifying a threshold $c(z)$ which is the smallest integer for which

$$
P(X > c(z)|Z = z, H_0) < \alpha
$$

holds, the rejection region is specified.

In a single stage randomized design, given the cumulative number of responses from both groups, if the difference of responses is significantly larger in the experimental group than the number of responses in the placebo group the null hypothesis should be rejected in favor of $H_1 : p_x > p_y$ if for a specific threshold r_1 given Z:

$$\{X - Y > r_1 | Z = z\} = \{X - (Z - X) > r_1 | Z = z\} = \{X > \frac{r_1 + z}{2} | Z = z\}$$

Note that the rejection area is similar to the one using the $c(z)$ threshold. Based on (2.27) the probability of the above event under H_0 or H_1 can be computed as:

$$P(X > \frac{r_1 + z}{2} | Z = z, H_i) = \sum_{x=[\frac{r_1+z}{2}]}^{m_+} f(x|z, H_i) \tag{2.28}$$

Now that notation and the basic formulas have been presented, the case for a two stage randomized phase II design is presented. One can design a single stage randomized design also but it is preffered for ethical reasons to include one stage for interim analysis and one for the final assessment of the intervention. For a two stage design the null hypothesis will be rejected when:

$$\begin{aligned}
&\{X_1 - Y_1 > r_1, X_1 + X_2 - Y_1 - Y_2 > r | Z_1 = z_1, Z_2 = z_2\} \\
&= \{X_1 > \frac{r_1 + z_1}{2}, X_1 + X_2 - z_1 + X_1 - z_2 + X_2 > r | Z_1 = z_1, Z_2 = z_2\} \\
&= \{X_1 > \frac{r_1 + z_1}{2}, X_1 + X_2 > \frac{r + z_1 + z_2}{2} | Z_1 = z_1, Z_2 = z_2\}
\end{aligned} \tag{2.29}$$

For the probability of event in (2.29) it can be computed that:

$$\begin{aligned}
&P(\{X_1 > \frac{r_1 + z_1}{2}, X_1 + X_2 > \frac{r + z_1 + z_2}{2} | Z_1 = z_1, Z_2 = z_2\}) \\
&= P(\{X_1 > \frac{r_1 + z_1}{2}, X_2 > \frac{r + z_1 + z_2}{2} - X_1 | Z_1 = z_1, Z_2 = z_2\}) \\
&= \sum_{x_1=\frac{r_1+z_1}{2}}^{m_{1+}} \sum_{x_2=\frac{r+z_1+z_2-x_2}{2}}^{m_{2+}} f(x_1, x_2 | z_1, z_2) \\
&= \sum_{x_1=m_{1-}}^{m_+} \sum_{x_2=m_{2-}}^{m_{2+}} I_{\{X_1 > \frac{r_1+z_1}{2}, X_2 > \frac{r+z_1+z_2}{2} - X_1\}} f(x_1|z_1) f(x_2|z_2)
\end{aligned} \tag{2.30}$$

Computing (2.30), the parameters that need to be calibrated so that the specified error constraints are satisfied are (n_1, n_2, r_1, r). In order to reduce the computations

it is advised to set $r_1 = 0$. Also the maximum possible total sample size should be specified. For each n_1, n_2 and for each z_1, z_2 we find $r(z_1, z_2)$ thresholds for which the error constraints are satisfied. This is repeated for all the possible combinations of n_1, n_2.

After finding $r(z_1, z_2)$ for each z_1, z_2 for each n_1, n_2 we select the pairs n_1, n_2 for which the expected power:

$$E(1 - b(Z_1, Z_2)|H_1) = \sum_{z_1=0}^{2n_1} \sum_{z_2=0}^{2n_2} f(z_1|H_1) f(z_2|H_1)(1 - b(z_1, z_2))$$

satisfies a specific power constraint ($\geq 1 - \beta$).Note that in order to compute the above formula one must actually specify the response probabilities. This creates a disadvantage that if the response probabilities are different the power of the test for the selected sample sizes will also be different and it could be lower. With this specification a number of quadruples $(n_1, n_2, 0, r_1)$ that satisfy the error constraints can be found. In order to select a specific quadruple similar optimality criteria to that of Simon's (1989) are used (minimax and optimal). The minimax remains the same but in the case of the optimal criteria, instead of using the probability of early termination, in this case the expected probability of early termination is used for the optimal criteria, that is:

$$\begin{aligned} P.E.T. = E(PET(Z_1)|H_0) &= \sum_{z_1=0}^{2n_1} P.E.T(z_1|H_0) f(z_1|H_0) \\ &= \sum_{z_1=0}^{2n_1} P(\{X_1 > \frac{r_1 + z_1}{2}|H_0) f(z_1|H_0) \end{aligned} \tag{2.31}$$

The optimal criteria will then select the design that minimizes the quantity:

$$EN = n_1 P.E.T. + (n_1 + n_2)(1 - P.E.T.)$$

2.3 Discussion of Methods

Simon's design: Simon's design has become the standard design for phase II trials especially in oncology clinical trials. Advantages of using Simon's two stage design are:

- It is ethical. If the therapy is not promising the trial terminates, thus the number of participants that are treated with an ineffective treatment is reduced. From

an economical viewpoint this is also an advantage since, if the drug is ineffective, the spendings reduce due to early termination.

- Since the design is based on binomial probalities and computer computations, the sample size needed for testing the hypotheses is less than a sample size derived through asymptotic methods.

Some disadvantages that are commonly seen in Simon's two stage are:

- The minimax design may have an excessively large expected sample size as compared to the optimal design, or the optimal design may have an excessively large maximum sample size n as compared to the minimax design. Additionally it is not always clear which one of the two criteria to use.

- Additionally problems than can occur is not specifying possible subgroups or using inaccurate historical controls. But essentially this is mostly a problem of using the wrong design.

Hetegoreneous Simon's design: Heterogeneous designs are essentially a modified Simon two stage design for cases where there are subgroups in the patients' population that have different response rates. In that notion heterogeneous designs contribute in testing for a general treatment effect while adjusting the threshold based on the number of patients from each subgroup. This is an important advantage since results without this adjustment would be biased. Howerver there is also a possible drawback which is specifying the prevalence of the subgroups correctly. When the prevalence of the subgroups is not considered correctly the results can also become biased.

Simon's Randomized Designs: The advantages and contributions of this design were thorougly explained in the respective section. Here we mention two drawbacks of this method:

- This design compares promising treatments with each other and the parameters of the comparisons are calibrated such that if we select the treatment with the most responses or randomly select one of the treatments that have equally the

largest number of responses we will actually select the most effective treatment. However there is a case that no treatment is actually more effective compared to a standard treatment. As Simon has mentioned there are still many advantages in comparing the treatments with each other but there is also the case that an ineffective treatment will proceed for a phase III trial.

- Since this design is not based on the notions of statistical power and type I error, no considerations about type I error are made.

Jung's Randomized Phase II designs: The advantage of this design is obvious when historical controls do not exist. Without specifying any proportion one can test for the effectiveness of a drug while keeping the type I error of the test below the specified constraint. Thus if a treatment is ineffective the probability of proceeding to a phase III trial (that is reject the null hypothesis) is not inflated.

However a main possible disadvantage is that in order to select the sample sizes at each stage the proportions must ultimately be specified. Since the sample sizes are selected based on the conditional expected power $E(1 - b(z_1, z_2))$ if the proportions are not specified correctly the actual power of the clinical trial can be less than $E(1 - b(z_1, z_2))$.

Chapter 3

Phase III trials

3.1 Classification of Phase III trials: Discussion

If an intervention succeeds in showing some efficacy during a phase II trial, a phase III trial will commence. As described in chapter 1 these trials require a larger sample size and their purpose is to establish with strong evidence the efficacy of the experimental intervention. If the results of a phase III trial are successfull the intervention will be approved and it will acquire market approval. Something noteworthy is, that for a Phase III trial to be considered successfull, the effectiveness of the drug need not be larger than the effect of the standard therapy (obviously it should be larger than that of a dummy intervention). This attribute classifies Phase III trials into three main categories, based on their goals:

Superiority trials: As the name suggests, superiority trials are designed to establish that the new experimental therapy is more efficient than the standard therapy that is used. In the abscence of a standard therapy a superiority trial establishes that those who receive the therapy respond significantly better than those that don't receive it. For example, in the case of comparing means of populations the null and alternative hypothesis would be of the form:

$$H_0 : \mu_1 \leq \mu_2$$
$$H_1 : \mu_1 > \mu_2$$

where μ_1 is the population mean of the experimental therapy and μ_2 the population mean of the standard therapy (or the population mean of those who don't receive therapy), respectively.

Non-inferiority trials: Noninferiority trials intend to demonstrate that the effect of the experimental treatment is not worse than that of the standard treatment by more than a prespecified margin, $\delta > 0$. In a noninferiority trial, the new treatment is expected to be at least similar to the existing therapy in terms of efficacy, while the advantages of the new treatment may include being more convenient to administer, inducing fewer side effects, or being less expensive [25]. In the case of comparing population means the hypotheses would be of the form:

$$H_0 : \mu_1 \leq \mu_2 - \delta$$
$$H_1 : \mu_1 > \mu_2 - \delta$$

If we intend to design an NI trial, then we must make the following two assumptions[25]:

- The constancy of the control effect
- The non-inferiority trial has assay sensitivity

The constancy assumption, is the assumption, that the effect of the standard treatment is the same in the current NI trial as that in historical superiority trials, which demonstrated the efficacy of the standard treatment. It has been shown previously that violation of this assumption leads to incorrect conclusion of non-inferiority.In simpler terms, simply concluding non-inferiority is not enough if there is no evidence that the standard therapy already has efficacy on the population that the experimental therapy will be given to. This can be logically explained as follows: An NI trial which fails to show a difference in treatment effects between T and C may mean that either both drugs were effective or neither drug was effective. Thus, the similarity of the current NI trial to past studies is also a consideration and use of historical data is necessary.

The second assumption, the assay sensitivity, is the ability of the trial to distinguish an effective treatment from a less effective or ineffective treatment if such a difference truly exists. In the case of a superiority trial, if it successfully demonstrates superiority, assay sensitivity is immediately demonstrated[20]. However, an NI trial that successfully finds the effects of the treatments to be similar has not necessarily demonstrated assay sensitivity. For example, considering a non-inferiority margin that is larger than the treatment effect difference between the standard therapy and placebo can probably yield a design with no assay sensitivity. Therefore, in addition to the constancy assumption, an NI trial must also rely on an assumption of assay sensitivity.

Apart from these assumptions, the choice of the control treatment is essential. Not only must the control's treatment effect be well documentated and observed on the same population, it is also essential that its effectiveness is based on placebo controlled trials or on superiority trials. Ignoring this parameter, may lead to an interesting phenomenon, known as"bio-creep" [31]. To make this case more understandable consider the following example:

Three experimental drugs A,B,C for a specific population exist. A was deemed effective through a placebo-controlled trial superiority trial. Then B was deemed non-inferior to A. Afterwards C was deemed non-inferior to B. It is evident that following this framework C may be worse compared to B which is worse than A and thus C may be inferior to A, but still gains market approval (when clearly it should not). Even, if C is non-inferior continuously following this framework will ultimately give an inferior intervention compared to A. This is an example of a "bio-creep" case.

Finally, a very important aspect of non-inferiority trials is the selection of the non-inferiority margin. According to the ICH guidelines:

- The determination of the margin in a non-inferiority trial is based on both statistical reasoning and clinical judgment, and should reflect uncertainties in the evidence on which the choice is based, and should be suitably conservative.

- This non-inferiority margin cannot be greater than the smallest effect size that the active drug would be reliably expected to have compared with placebo in the setting of a placebo-controlled trial.

The process of selecting an appropriate M may be broken down into two steps. As [25] describes:

- First, statistical methodology must be used to determine the smallest possible effect size of the control intervention compared with placebo.

- Second, clinical judgment should be used to determine the largest loss of effectiveness of the new treatment relative to the active-control that would still be considered as clinically insignificant, considering the smallest effect size.

In order to decide the smallest possible effect size (denote it as δ_1) of the control many methodologies have been proposed. One of them would be to collect relevant clinical trials that used the control treatment and compared it to placebo and use meta-analysis methodologies to construct confidence intervals for the true effect size. Afterwards select the lowest boundary point of the C.I. as the smallest possible effect size of the active control.

According to the F.D.A. after selecting δ_1 one will consider the non-inferiority margin as

$$\delta = (1 - l)\delta_1$$

Where l takes values between 0 and 1 . This is where clinical judgement will be used to select the l. Essentially, What this formula describes is that if the experimental treatment has an effect of at least $l\delta_1$ (e.g. it has efficiency equal to the 90 percent of efficiency of the standard treatment) then the two treatments do not differ in effetiveness in a clinically significant way (which implies non-inferiority).

Equivalence trials: An equivalence trial intends to establish that two therapies are similar; the difference between the experimental therapy and the standard therapy are not large in either direction. This is a subtle difference between the terms of equivalence and non-inferiority, which are two different types of clinical trials but cause some confusion and in many cases are used interchangeably. To establish equivalence of two treatments, we need to specify the equivalence margin $\delta > 0$, which is the maximal difference (e.g. difference in sample means, response rates etc) between the two therapies that is considered clinically insignificant. The process of selection for δ is similar to the non-inferiority case. In the case of comparing population means the hypotheses would be of the form:

$$H_0 : |\mu_1 - \mu_2| \geq \delta$$

$$H_1 : |\mu_1 - \mu_2| < \delta$$

Thus, equivalence trials are designed to establish whether the experimental treatment is better or worse than the standard one by a prespecified margin that is clinically acceptable and thus that the two therapies have more or less the same effect.

Equivalency designs are not very common in therapeutic trials evaluating effectiveness because the study objective often is to show that a new treatment is not inferior to a standard, which corresponds to a noninferiority design[32]. Yet equivalency trials may arise in cases where a different formulation of an already existing

treatment is to be tested (for example, a new version of a treatment administered orally versus the same treatment in liquid form)[33]. In many cases, equivalency trials also occur when pharmacokinetic characteristics of two therapies are to be compared. In that case, however parallel designs are not used and crossover designs are preferred. However, parallel designs can still be used in such cases if, for specific reasons, there is no evidence of large intra-subject variability or the trial must terminate early (for example, when very ill patients are participating).[33]

3.2 Endpoints In Phase III Trials

In this section statistical methodoogy for different types of endpoints will be presented. The most common endpoints in phase III clinical trials are quantitative, dichotomous and survival endpoints.

3.2.1 Survival Endpoints

Many clinical trials are designed to assess the effect of an intervention on therapeutic areas such as oncology. In such therapeutic areas the endpoints that need to be assessed are either the survivability of patients or factors that are related to the survivability of patients. More specifically common endpoints are[34]:

- Overall Survival (OS): The most common endpoint in survival analysis applications.Overall survival (OS) is defined as the time from randomization to death. Since the goal of cancer treatment is generally to extend survival, OS is often referred to as the gold standard endpoint in oncology clinical trials.
- Progression free survival (PFS) is defined as the time from randomization until first evidence of disease progression or death. PFS is measured by censoring and patients who are still alive at the time of evaluation or those who were lost to follow up.
- Time to progression (TTP) is defined as the time from randomization until first evidence of disease progression. Since PFS and TTP are similar, it is important for studies to clarify what is meant by evidence of disease progression.
- Disease free survival (DFS) is defined as the time from randomization until evidence of disease recurrence.
- Event-free survival (EFS) is defined as the time from randomization to an event which may include disease progression, discontinuation of the treatment for any reason, or death. While EFS and DFS used to be interchangeable, the patient is not technically “disease-free” at the time of randomization.

It is evident that in any of the above endpoints, there is a specific outcome of interest (death,progression of disease etc) and the purpose of a clinical trial using this endpoint is to evaluate the length of time until its occurrence. Thus, using such endpoints, time-to-event data arises, that is, data that measures the total time it takes for an event to occur. During these trials, follow-up of patients is regular. A strict definition of follow-up, is according to the National Cancer Institute the procedure of monitoring a person's health over time after treatment. This includes keeping track of the health of people who participate in a clinical study or clinical trial for a period of time, both during the study and after the study ends [34].

It should be noted that time-to-event studies are usually Phase III trials and thus they take a lot of years to complete. Under this context, it is for loss to follow-up to arise, that is, patients may stop participating in the ongoing trial and follow-up of the patient is infeasible. This results in a type of missing data, since it is not known whether the event of interest has occured and at what time it has occured. This is known as censoring, and specifically as right censoring because if t is the censoring time of a patient, then the interval $[t, \infty)$ indicates that there is no more data on this patient (for any time-point on the right of t). Notice however that even if a patient was not lost to follow up, at some point T_0 the trial would still stop and after this point we would still not know the true time to event of a patient that survived until the end of the trial. Thus, right censoring occurs in two cases:

- For a trial with prespecified duration T_0 a patient will be right censored if he survives until the end of the trial (or up to time T_0).
- During the trial, the patient stops participating. He wil be right censored at the time of the last-follow up u.

In applications with censoring the nature of the censoring rules may vary from observation to observation. Thus, the censoring mechanism can be considered as a random variable. Considering the variables below:

- T_i is the survival time of the i-th.
- Y_i is the time of censoring i-th.

- $X_i = min(T_i, Y_i)$ is the timepoint of death or the timepoint where the i-th patient was right censored.

Additionally consider that when X_i=Y_i we consider that the patient was not censored.

3.2.1.1 Survival Functions

Consider the random variable T, that measures the time to a specific event. Then, $F(t) = P(T \leq t)$ is its cdf. In this section we define:

- $S(t) = 1 - F(t) = P(T > t) = P(T \geq t)$, as the continuou survival function
- $h(t) = \lim_{dt \to 0} \frac{P(t<T<t+dt|T \geq t)}{dt} = \lim_{dt \to 0} \frac{P(t<T<t+dt, T \geq t)}{dtP(T \geq t)} = \lim_{dt \to 0} \frac{P(t \leq T<t+dt)}{dtS(t)} = \frac{f(t)}{S(t)}$, as the hazard function.
- $H(t) = \int_{-\infty}^{+\infty} h(t)dt$, the cumulative hazard function.

Notice that since $H'(t) = h(t) = \frac{f(t)}{S(t)}$ an antiderivative of $\frac{f(t)}{S(t)}$ is $-log(S(t))$. Indeed:

$$(-log(S(t))') = -\frac{1}{S(t)}(1 - F(t))' = \frac{1}{S(t)}f(t) = h(t) = H'(t)$$

. Thus the relation:

$$H(t) = -log(S(t))$$

occurs.

These are the formulas for the continuous cases. Some relations in the discrete case will now be presented. Consider the discrete random variable T, with mass $f_1, .., f_j$ at time points $t_1, ..t_j$. Defining $F(t_j) = P(T \leq t_j)$, as the cdf, the survival distribution will be:

$$S(t_j) = 1 - F(t_j) = P(T > t_j)$$

The hazard function will then be defined as:

$$h_j = P(T = t_j | T \geq t_j) = \frac{P(T = t_j, T \geq t_j)}{S(t_{j-1})} = \frac{P(T = t_j)}{S(t_{j-1})} = \frac{f_j}{f_j + f_{j+1} + ...}$$

An additional relation for the hazard function is:

$$\begin{aligned} h_j = P(T = t_j | T \geq t_j) = \frac{P(T \leq t_j) - P(T \leq t_{j-1})}{S(t_{j-1})} = \\ = \frac{S(t_{j-1}) - S(t_j)}{S(t_{j-1})} = 1 - \frac{S(t_j)}{S(t_{j-1})} \end{aligned} \tag{3.1}$$

Using the two previous relations one can compute a relation for the discrete survival function:

$$\begin{aligned} S(t_j) = P(T > t_j) = P(T > t_j, T > t_{j-1}) = P(T > t_j | T > t_{j-1}) P(T > t_{j-1}) = \\ = (1 - P(T \leq t_j | T \geq t_j)) P(T > t_{j-1}) = (1 - P(T = t_j | T \geq t_j)) P(T > t_{j-1}) \\ = (1 - h_j) P(T > t_{j-1}) = (1 - h_j) S(t_{j-1}) \end{aligned} \tag{3.2}$$

Additionally, if h_j is small one can consider that the discrete cumulative hazard function $H(t) = \sum_{t_j \leq t} h_j$ In the second row of (3.2) it is true that $P(T > t_j | T > t_{j-1}) = P(T > t_j | T \geq t_j)$. Indeed it is obvious that the two conditions imposed do not affect differently the probability of $\{T > t_j\}$.Additionally using this formula inductively will result in:

$$S(t_j) = P(T > t_j) = \prod_{i: t_i \leq t_j} (1 - h_i)$$

3.2.1.2 Estimation of Survival Curves

In most cases in order to estimate the survival function non-parametric methods are employed. Specifically, the Kaplan-Meier estimator of the survival function is used [35]. Below the derivation of the estimator's formula is presented.

Kaplan and Meier, based their proof on the definition of the nonparametric likelihood. Specifically, If $X_1, ..., X_n$ is a random sample with distribution function $F(.)$ then the non-parametric likelihood (NPL)is defined as:

$$L(F | x_1, ..., x_n) = \prod_{i=1}^{n} P(X_i = x_i)$$

This is similar to the parametric likelihood but instead of a parameter the distibution function is estimated. Additionally, the candidate cdf's that maximize the NPL must be of discete nature otherwise:

$$P(X_i = x_i) = 0$$

Thus, finding the non parametric maximum likelihood estimator (NPLME) translates into how much mass should be placed at each observed data value x_i. In the case of right censoring consider the distinct survival times $T_1 = t_1 < ... < T_n = t_n$ and set $t_0 = 0$ and $t_{n+1} = \infty$. Consider also as λ_i the number of censored observations in the interval $[t_i, t_{i+1})$. Consider also the censoring time of the i patient as y_i. Also for convenience set $\mathcal{D}$ and $\mathcal{C}$ the sets that contain the failure times and the censoring times respectively. For each patient either his survival time or the fact that he lived more then the time at which he was censored will be observed. The non-parametric likelihood can then be written as:

$$\begin{aligned} L(F|\mathcal{C}, \mathcal{D}) &= \prod_{i \in \mathcal{C}} P(T_i > y_i) \prod_{j \in \mathcal{D}} P(T_j = t_j) \\ &= \prod_{i \in \mathcal{C}} P(T_i > y_i) \prod_{j \in \mathcal{D}} (S(t_{j-1}) - S(t_j)) \end{aligned} \tag{3.3}$$

Notice that in order to maximize the above likelihood $S(t_{j-1}), P(T_i > y_i)$ needs to become as large as possible and $S(t_j)$ as small as possibe. Since every censoring time is inside some interval $[T_j, T_{j+1})$ and since $S(t_j) = 1 - F(t_j)$ is a decreasing function:

$$S(t_{j+1}) < S(y_i) \leq S(t_j)$$

this implies that to enlarge the likelihood we can set $S(t_j) = S(y_i)$ or in other words "put weight" only on the failure times. Utilizing this and the fact that $P(T_j > t_j) = P(T_i > t_j)$ it is straightforward that (3.3) can be written as:

$$\begin{aligned} P(T_1 > t_1)^{\lambda_1} \times ... \times P(T_n > t_n)^{\lambda_n} \times & P(T_1 = t_1)^{\delta_1} \times ... \times P(T_n = t_n)^{\delta_n} \\ &= \prod_{j=1}^{n} [P(T_j > t_j)]^{\lambda_j} [P(T_j = t_j)]^{\delta_j} = \\ &= \prod_{j=1}^{n} [P(T_j > t_j)]^{\lambda_j} [P(T_j \geq t_{j-1}) - P(T_j \geq t_j)]^{\delta_j} \end{aligned} \tag{3.4}$$

setting $\frac{S(t_j)}{S(t_{j-1})} = p_j$ for simplicity we have:

$$\begin{aligned}\prod_{j=1}^{n}(p_1 \dots p_j)^{\lambda_j}(p_1 \dots p_{j-1}h_j)^{\delta_j} &= \prod_{j=1}^{n}(\frac{1}{p_j})^{\delta_j}(p_1 \dots p_j)^{\lambda_j}(p_1 \dots p_j)^{\delta_j}h_j^{\delta_j} \\ &= \prod_{j=1}^{n}(\frac{1}{p_j})^{\delta_j}(p_1 \dots p_j)^{\delta_j+\lambda_j}(h_j)^{\delta_j}\end{aligned} \tag{3.5}$$

expanding (3.5) yields:

$$\begin{aligned}&(\frac{1}{p_1})^{\delta_1}(p_1)^{\delta_1+\lambda_1}(h_1)^{\delta_1}(\frac{1}{p_2})^{\delta_2}(p_1p_2)^{\delta_2+\lambda_2}(h_2)^{\delta_2} \times \dots \times (\frac{1}{p_n})^{\delta_n}(p_1....p_n)^{\delta_n+\lambda_n}(h_n)^{\delta_n} = \\ &(\frac{1}{p_1})^{\delta_1}(p_1)^{\sum_{i=1}^{n}\delta_i+\lambda_i}(h_1)^{\delta_1}(\frac{1}{p_2})^{\delta_2}(p_2)^{\sum_{i=2}^{n}\delta_i+\lambda_i}(h_2)^{\delta_2} \dots\end{aligned} \tag{3.6}$$

setting

$$n_j = \sum_{i=j}^{n}\delta_i + \lambda_i \tag{3.7}$$

it is easy to see that using (3.6) and (3.7), (3.5) can be written as:

$$\prod_{j=1}^{n}(\frac{1}{p_j})^{\delta_1}(p_j)^{n_j}(h_j)^{\delta_j} = \prod_{j=1}^{n}(p_j)^{n_j-\delta_j}(h_j)^{\delta_j} = \prod_{j=1}^{n}(1-h_j)^{n_j-\delta_j}(h_j)^{\delta_j} \tag{3.8}$$

Thus $L(F|\mathcal{C},\mathcal{D})$ is maximized when $L(h_1,...h_n|\mathcal{C},\mathcal{D})$ is maximized. Thus, It remains to differentiate wrt to the discrete hazard functions:

$$lnL(\mathbf{h}|\mathcal{C},\mathcal{D}) = \sum_{j=1}^{n}(n_j-\delta_j)ln(1-h_j) + \sum_{j=1}^{n}\delta_j ln(h_j) \tag{3.9}$$

But it is easy to notice that this log-likelihood is the same as the binomial log-likelihood without the binomial coefficient. Thus, it follows that NPMLE of h will be:

$$\hat{h_j} = \frac{d_j}{n_j}$$

Since the survival functions that maximize the NPMLE are discrete and

$$S(t_j) = P(T > t_j) = \prod_{i:t_i \leq t_j} (1 - h_i)$$

the NMPLE will be

$$\hat{S}(t_j) = \prod_{i:t_i \leq t_j} (1 - \hat{h}_i)$$

Following Cox [36] the Kaplan Meier estimate can be shown to be maximize the NPMLE function in the space of all distribution functions. However, he also notes that is of small statistical implications. For example, it is straightforward that if some patients reached the end of the study then they will be censored. By the previous formulas the Kaplan Meier estimate will be undefined for time values larger than the last survival time (that is the estimate will always be positive). Thus every possible discrete survival function that puts the same weight at the observed survival times will also maximize the likelihood.

Following the estimation of the Survival Function one can estimate the hazard function through plugging the estimate of $\hat{h_j} = \frac{d_j}{n_j}$ in the formula for the discrete cumulative hazard function. This will yield the Nelson-Aalen estimator. Another estimator for the cumulative hazard function is $log(\hat{S}(t))$ where we plug the Kaplan Meier estimate on the formula for the continuous $H(t)$.

3.2.1.3 Cox's Proportional Hazards Model

Consider a vector of covariates $(x_1, \ldots, x_n)$ and let T be the failure time. Consider a class of models termed relative risk models or Cox models which are specified by the hazard relationship:

$$h(t|x_1, \ldots, x_n) = \frac{P(t \leq T \leq t + \delta t | T \geq t, x_1, \ldots, x_n)}{dt} = h_0(t)h(t, \mathbf{x})$$

where $h_0(t)$ is an arbitrary unspecified baseline hazard function and the relative risk function $h(t, x)$ specifies the relationship between the covariate x and the failure rate or hazard function. For the relative risk it is common to be defined as:

$$h(t, x_1, ..., x_n) = e^{b_1 x_1 + \cdots + b_n x_n}$$

.

In a clinical trial setting the number of covariates is usually one and it is a binary covariate indicating whether the patient received the experimental therapy or the standard therapy. Assuming the previous model for the hazard function given the covariate:

$$h(t|x) = h_0(t)e^{bx}$$

There are two parameters in this model that need to be estimated $h_0(t)$ and b, but we are interested in estimating b. This is because for two observations, one in the experimental group and one in the placebo group:

$$\frac{h(t|x=1)}{h(t|x=0)} = \frac{h_0(t)e^b}{h_0(t)} = e^b$$

Thus in order to assess the differences in hazards between patients we only need to estimate b using the data acquired during the clinical trial. More importantly however, the above formula implies that the hazard ratios do not depend on time and are fixed. This is called the proportionality of hazards and it is a very important assumption for the model. Of course, the above results and assumptions still occur for the case of having multiple covariates.

In order to perform inference for b a description of the term "partial likelihood" (a term coined by Cox and used in the inferential procedures of the proportional hazards model) is presented [37]. Let Y be a random variable with density $f(y;\theta)$. Assume that Y is transformed (without using the parameter θ) into a set of random variables $(X_1, S_1, \ldots, X_n, S_n)$. Then the likelihood of $(X_1, S_1, \ldots X_n, S_n)$ is:

$$\begin{aligned} f(x_1, s_1, \ldots, x_n, s_n; \theta) = & \\ &= f(x_n|x_1, \ldots, x_{n-1}, s_1, .., s_n; \theta) f(x_1, \ldots, x_{n-1}, s_1, .., s_n; \theta) \\ &= f(x_n|x_1, ..., x_{n-1}, s_1, .., s_n; \theta) \times \\ &\times f(s_n|x_1, ..., x_{n-1}, s_1, .., s_{n-1}; \theta) f(x_1, ..., x_{n-1}, s_1, ..., s_{n-1}; \theta) \\ &= \prod_{j=1}^{n} f(x_j|x_1, ..., x_{j-1}, s_1, ..., s_j; \theta) \prod_{j=1}^{n} f(s_j|x_1, ..., x_{j-1}, s_1, \ldots, s_{j-1}; \theta) \end{aligned} \tag{3.10}$$

This is the full likelihood. The partial likelihood is defined by Cox as the first product of equation (3.10). That is instead of using the whole likelihood for inferences

regarding θ we only use the first product. When the first product contains almost all the information about θ this framework simplifies the computations needed but also there is not substantial information loss that could contribute to the estimation of θ.

In the case of a clinical trial with time to event data and no ties in failure time (i.e. those who failed, failed at different times) we observe a set of survival $t_1, ...t_n$ and censoring $c_{n+1}, ...c_k$ times. The censoring times might have a different order but that does not matter. We also define

- $(\tau_1, \ldots, \tau_n)$, the ordered failure times
- $(i_1, ..., i_n)$, where i_j is the item that had the the j-th largest survival time.
- $\mathbf{C}_1, \ldots, \mathbf{C}_n$ where $\mathbf{C}_j$ is the censoring information (i.e. who was censored) at the time interval $[\tau_{j-1}, \tau_j)$.

One can also denote:

$$\mathbf{A}_j = (\mathbf{C}_j, \tau_j)$$

which is actually the information of who was censored in the time interval $[\tau_{j-1}, \tau_j)$ plus the information that there was one failure time at τ_j. It is evident that each survival time corresponds to an ordered survival time and to a label, that is if $T_{(1)} = \tau_1, I_1 = 5$ then this is equivalent to $T_5 = \tau_1$. Thus using the previous the sequence $t_1, ...t_n, c_{n+1}, ...c_k$ and consequently the corresponding likelihood can be transformed into:

$$\begin{aligned} f(i_1, \mathbf{A}_1, \ldots, i_n, \mathbf{A}_n; b, h_0(t)) = \prod_{j=1}^{n} f(i_j | i_1, ...i_{j-1}, \mathbf{A}_1, \ldots, \mathbf{A}_j; b, h_0(t)) \times \\ \times \prod_{j=1}^{n} f(\mathbf{A}_j | i_1, ...i_{j-1}, \mathbf{A}_1, ..., \mathbf{A}_{j-1}; b, h_0(t)) \end{aligned} \tag{3.11}$$

An intuitive comment one can make, is that the parameter b does affect the survivability of patients but it does not affect the probability of censoring since censoring is considered independent of the covariates. Additionally, since the functional form of $h_0(t)$ is not restricted, the baseline hazard function can be selected in such a way that it would describe the survival times arbitrarily well (for example we could set the baseline hazard function as a spike function at the survival times)[38].As a result

there is not a lot of information in the second product for β so most of the information than can be acquired for *beta* is located in the first product. Thus using inference only in the first product does not result in great information loss. Cox followed this logic and used only the first product for the estimation of b. Expanding $\mathbf{A}_j$ at time τ_j the contribution to the partial likelihood will be:

$$p(i_j|\tau_1, ..., \tau_j, i_1, \ldots, i_{j-1}, \mathbf{C}_1, ..., \mathbf{C}_j)$$

By this relation one can notice that $\{\tau_1, ..., \tau_j, i_1, \ldots, i_{j-1}, \mathbf{C}_1, ..., \mathbf{C}_j\}$ implies who was at risk at time τ_j. From now one, we will denote:

$$\mathcal{R}(\tau_j) = \{\text{people who were at risk at time } \tau_j\}$$

Notice that $\mathcal{R}(\tau_j)$ contains the information on subjects that lived at least τ_j time. Additionally, $\{i_1, \ldots, i_{j-1}, \mathbf{C}_1, \ldots, \mathbf{C}_j, \tau_1, \ldots, \tau_j\}$ provides information about who was censored just before time τ_j and also who has died at times $\tau_1, ...\tau_{j-1}$. It is obvious thus that $(\mathcal{R}(\tau_j), \tau_j)$ contains all the information that $\{i_1, \ldots, i_{j-1}, \mathbf{C}_1, \ldots, \mathbf{C}_j, \tau_1, \ldots, \tau_j\}$ provide for the event of i_j (notice that $R(t_j)$ does not provide a label of who was censored or censored before but the labels themselves do not affect the probability of i_j). with these in mind we have that

$$\begin{aligned}
&p(i_j|\tau_1, ..., \tau_j, i_1, ..., i_{j-1}, \mathbf{C}_1, ..., \mathbf{C}_j) \\
&p(i_j|\tau_1, ..., \tau_j, i_1, ..., i_{j-1}, \mathbf{C}_1, ..., \mathbf{C}_j, \mathcal{R}(\tau_j)) \\
&= p(i_j|\tau_j, \mathcal{R}(\tau_j)) = \frac{p(i_j, \tau_j, \mathcal{R}(\tau_j))}{p(\tau_j, \mathcal{R}(\tau_j)} \\
&= \frac{p(i_j, \tau_j|\mathcal{R}(\tau_j))}{p(\tau_j|\mathcal{R}(\tau_j)} \times 1
\end{aligned} \tag{3.12}$$

Now notice that $P(I_j = i_j, T_{(j)} = \tau_j) = P(T_{i_j} = \tau_j)$ thus continuing for (3.12)

$$\begin{aligned}
&= \frac{p(i_j, \tau_j|\mathcal{R}(\tau_j)}{p(\tau_j|\mathcal{R}(\tau_j)}) = \\
&= \frac{p(i_j \text{ dies at time } \tau_j|\text{those who were at risk at time were})}{p(\text{someone dies at time } \tau_j|\text{those who were at risk at time } \tau_j : \mathcal{R}(\tau_j)}) \\
&= \frac{P(T_{i_j} = \tau_j|T_{i_j} > \tau_j)}{\sum_{l \in R(\tau_j)} P(T_{i_l})}
\end{aligned} \tag{3.13}$$

Now notice that the formula in (3.13) implies

$$\frac{P(T_{i_j} = \tau_j | T_{i_j} > \tau_j)}{\sum_{l \in R(\tau_j)} P(T_{i_l})} = \frac{h_0(t) h_{i_j}(t_j)}{\sum_{l \in R(\tau_j)} h_0(t) h_{i_l}(t_j)} = \frac{h_{i_j}(t_j)}{\sum_{l \in R(\tau_j)} h_{i_l}(t_j)} \tag{3.14}$$

For the case of tied observations, denote $d_{\tau_1}, ..., d_{\tau_k}$ the number of multiplicities at each failure time point. Using similar arguments for the partial likelihood as before the contribution to the partial likelihood will eventually be:

$$p(i_{j,1}, ..., i_{j,d_{\tau_j}} | \tau_1, \tau_2, ...d_{\tau_1}, ..., d_{\tau_j}, ...i_1, ...i_j, \mathbf{C}_1, ..., \mathbf{C}_j) \tag{3.15}$$

This is the probability of observing items $i_{j,1}, ..., i_{j,d_{\tau_j}}$ failing while having observed the previous failure times, all the censoring information just before τ_j, the items that failed previously and the number of failures at times $\tau_j, ..., \tau_1$.

Following the same framework as before expanding the conditional probability will eventually yield the formula for the j-th contribution as:

$$\frac{h_{i_{j,1}}(t_j)...h_{i_{j,d_{\tau_j}}}(t_j)}{\sum_{l \in s(j,d)} h_{l_1}(t_j)...h_{l_{d_{\tau_j}}}(t_j)}$$

where $s(j, d)$ is the set of all possible d-combinations from the risk set at time τ_j (i.e. all the possible ways that d_{τ_j} can be observed when the people at risk at τ_j is R_{τ_j}).

For the case of untied observations, assuming that $h(t) = h_0(t)e^{\theta x}$ the partial likelihood function becomes:

$$L(\theta) = \prod_{j=1}^{n} \frac{e^{\theta x_j}}{\sum_{l \in R(t_j)} e^{\theta x_l}}$$

taking ln's will yield

$$lnL(\theta) = \sum_{j=1}^{n} (\theta x_j - ln(\sum_{l \in R(t_j)} e^{\theta x_l})) = \sum_{j=1}^{n} \theta x_j - \sum_{j=1}^{n} ln(\sum_{l \in R(t_j)} e^{\theta x_l}))$$

Defining the previous set:

$$\{\tau_1, ..., \tau_j, i_1, ..., i_{j-1}, \mathbf{C}_1, ..., \mathbf{C}_n\}$$

as r_j-the history of the event up until the j-th death and differentiating wrt to θ:

$$\frac{dlnL(\theta)}{d\theta} = \sum_{j=1}^{n} \frac{d \log f(i_j|r_j;\theta)}{d\theta} = \sum_{j=1}^{n} x_j - \sum_{j=1}^{n} [\frac{1}{\sum_{l \in R(\tau_j)}^{n} e^{\theta x_l}} \sum_{l \in R(\tau_j)} e^{\theta x_l} x_l]$$

in order to solve

$$\frac{dlnL(\theta)}{d\theta} = 0$$

computational algorithms and specifically the Newton-Rapshon methods must be used. Differentiating again wrt to θ we will have:

$$\frac{d^2 lnL(\theta)}{d\theta^2} = -((\frac{1}{\sum_{l \in R(\tau_j)}^{n} e^{\theta x_l}})^2 \sum_{l \in R(\tau_j)} e^{\theta x_l} x_l + \frac{1}{\sum_{l \in R(\tau_j)}^{n} e^{\theta x_l}} \sum_{l \in R(\tau_j)} e^{\theta x_l} x_l^2)$$

The same methodology is followed in the case of tied observations or in the case of multiple covariates.

Then the score function of the partial likelihood will be

$$U = \sum_{j=1}^{n} \frac{d \log f(i_j|r_j;\theta)}{d\theta} = \sum_{j=1}^{n} U_j$$

If one takes the expected value of U, $E(U)$ then it will be with respect to both I_j and $\mathcal{R}_j$ since the history is actually a set of random variables. However by taking a conditional expectation of

$$E(U_j|\mathcal{R}_j = r_j) = E(\frac{d \log f(I_j|r_j;\theta)}{d\theta}) = \sum_{j=1}^{n} \frac{d \log f(i_j|r_j;\theta)}{d\theta} f(i_j|r_j;\theta) = 0$$

The last equality follows immediately after expanding the differentiation of $\frac{d \log f(i_j|r_j;\theta)}{d\theta}$.Here n denotes for simplicity the observed number of people at risk at $\tau_{(j)}$.

Also for $j > k$

$$E(U_j U_k|\mathcal{R}_j = r_j) = U_k E(U_j|\mathcal{R}_j = r_j) = 0$$

Following similar arguments it can be also shown that:

$$E(\frac{d^2 \log f(i_j|r_j;\theta)}{d\theta^2}|r_j) = -E((\frac{d \log f(i_j|r_j;\theta)}{d\theta})^2|r_j)$$

Thus,

$$Var(U_j|\mathcal{R}_j = r_j) = E(U_j^2) = -E((\frac{d \log f(i_j|r_j;\theta)}{d\theta})^2|\mathcal{R}_j = r_j)$$

Since all these previous relations apply for any observed history it will also be true that:

$$E(U_j) = E(E(U_j|\mathcal{R}_j)) = 0$$

and similarly

$$Var(U_j) = -E((\frac{d \log f(i_j|r_j;\theta)}{d\theta})^2)$$
$$E(U_j U_k) = 0$$

for $j > k$, finally yielding

$$E(U) = 0$$
$$Var(U) = \sum_{j=1}^{n} Var(U_j)$$

Note that U_j, U_k are uncorrelated but that does not mean that they are independent. Assuming however that there is some mild independence between them and that the partial likelihood estimate $\hat{\theta}$ is consistent (which initially was not proved rigorously but intead many simulations showed that the partial likelihood estimate was close to the real parameters when the sample size was large. The consistency of these estimators was proved in a strict manner first by Tsiatis in 1981, 9 years after Cox's paper on proportinal hazards[39]), one can use the same arguments for the MLE's normal asymptotic distribution to derive that the partial likelihood estimate is asymptotically normally distributed. This result implies that Wald's test, Rao's test and likelihood ratio test for the coefficients and model comparisons are valid. This was also proved by Tsiatis [39] without using this type of conditions.

3.2.1.4 Residuals and Diagnostics For Cox's Model

A number of residuals can be used to assess the goodness of fit of a proportional hazards model, the condition of proportional hazards itself and other parameters such as outliers and influential observations. This dissertation is mostly centered on the Schoenfeld residuals which can be used to assess the proportionality of hazards assumption and the Cox-Snell residuals to assess the goodness of fit of the model.

The proportionality of hazards assumption should be checked always during a trial since a violation of this assumption will make the log rank test less powerful. If a violation of proportional hazards is observed than another test (weighted log-rank test) and a sample-size re-calculation could be used to acquire the desirable power.

Following Therneau and Grambsch [40] suppose that we have one covariate which is the type of treatment (standard vs experimental). The theory below can be generalized to multiple covariates. Additionally, assume that no-ties in survival times are observed. Define $\mathcal{R}_j$ (see section of Cox's proportional hazards model),the history, as before. Also define x_j as the covariate value of the j-th person that died. Given $\mathcal{R}_j$ we do not know who dies and thus not his covariate value which means that x_j is also a random variabe and it can be easily calculated that:

$$E(x_j|\mathcal{R}_j = r_j) = E(x_j|T_j = \tau_j, \mathcal{R}(\tau_j) = r_{\tau_j}) = \frac{\sum_{l \in r_{\tau_j}} x_l e^{bx_l}}{\sum_{l \in r_{\tau_j}} e^{bx_l}}$$

Note that the expected value does not actually depend on τ_j. Define then the Schoenfeld residuals [41],

$$c_j = x_j - E(x_j|\mathcal{R}_j = r_j)$$

Now suppose that the proportional hazards could be violated which means that the effect of treatment changes with time and specifically:

$$b(t) = b + g(t)\theta$$

Additionaly the model describing the hazard function given the covariates will be

$$h(t) = h_0 e^{b(t)x}$$

where x is a dummy variable denoting what treatment was given. As with the proportional hazards the expected value of the covariate value given the history is:

$$E(x_j|\mathcal{R}_j = r_j) = E(x_j|T_j = \tau_j, \mathcal{R}(\tau_j) = r_{\tau_j}) = \frac{\sum_{l \in r_{\tau_j}} x_l e^{b(\tau_j)x_l}}{\sum_{l \in r_{\tau_j}} e^{b(\tau_j)x_l}} = M(b(\tau_j), \mathcal{R}(\tau_j))$$

since the history contains the value of the j-th survival time and the risk set at that time (also, the last equality above could be seen, in a gerenal scenario, as defining a function $g(f(t), y) = E(X|T = t, Y = y)$ where f is only a function of t). The Schoenfeld residuals described previously can be written as

$$\begin{aligned} c_j &= [x_j - M(b(\tau_j), r_{\tau_j})] + [M(b(\tau_j), r_{\tau_j}) - M(b, r_{\tau_j})] \\ &\approx [x_j - M(b(\tau_j), r_{\tau_j})] + M(b, r_{\tau_j}) + \frac{dM(b(\tau_j), r_{\tau_j})(b)}{db(\tau_j)}(b(\tau_j) - b) - M(b, r_{\tau_j})) \\ &\approx [x_j - M(b(\tau_j), r_{\tau_j})] + \frac{dM(b(\tau_j), r_{\tau_j})(b)}{db(\tau_j)} g(\tau_j)\theta \end{aligned} \tag{3.16}$$

where the second equality occurs by expnding $M(b(\tau_j), \tau_j)$ around $b(\tau_j) = b$. Additionally, following simple algebraic calculations it can be shown that:

$$\frac{dM(b(\tau_j), r_{\tau_j})}{db(\tau_j)}(b) = Var(c_j|\mathcal{R}_j = r_j)$$

$$E(c_j|\mathcal{R}_j = r_j) \approx g(\tau_j)\theta Var(c_j|\mathcal{R}_j = r_j)$$

where $Var(c_j|\mathcal{R}_j = r_j)$ here denotes the variance of the schoenfeld residual under the proportional hazard model. By using the fact that $\theta = \frac{b(t)-b}{g(t)}$ we finally get:

$$E(c_j|\mathcal{R}_j = r_j) \approx (b(\tau_j) - b)Var(c_j|\mathcal{R}_j = r_j) \tag{3.17}$$

In order to get rid of the variance term one could define the scaled Schoenfeld residuals as:

$$c^* \approx Var(c_j|\mathcal{R}_j = r_j)^{-1} c_j$$

for which

$$E(c_j^*|\mathcal{R}_j = r_j) \approx b(\tau_j) - b$$

That final relation implies that if $b(t) = b$ that is, the proportional hazards assumption in not violated then we would expect $c_t^* + b$ given the history to be a horizontal line. Thus estimating the Schoenfeld residuals and their variance as well as the coefficient b and plotting them against time is a graphical test to test proportionality of hazards.

Note that the Schoenfeld residuals can be used for any type of covariate, continuous or dichotomous. For a dichotomous variable one could simply visualize the estimated survival function for each group and note whether the distance between the survival functions changes significantly wrt time (for example the survival functions cross, or initially the survival functions are close to each other but as time passes the difference in survival becomes noticeably larger). However plotting the survival functions could only work in the case of dichotomous covariates.

The above procedure and proof can also be extended for multiple covariates. In such a case, calculate the vector of Schoenfeld residuals for the different covariates as:

$$E(\mathbf{x}|r_j) = \frac{\sum_{l \in r_{\tau_j}} \mathbf{x}_l' e^{\mathbf{b}'\mathbf{x}_l}}{\sum_{l \in r_{\tau_j}} e^{\mathbf{b}'\mathbf{x}_l}}$$

Proceeding with similar frameworks yields:

$$E(\mathbf{c}|r_j) = \mathbf{V}(\mathbf{c}_j|r_j)\mathbf{G}(\tau_j)\theta$$

where $\mathbf{V}$ is the covariance matrix of the Schoenfeld residuals, $\mathbf{G}$ is a diagonal matrix that contains the function $g(t)$ and θ is a constant vector. Consequently the residuals for the Schoenfeld residuals are:

$$E(\mathbf{c}^*|r_j) = \mathbf{G}(\tau_j)\theta$$

$\mathbf{V}$ can be estimated by $d\hat{\mathbf{V}}(\hat{\mathbf{b}})\hat{\mathbf{c}}$, where d is the number of events from which it follows that a graphical test for the proportionality of hazards would be to check if the sum $\hat{c}_{kt}^* + \hat{b}_k$ (where $\hat{c}_{kt}^*$ denotes the estimated Schoenfeld residual of the k-th covariate at time t) given the history forms a paralell line to the axis of time.

Moving on with the second type of residuals, the Cox-Snell residuals, consider first the random variable

$$-\log[S(T)] = -\log[1 - F(T)]$$

where T is a random variable and S is its survival function and F its cdf. Then it is known that $-\log S(T) \sim \mathcal{E}(1)$, that is, it is exponentially distributed with mean 1. If the model fitted to the data is satisfactory the estimated probability of survival time of the i-th patient, $\hat{S}_i(t_i)$ will be close to the true probability of survival time $S_i(t_i)$. Thus $-log[\hat{S}_i(t)]$ will also be close to $-log[S_i(t)]$. Thus, if the model is adequate it will be expected that the $-log[\hat{S}_i(t_i)]$'s are described by an exponential distribution.Setting $r_{cs} = -log[\hat{S}_i(t_i)]$ we get the Cox-Snell residuals [49]. Since, for an exponential distribution with rate λ the cumulative hazard function is $H(t) = \lambda t$ we should expect that:

$$H(r_{cs}) \approx r_{cs}$$

In order to construct a graphical plot to assess the goodness of fit for a model, and since we a have censored sample from a unit exponential distribution, we can replace the Hazard function H of the cox-snell residuals with its estimate $\hat{H}(r_{cs}) = -log\hat{S}(r_{cs})$.Then, it remains to assess whether approximately:

$$\hat{H}(r_{cs}) \approx r_{cs}$$

or if by plottig $\hat{H}(r_{cs})$ vs r_{cs} we get a straight line with unit slope.

Note that $-log[\hat{S}_i(t)] = \hat{H}_0(t)e^{\hat{b}' x_i}$, thus an estimator for the cumulative baseline hazard function is needed. One can use Breslow's estimator which is given as:

$$\hat{H}_0(t) = \sum_{t_{(j)} \leq t} \frac{d_j}{\sum_{l \in \mathcal{R}(\tau_j)} e^{\hat{b}' x_l}}$$

The derivation of this estimator uses the NPMLE approach as the Kaplan-Meier estimation procedure. A detailed descripition of how this estimator occurs is given in [48].

3.2.1.5 Log Rank Test

The treatment effect of a trial with survival endpoints is often detected by the log-rank test, which is a nonparametric procedure for testing the hypothesis of equality of two or more survival distribution functions. For example, the hypothesis of interest in a two-arm trial is

$$H_0 : S_1(t) = S_2(t)$$
$$H_1 : S_1(t) \neq S_2(t)$$

where $S_1(t)$ and $S_2(t)$ represent the survival distribution of two groups with indexes 1 and 2.

Assume that the unique and ordered failure times for two groups are denoted by $t_1 < t_2 < ..... < t_k$. Let d_{1j} be the number of failures and n_{1j} be the number at risk in group 1 at time t_j . Let d_{2j} and n_{2j} be the corresponding number in group 2. Then, $d_j = d_{1j} + d_{2j}$ represents the number of failures in both groups at time t_j , and $n_j = n_{1j} + n_{2j}$ is the number at risk in both groups at time t_j .

Within the interval $[t_j, t_j + dt)$, where dt is defined as a small increment of time, then we have, conditional on n_{ij} , that d_{ij} has, approximately, a binomial distribution with mean $E(d_{ij}) = n_{ij}h_{ij}dt$, where $h_{ij} = h_i(t_j)$ is the hazard rate for group i at time t_j. Following Cook and DeMets (2005)[42], the joint distribution of d_{1j} and d_{2j} conditional on n_{1j} and n_{2j} is

$$\begin{aligned} p(d_{1j}, d_{2j}|n_{1j}, n_{2j}) &= p(d_{1j}|n_{1j})p(d_{2j}|n_{2j}) \\ &= \binom{n_{1j}}{d_{1j}}\binom{n_{2j}}{d_{2j}}(\lambda_{1j}dt)^{d_{1j}}(1 - h_{1j}dt)^{n_{1j}-d_{1j}}(h_{2j}dt)^{d_{2j}}(1 - h_{2j}dt)^{n_{2j}-d_{2j}} \\ &\approx \binom{n_{1j}}{d_{1j}}\binom{n_{2j}}{d_{2j}}h_{1j}^{d_{1j}}h_{2j}^{d_{2j}}dt^{d_j} \end{aligned} \tag{3.18}$$

Additionally the conditional distribution of d_{1j}, d_{2j} is:

$$\begin{aligned} p(d_{1j}, d_{2j}|n_{1j}, n_{2j}, d_j) =& \frac{p(d_{1j}, d_{2j}|n_{1j}, n_{2j})}{\sum_{s=0}^{d_j} p(s, d_j - s|n_{1j}, n_{2j})} \approx \frac{\binom{n_{1j}}{d_{1j}}\binom{n_{2j}}{d_{2j}}h_{1j}^{d_{1j}}h_{2j}^{d_{2j}}}{\sum_{s=0}^{d_j}\binom{n_{1j}}{s}\binom{n_{2j}}{d_j-s}h_{1j}^{s}h_{2j}^{d_j-s}} \\ =& \frac{\binom{n_{1j}}{d_{1j}}\binom{n_{2j}}{d_{2j}}\delta^{d_{2j}}}{\sum_{s=0}^{d_j}\binom{n_{1j}}{s}\binom{n_{2j}}{d_j-s}\delta^{d_j-s}} \end{aligned} \tag{3.19}$$

where δ is the hazard ratio (hazard of experimental group on numerator).

Since the null hypothesis $H_0 : h_1(t) = h_2(t)$ is equivalent to $H_0 : \delta = 1$ (or $\log(\delta) = 0$). Thus, under the null hypothesis, the conditional distribution simplifies as

$$p(d_{1j}, d_{2j}|d_j, n_{1j}, n_{2j}) = \frac{\binom{n_{1j}}{d_{1j}}\binom{n_{2j}}{d_{2j}}}{\sum_{s=0}^{d_j}\binom{n_{1j}}{s}\binom{n_{2j}}{d_j-s}} = \frac{\binom{n_{1j}}{d_{1j}}\binom{n_{2j}}{d_{2j}}}{\binom{n_j}{d_j}} \tag{3.20}$$

which is the hypergeometric distribution. As a results Under H_0 the conditional expectation and variance of d_{1j} are:

$$e_{1j} = E(d_{1j}|n_{1j}, n_{2j}, d_j) = \frac{n_{1j}d_j}{n_j}$$

$$u_{1j} = Var(d_{1j}|n_{1j}, n_{2j}, d_j) = \frac{n_{1j}n_{2j}d_j(n_j - d_j)}{n_j^2(n_j - 1)}$$

Thus one can intuitively construct the test statistic:

$$U = \sum_{j=1}^{k}(d_{1j} - e_{1j}) \tag{3.21}$$

which can be considered as a measure of divergence from the expected deaths under the null hypothesis. Notice that $d_{1j} - e_{1j}$ are not independent over the observed failure times; however, it can be shown that they are uncorrelated [47]. The variance of U can then be estimated by

$$V = \sum_{j=1}^{k} \frac{n_{1j}n_{2j}d_j(n_j - d_j)}{n_j^2(n_j - 1)}$$

Using the Martingale CLT it can be shown that asymptotically under H_0 the log-rank test L [43]:

$$L = \frac{U}{\sqrt{V}} \xrightarrow{D} \mathcal{N}(0, 1) \tag{3.22}$$

This proof is ommited because it contains measure theoretic and martingale approaches that are out of the scope of this dissertation.

In order to derive an asymptotic distribution for the alternative hypothesis local alternatives must be used. Following Schoenfeld (44) consider a local alternative of the form: $H_{1n} : -log(\delta) = \frac{b}{\sqrt{n}}$ with $b < \infty$.

It can be shown that under this local alternative the asymptotic distribution of the log rank test L is approximately

$$L \xrightarrow{D} \mathcal{N}(E, 1) \tag{3.23}$$

Where ω_1, ω_2 are the proportions of patients randomized to each arm $E = b\sqrt{\omega_1\omega_2 P}$ and

$$P = \omega_1 p_1 + \omega_2 p_2$$

where

$$p_i = \int_0^\infty h_i(t) S_i(t) G(t) dt$$

where $G(t)$ is the common survival distribution of censoring times.

3.2.2 Quantitative and Dichotomous Endpoints

Irrespective of its classification a clinical trial may be designed to assess a quantitative effect. A comparison of mean difference between the experimental and standard treatment is then applied. Even though it is not very common phase III trials that compare risk difference using response rates might also be used. The hypotheses For each classification in the case of quantitative outcomes were presented previously as examples. Similaly, for assessing risk difference the null and alternative hypothesis will be:

$$H_0 : p_1 \leq p_2$$
$$H_1 : p_1 > p_2$$

$$H_0 : p_1 \leq p_2 - \delta$$
$$H_1 : p_1 > p_2 - \delta$$

$$H_0 : |p_1 - p_2| \geq \delta$$
$$H_1 : |p_1 - p_2| < \delta$$

For superiority, non-inferiority and equivalence trials respectively.

In the case of quantitative outcomes proper test statistics need to be acquired to test the hypotheses of superiority, non-inferiority and equivalency. Fora each of these cases consider that there are two arms and

$$X_i \sim \mathcal{N}(\mu_1, \sigma_1^2), \text{ and } Y_i \sim \mathcal{N}(\mu_2, \sigma_2^2)$$

are the quantitative responses of the ith participant in group 1 and 2 respectively. Notice that assuming these distributions a convention is made that the variances are known and that they are equal. The equality of variance is not obligatory and formulas in the cases of unequal variance can be derived similarly. The fact that the variances are considered known however is not trivial. This is because the sample size formulas depend on the variance and thus they must be considered known. With these conventions the unknown parameters in the sample $(X_1, ..., X_n, Y_1, ..., Y_m)$ are the true means. Thus the parameter space will be of the form:

$$\Theta = \{(\mu_1, \mu_2) : \mu_1 \in \mathcal{R}, \mu_2 \in \mathcal{R}\}$$

When the parameters spaces of the null and alternative hypothesis have union equal to the parameter space the LRT test statistic can be applied by computing:

$$\lambda = \frac{sup_{\theta \in \Theta_0} L(\theta | x_1, \ldots, y_m)}{sup_{\theta \in \Theta} L(\theta | x_1, \ldots, y_m)}$$

where L is the likelihood function. The null hypothesis will be rejected when:

$$\lambda < c$$

where c is a specific constant.

For any of the classifications of clinical trials mentioned, the union of the alternative and null hypothesis gives the parameter space. Defining Θ_0,Θ_1 as the parameter spaces for the null and alternative hypothesis respectively, in the case of non-inferiority we will have:

$$\Theta_0 = \{(\mu_1, \mu_2) : \mu_1 \leq \mu_2 - \delta\}$$

$$\Theta_1 = \{(\mu_1, \mu_2) : \mu_1 > \mu_2 - \delta\}$$

$$\Theta = \Theta_0 \cup \Theta_1$$

Following the procedure of finding a maximum in Θ for the likelihood:

$$lnl(\theta | x_1, ..., y_m) = -\frac{n}{2} ln(2\pi\sigma_1^2) - \frac{m}{2} ln(2\pi\sigma_2^2) - \frac{\sum_{i=1}^{n}(x_i - \mu_1)}{2\sigma_1^2} - \frac{\sum_{i=1}^{n}(y_i - \mu_2)}{2\sigma_2^2}$$

differentiating w.r.t. μ_i and setting the derivative equal to zero will yield:

$$\frac{\partial lnl(\theta|x_1,\ldots,y_m)}{\partial \mu_i} = \frac{\sum_{i=1}^{n}(x_{1i}-\mu_i)}{2\sigma_i^2} = 0$$

and thus the estimators of μ_i is $\hat{\mu}_1 = \bar{X}$ and $\hat{\mu}_2 = \bar{Y}$.

However if $\bar{X} > \bar{Y} + \delta$ then $(\bar{X}, \bar{Y}) \in \Theta$. Thus another set of values needs to be found. For any fixed μ_2 the likelihood is a symmetric function that has its maximum on $\hat{\mu}_1 = \bar{X}$. For smaller values of $\bar{X}$ the likelihood is an increasing function and for larger values the likelihood is decreasing. It is evident that if $\hat{\mu}_1 = \bar{X} > \mu_2 + \delta$, because of the monotonicity of the likelihood function, the next possible maximum for which the constraint is satisfied would be to set $\hat{\mu}_1 = \mu_2 + \delta$. Using these facts it holds that for any $(\mu_1, \mu_2) \in \Theta_0$:

$$l(\mu_2,\mu_1|x_1,\ldots,y_m) \leq l(\mu_2,\mu_2+\delta|x_1,\ldots,y_m) \leq max_{\mu_2} l(\mu_2,\mu_2+\delta|x_1,\ldots,y_m) \leq l(\bar{Y},\bar{X}|x_1,\ldots,y_m)$$

This implies that if $\bar{X} > \bar{Y} + \delta$ than for any $(\mu_1, \mu_2) \in \theta_0$

$$l(\mu_2,\mu_1|x_1,\ldots,y_m) < max_{\mu_2} l(\mu_2,\mu_2+\delta|x_1,\ldots,y_m)$$

The loglikelihood function will then be

$$lnl(\theta|x_1,\ldots,y_m) = \frac{-n}{2}ln(2\pi\sigma^2) - \frac{n}{2}ln(2\pi\sigma^2) - \frac{\sum_{i=1}^{n}(x_i-\mu_2-\delta)^2}{2\sigma_1^2} - \frac{\sum_{i=1}^{m}(y_i-\mu_2)^2}{2\sigma_2^2}$$

Differentiating again will yield:

$$\frac{\partial lnl(\theta|x_1,...,y_m)}{\partial \mu_2} = \frac{\sum_{i=1}^{n}(x_i-\mu_2-\delta)}{\sigma_1^2} + \frac{\sum_{i=1}^{m}(y_i-\mu_2)}{\sigma_2^2} = 0$$

or

$$\frac{n\bar{X}-n\mu_2-n\delta}{\sigma_1^2} + \frac{m\bar{Y}-m\mu_2}{\sigma_2^2} = 0$$

Thus solving for μ_2 yields:

$$\hat{\mu}_2 = \frac{\frac{n}{\sigma_1^2}\bar{X} + \frac{m}{\sigma_2^2}\bar{Y} - \frac{n}{\sigma_1^2}\delta}{\frac{n}{\sigma_1^2}+\frac{m}{\sigma_2^2}} \text{ and } \hat{\mu_1} = \hat{\mu}_2 + \delta \hat{=} \frac{\frac{n}{\sigma_1^2}\bar{X} + \frac{m}{\sigma_2^2}\bar{Y} + \frac{m}{\sigma_2^2}\delta}{\frac{n}{\sigma_1^2}+\frac{m}{\sigma_2^2}}$$

as a result expanding the likelihood will give:

$$\lambda = \frac{sup_{\theta\in\theta_0} L(\theta|x_1, ..., y_m)}{sup_{\theta\in\theta} L(\theta|x_1, ..., y_m)} = \frac{\frac{1}{2\pi\sigma_1^2}\frac{1}{2\pi\sigma_2^2}\exp(-\sum_{i=1}^{n}\frac{(x_i-\hat{\mu}_1)^2}{2\sigma_1^2} - \sum_{i=1}^{m}\frac{(y_i-\hat{\mu}_2)^2}{2\sigma_2^2})}{\frac{1}{2\pi\sigma_1^2}\frac{1}{2\pi\sigma_2^2}\exp(-\sum_{i=1}^{n}\frac{(x_i-\bar{X})^2}{2\sigma_1^2} - \frac{\sum_{i=1}^{m}(y_i-\bar{Y})^2}{2\sigma_2^2})} \tag{3.24}$$

or

$$\lambda = \exp(-\sum_{i=1}^{n}\frac{(x_i-\hat{\mu}_1)^2}{2\sigma_1^2} - \sum_{i=1}^{m}\frac{(y_i-\hat{\mu}_2)^2}{2\sigma_2^2})\exp(\sum_{i=1}^{n}\frac{(x_i-\bar{X})^2}{2\sigma_1^2} + \sum_{i=1}^{m}\frac{(y_i-\bar{Y})^2}{2\sigma_2^2}) \tag{3.25}$$

By taking advantage of the fact that:

$$\frac{\sum_{i=1}^{n}(x_i-\hat{\mu}_1)^2}{2\sigma_1^2} - \frac{\sum_{i=1}^{m}(x_i-\bar{X})^2}{2\sigma_1^2} = \frac{-\frac{n}{2\sigma_1^2}(\frac{m}{\sigma_2^2}\bar{Y} - \frac{m}{\sigma_2^2}\bar{X} + \frac{m}{\sigma_2^2}\delta)^2}{(\frac{n}{\sigma_1^2}+\frac{m}{\sigma_2^2})^2}$$

and:

$$\frac{\sum_{i=1}^{n}(y_i-\hat{\mu}_2)^2}{2\sigma_1^2} - \frac{\sum_{i=1}^{m}(y_i-\bar{Y})^2}{2\sigma_1^2} = \frac{-\frac{m}{2\sigma_2^2}(\frac{n}{\sigma_1^2}\bar{X} - \frac{n}{\sigma_1^2}\bar{Y} - \frac{n}{\sigma_1^2}\delta)^2}{(\frac{n}{\sigma_1^2}+\frac{m}{\sigma_2^2})^2}$$

thus,

$$\begin{aligned}
\lambda &= \exp\,(\frac{\sum_{i=1}^{n}(x_i-\hat{\mu}_1)}{2\sigma_1^2} - \frac{\sum_{i=1}^{m}(y_i-\hat{\mu}_2)}{2\sigma_1^2} - \frac{\sum_{i=1}^{n}(x_i-\bar{X})}{2\sigma_1^2} - \frac{\sum_{i=1}^{m}(y_i-\bar{Y})}{2\sigma_1^2}) \\
&= \exp\frac{(-\frac{n}{2\sigma_1^2}(\frac{m}{\sigma_2^2}\bar{Y} - \frac{m}{\sigma_2^2}\bar{X} + \frac{m}{\sigma_2^2}\delta)^2 - \frac{m}{2\sigma_2^2}(\frac{n}{\sigma_1^2}\bar{X} - \frac{n}{\sigma_1^2}\bar{Y} - \frac{n}{\sigma_1^2}\delta)^2}{\frac{n}{\sigma_1^2}+\frac{m}{\sigma_2^2}}) \\
&= \exp\frac{(-\frac{n}{2\sigma_1^2}(\frac{m}{\sigma_2^2})^2(\bar{X}-\bar{Y}-\delta)^2 - \frac{m}{2\sigma_2^2}(\frac{n}{\sigma_1^2})^2(\bar{X}-\bar{Y}-\delta)^2}{(\frac{n}{\sigma_1^2}+\frac{m}{\sigma_2^2})^2}) \\
&= \exp\,(\frac{-\frac{n}{\sigma_1^2}(\frac{m}{2\sigma_2^2})(\bar{X}-\bar{Y}-\delta)^2(\frac{m}{\sigma_2^2}+\frac{n}{\sigma_1^2})}{(\frac{n}{\sigma_1^2}+\frac{m}{\sigma_2^2})^2} \\
&= \exp(\frac{-1}{2}\frac{(\bar{X}-\bar{Y}-\delta)^2}{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}})
\end{aligned} \tag{3.26}$$

so the likelihood ratio can be written as:

$$\lambda = \exp(\frac{-1}{2}\frac{(\bar{X}-\bar{Y}-\delta)^2}{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}) < c \iff \frac{-1}{2}\frac{(\bar{X}-\bar{Y}-\delta)^2}{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}} < lnc \iff \frac{(\bar{X}-\bar{Y}-\delta)^2}{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}} > -2lnc \tag{3.27}$$

Considering that this is the case for $\bar{X}-\bar{Y}-\delta > 0$, taking square roots yields:

$$\frac{\bar{X}-\bar{Y}-\delta}{\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}} > \sqrt{-2lnc} = c'$$

Under the nul hypothesis the exact distribution of the test statistic is not known, but in any case it will be of the form $\mathcal{N}(-c^*, 1)$ where $c^* \geq 0$ a constant. However, if it is assumed that its distribution is $\mathcal{N}(0, 1)$ then, rejecting the null hypothesis for this distribution implies that the null hypothesis would be rejected for any other distribution of the null hypothesis. This is because as the mean becomes more negative the respective distributions will always contain the rejection area of the $\mathcal{N}(0, 1)$ (because they are "moved" to the left). Thus one can use the points of a standard normal distribution for the comparison with the test statistic.

For equivalence studies the parameter space is in this case:

$$\Theta_0 = \{(\mu_1, \mu_2) : \mu_1 \leq \mu_2 - \delta \text{ , or, } \mu_1 \leq \mu_2 + \delta\}$$

$$\Theta_1 = \{(\mu_1, \mu_2) : \mu_2 - \delta \leq \mu_1 \leq \mu_2 + \delta\}$$

$$\Theta = \Theta_0 \cup \Theta_1$$

One can consider the procedure of Two-One Sided Testing (TOST)[45], which is based on the comment below. The null and alternative hypotheses can be written as:

$$H_0 : \mu_1 - \mu_2 \leq -\delta \text{ OR } \mu_1 - \mu_2 \geq \delta$$
$$H_1 : -\delta < \mu_1 - \mu_2 < \delta$$

by defining the null and alternative hypotheses:

$$H_{01} : \mu_1 - \mu_2 \leq -\delta$$

$$H_{11} : \mu_1 - \mu_2 > -\delta$$

and

$$H_{02} : \mu_1 - \mu_2 \geq \delta$$
$$H_{12} : \mu_1 - \mu_2 < \delta$$

we can write:

$$H_0 : \cup H_{0i}$$
$$H_1 : \cap H_{1i}$$

Thus, we would reject H_0 if both H_{01}, H_{02} were rejected. By building separately an LRT statistic for each H_{0i}, H_{1i} with significance levels at most α and based on the fact that to reject H_0 we must simultaneously reject the other two hypotheses we can construct a test statistic of the form:

$$\{\frac{\bar{X} - \bar{Y} + \delta}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} > z_a\} \cap \{\frac{\bar{X} - \bar{Y} - \delta}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} < -z_a\}$$

Note that even though it seems that we perform a repeating testing procedure, which would inflate the type I error, the type I error actually remains below the prespecified significance level α:

$$P(\{\frac{\bar{X} - \bar{Y} + \delta}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} > z_a\} \cap \{\frac{\bar{X} - \bar{Y} - \delta}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} < -z_a | H_0) \leq P(\frac{\bar{X} - \bar{Y} + \delta}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} > z_a | H_0) \leq a$$

where the second inequality occurs because the interstetion of two event is less or equal to the probability of one of them. As a result no error controlling methods are needed.

For dichotomous outcomes r.v.'s:

$$X \sim \mathcal{B}(p_1, n_1), \text{ and } Y \sim \mathcal{B}(p_2, n_2)$$

where X, Y is the number of successes in in the experimental and placebo or standard therapy group respectively.

Since in phase III trials the sample size is large one can use the normal approximation that is:

$$X \sim \mathcal{N}(n_1 p_1, n_1 p_1(1-p_1)), \text{ and } Y \sim \mathcal{N}(n_2 p_2, n_2 p_2(1-p_2))$$

and dividing by the sample sizes

$$\bar{X} \sim \mathcal{N}(p_1, \tfrac{p_1(1-p_1)}{n_1}), \text{ and } \bar{Y} \sim \mathcal{N}(p_2, \tfrac{p_2(1-p_2)}{n_2})$$

At this point the same arguments for chossing a proper boundary (as in the case of the LRT for normal variales) can be used.

As a result, for an inferioty trial we will have:

$$\frac{\bar{X} - \bar{Y} + \delta}{\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{n}}} > z_a$$

For an equivalency trial we will have:

$$z_a - \delta' \leq \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{n}}} \leq -z_a + \delta'$$

where

$$\delta' = \frac{\delta}{\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{n}}}$$

Notice in this case that the variance of the binomial random variables is not known since it consists of the parameters. The design of such a study is mostly based on what is already known for the standard therapy and what is expected from the experimental therapy. This is specified during the design stage. For example, if it is known that the proportion of response is 30 percent in the placebo group and it is expected that the experimental drug will be better by 10 percent, then for the calculation of variance it will be $p_1 = 0.4, p_2 = 0.3$. In order to estimate the proportions of standard therapy previous studies where the standard treatment was compared with placebo must be used.

3.3 Sample Size Estimation

Based on the test statistics derived in the previous section, the sample size of the clinical trial must be calibrated in such a way that the test will have desirable power. Starting with quantitative outcomes, we present the methodology of sample size estimation for superiority, non-inferiority and equivalence:

For a superiority trial under H_1:

$$\bar{X}_1 - \bar{X}_2 \sim \mathcal{N}(\theta, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n})$$

for a $\theta > 0$. In order to computed the power, this θ needs to be specified. Its specification is based on clinical judgement and on the expectations of the experimental treatment compared to the standard one. After specifying θ it holds for the power that:

$$\begin{aligned}
1-\beta &= P(\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > z_a | \mu_1 = \mu_2 + \theta) = P(\frac{\bar{X}_1 - \bar{X}_2 - \theta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > z_a - \frac{\theta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} | \mu_1 = \mu_2 + \theta) \\
1-\beta &= \Phi(-z_a + \frac{\theta}{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} | \mu_1 = \mu_2 + \theta) = 1 - P(Z > -z_a + \frac{\theta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} | \mu_1 = \mu_2 + \theta) \\
\beta &= P(Z > -z_a + \frac{\theta}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n_1}}} | \mu_1 = \mu_2 + \theta) \\
&- z_a + \frac{\theta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = z_\beta
\end{aligned} \tag{3.28}$$

where the third equality occured because for a normal distribution it holds due to symmetry:

$$P(Z > z) = P(Z < -z) \tag{3.29}$$

Usually, equal variances ($\sigma_1 = \sigma_2 = \sigma$) and sample sizes ($n_1 = n_2 = n$) are assumed. Thus from (3.28) solving for n yields:

$$n = \frac{4\sigma^2(z_\beta + z_a)^2}{\theta^2} \tag{3.30}$$

For a non-inferiority trial, as described in the first section, the margin δ is actually the margin for which the effect of the two treatments (standard ve experimental) is clinically insignificant. This partly implies that under H_1, $\theta = 0$. Under H_1 superiority can also be implied and it is actually checked if non-inferiority is established but at the design stage sample size is estimated mostly for the case of $\theta = 0$. These conditions yield that under H_1:

$$\bar{X}_1 - \bar{X}_2 \sim \mathcal{N}(0, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$$

and the power formula is:

$$\begin{aligned}
1 - \beta &= P(\frac{\bar{X}_1 - \bar{X}_2 + \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > z_a)|\theta = 0) = P(\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > z_a - \frac{\delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}|\theta = 0) \\
1 - \beta &= \Phi(-z_a + \frac{\delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}|\theta = 0) = 1 - P(\frac{\bar{X}_1 - \bar{X}_2 + \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > -z_a + \frac{\delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}) \\
z_\beta &= -z_a + \frac{\delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}
\end{aligned} \tag{3.31}$$

Setting equal variances and sample sizes as in the superiority case will yield:

$$n = \frac{4\sigma^2(z_\beta + z_a)^2}{\delta^2} \tag{3.32}$$

For the case of equivalency, the margins are selected similarly to the non-inferiority trial. Thus, under the alternative hypothesis it is common to consider that $\theta = 0$.For convenience we set

$$\delta' = \frac{\delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

The power formula will then be:

$$\begin{aligned}
1-\beta &= P(z_a - \delta' \leq \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \leq -z_a + \delta') = \phi(-z_a+\delta') - \phi(z_a - \delta') \\
&= \Phi(-z_a+\delta) - 1 + P(\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} > z_a - \delta) = \\
&= \Phi(-z_a+\delta) - 1 + P(\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} < -z_a + \delta) \\
&= 2\Phi(-z_a+\delta) - 1 \\
&= 2(1 - P(\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} > -z_a + \delta)) - 1
\end{aligned} \tag{3.33}$$

Continuing on both sides:

$$\begin{aligned}
2-\beta &= 2(1 - P(\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} > -z_a + \delta)) \\
\frac{\beta}{2} &= P(\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} > -z_a + \delta)) \\
z_{\frac{\beta}{2}} &= -z_a + \delta' = -z_a + \frac{\delta}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}
\end{aligned} \tag{3.34}$$

Assuming equal variances and sample sizes the formula for the sample size will be:

$$n = \frac{4\sigma^2(z_{\frac{\beta}{2}} + z_a)^2}{\delta^2} \tag{3.35}$$

For survival outcomes using Schoenfeld's asymptotic formula which was presented in the Section of the Log rank test we have that we will reject the null hypothesis under the local alternative when:

$$\begin{aligned}
1-\beta &= P(|L| > z_{\frac{a}{2}} | H_{1n}) = P(L > z_{\frac{a}{2}} \text{ or } L < -z_{\frac{a}{2}}) \\
&= P(L - E > z_{\frac{a}{2}} - E \text{ or } L - E < -z_{\frac{a}{2}} - E || H_{1n}) \\
&\approx P(L - E > z_{\frac{a}{2}} - E | H_{1n}) = P(Z > z_{\frac{a}{2}} - E | H_{1n})
\end{aligned} \tag{3.36}$$

where the last equality occurs because intuitively under H_{1n} the hazard rate of the control group is larger then the hazard rate of the group that takes the experimental therapy. Thus the event of the logrank statistic (which is computed based on the deaths in the control group) being smaller than its expected value becomes unlikely and can be omitted.

Using the same methodology as before from (3.36) we get:

$$E = z_{\frac{a}{2}} + z_{\beta}$$

and expanding $E = b\sqrt{\omega_1\omega_2 P} = -log(\delta)\sqrt{n}\sqrt{\omega_1\omega_2 P})$

Notice that since P is the probability of someone dying in either group 1 or group 2, we can write: $E = -log(\delta)\sqrt{\omega_1\omega_2 d}$ since $d = nP$. Thus solving for d we get:

$$d = \frac{(z_{\frac{a}{2}} + z_{\beta})^2}{[log(\delta)]^2\omega_1\omega_2} \tag{3.37}$$

Note that that even though formula (3.37) is more convenient to calculate, it usually isn't enough for the proper design of a survival clinical trial. One reason is that eventually we do not know how much time is needed to observe this number of events and we also do not know how many patients need to be accrued in order to observe them. Thus the total sample size and the duration of the trial must eventually be determined.

In order to acquire the sample size for a survival clinical trial the probabilities of observing an event in the control or the experimental group need to be computed. Denoting as $p_i, i = 1, 2$ the probability of observing an event in group i ($i = 1$ is the control group)[43]:

$$\begin{aligned} p_i = P_i(T < C) &= \int_0^{+\infty}\int_t^{+\infty} f(c,t)dcdt \\ &= \int_0^{+\infty} f_i(t)\int_t^{+\infty} f(c)dcdt \\ &= \int_0^{+\infty} f_i(t)P(C > t)dt \\ &= \int_0^{+\infty} h_i(t)S_i(t)P(C > t)dt = \int_0^{+\infty} h_i(t)S_i(t)G(t)dt \end{aligned} \tag{3.38}$$

where $G(t) = P(C > t)$. Note that there is not an index for this distribution function, because we assume that the censoring distribution is common for both groups. If the censoring distributions were different this would imply that censoring is dependent on which group someone belongs and informative censoring would occur.

To calculate this probability, we have to specify the censoring distribution $G(t)$, which is usually difficult during the design stage. However, it is typically assumed that subjects are accrued uniformly over an accrual period of length t_a and followed for a period of length t_f , and that no subject is lost to follow-up during the study. This way the distribution of censoring depends only on the time of entry. For example, under the previous conditions of no loss to follow up, for a trial with accrual $t_a = 2$ years and follow up $t_f = 2$ the probability of someone having censoring time larger than 2.5 years implies that he shoudl have entered the first 1.5 years of the study. Since the accrual times follow a uniform distribution it is straightforward that $G(t) = A(t_a+t_f-t)$,where $A(t)$ is the uniform cdf of the accrual times. Expanding $G(t)$ into branches we get:

$$G(t) = \begin{cases} 0 & , t > t_a + t_f \\ \frac{t_a+t_f-t}{t_a} & , t_f < t \leq t_a + t_f \\ 1 & , 0 < t \leq t_f \end{cases}$$

Thus the computations become:

$$\begin{aligned} p_i &= \int_0^{+\infty} h_i(t)S_i(t)G(t)dt = \int_0^{t_f} h_i(t)S_i(t)dt + \int_{t_f}^{t_a+t_f} h_i(t)S_i(t)\frac{t_a+t_f-t}{t_a}dt \\ &= \int_0^{t_f} dF_i(t) + \frac{t_a+t_f}{t_a}\int_{t_f}^{t_f+t_a} dF_i(t) - \frac{1}{t_a}\int_{t_f}^{t_f+t_a} tdF_i(t) \\ &= F_i(t_f) - 0 + \frac{t_a+t_f}{t_a}(F_i(t_f+t_a) - F_i(t_f)) - \\ &- \frac{1}{t_a}((t_f+t_a)F_i(t_f+t_a) - t_fF_i(t_f)) + \frac{1}{t_a}\int_{t_f}^{t_f+t_a} F_i(t)dt \\ &= \frac{1}{t_a}\int_{t_f}^{t_f+t_a} F_i(t)dt \\ &= 1 - \frac{1}{t_a}\int_{t_f}^{t_f+t_a} S_i(t)dt \end{aligned} \tag{3.39}$$

Using Simpson's approximation and after some algebraic calculations we get:

$$p_i \approx 1 - \frac{1}{6}(S_i(t_f) + 4S_i(0.5t_a + t_f) + S_i(t_a + t_f))$$

Using historical data from previous relevant trials the p_i's can be estimated and used for the calculation of the total probability of observing a death:

$$P = \omega_1 p_1 + \omega_2 p_2$$

Thus the required sample size can be estimated as $n = d/P$, where d is the required number of observed events derived from the formula (3.37).

3.4 Group Sequential Designs

As noted the duration of a phase III clinical trial is multiple years. Instead, of performing a final test when the trial ends one could actually perform interim analyses (that is, analyses during the trial) to arrive to a conclusion sooner. This framework is similar to the one discussed in Chapter 2 with the 2-stage trials but in a phase III trial more interim analysis stages are needed. Group sequential methods are rules for stopping a trial early based on treatment differences that are observed during interim analyses.

Group sequential methods pose numerous advantages compared to designs that utilize only one test at the end of the trial. One of them is related to ethical considerations. For example, if a drug is effective then this result may be observed earlier and patients in the placebo group will be able to receive the better treatment sooner. Group sequential methods also offer logistical advantages since stopping a trial earlier reduces the resources spent.

In order to perform group sequential methods, a test statistic at each monitoring time needs to be computed, thus a sequence of test statistics arise. Turnbull (2000) [46] proposed to use sequences of standardized test statistics that have the identities below:

- $(Z_1, ..., Z_k)$ follow a multivatiare normal
- $E(Z_k) = \theta\sqrt{\mathcal{I}_k}$
- $Cov(Z_{k_1}, Z_{k_2}) = \sqrt{\frac{\mathcal{I}_{k_1}}{\mathcal{I}_{k_2}}}$

Where $\mathcal{I}_{k_2}, ..., \mathcal{I}_{k_n}$ are the observed information levels and θ is a parameter of interest. Notice that by standardized we mean test statistics that have variance 1. Thus, $Z_k \sim N(\theta\sqrt{\mathcal{I}_k}, 1)$. If these conditions are satisfied than these statistics will be called the canonical joint distribution with information levels $\mathcal{I}_{k_1}, ..., \mathcal{I}_{k_n}$ for the parameter θ. An example of such statistic is given below. Suppose that in a two armed trial the participants in group A have quantitative responses $X_{Ai} \sim N(\mu_A, \sigma_A^2)$ and the participants in group B have quantitative responses $X_{Bi} \sim N(\mu_B, \sigma_B^2)$. The hypotheses to be tested are:

$$H_0 : \mu_A = \mu_B$$

$$H_1 : \mu_A \neq \mu_B$$

At time t_k an interim analyses is performed. Consider the statistic

$$\bar{X}_A^{(k)} - \bar{X}_B^{(k)} \sim N(\mu_A - \mu_B, \frac{\sigma_A^2}{n_{Ak}} + \frac{\sigma_B^2}{n_{Bk}})$$

The information level at stage k is then

$$\mathcal{I}_k = (\frac{\sigma_A^2}{n_{Ak}} + \frac{\sigma_B^2}{n_{Bk}})^{-1}$$

Setting $\theta = \mu_A - \mu_B$ it occurs that:

$$Z_k = (\bar{X}_A^{(k)} - \bar{X}_B^{(k)})\sqrt{\mathcal{I}_k} \sim N(\theta\sqrt{\mathcal{I}_k}, 1)$$

So the families $(Z_1, ...Z_n)$ satisfy the the conditions (1) and (2). For (3) note that:

$$Cov(Z_{k_1}, Z_{k_2}) = \sqrt{\mathcal{I}_{k_1}}\sqrt{\mathcal{I}_{k_2}} Cov(\bar{X}_A^{(k_1)} - \bar{X}_B^{(k_1)}, \bar{X}_A^{(k_2)} - \bar{X}_B^{(k_2)})$$

Since the $\bar{X}_A^{(k_i)}, \bar{X}_A^{(k_j)}$ for $i, j = 1, 2$ are independent the above relation will simplify into:

$$\begin{aligned} Cov(Z_{k_1}, Z_{k_2}) &= \sqrt{\mathcal{I}_{k_1}}\sqrt{\mathcal{I}_{k_2}}(Cov(\bar{X}_A^{(k_1)}, \bar{X}_A^{(k_2)}) + Cov(\bar{X}_B^{(k_1)}, \bar{X}_B^{(k_2)})) \\ &= \sqrt{\mathcal{I}_{k_1}}\sqrt{\mathcal{I}_{k_2}}(Cov(\frac{1}{n_{A_{k_1}}}\sum_{i=1}^{n_{A_{k_1}}} X_{Ai}, \frac{1}{n_{A_{k_2}}}\sum_{i=1}^{n_{A_{k_2}}} X_{Ai}) + Cov(\frac{1}{n_{B_{k_1}}}\sum_{i=1}^{n_{B_{k_1}}} X_{Bi}, \frac{1}{n_{B_{k_2}}}\sum_{i=1}^{n_{B_{k_2}}} X_{Bi}) \end{aligned} \tag{3.40}$$

Notice for the group A (the same follows for group B) that:

$$Cov(\frac{1}{n_{A_{k_1}}}\sum_{i=1}^{n_{A_{k_1}}} X_{Ai}, \frac{1}{n_{A_{k_2}}}(\sum_{i=1}^{n_{A_{k_1}}} X_{Ai} + \sum_{i=n_{A_{k_1}}+1}^{n_{A_{k_2}}} X_{Ai}) = \frac{1}{n_{A_{k_1}}}\frac{1}{n_{A_{k_2}}} Cov(\sum_{i=1}^{n_{A_{k_1}}} X_{Ai}, \sum_{i=1}^{n_{A_{k_1}}} X_{Ai} + \sum_{i=n_{A_{k_1}}}^{n_{A_{k_2}}} X_{Ai})$$

$$= \frac{1}{n_{A_{k_1}}}\frac{1}{n_{A_{k_2}}} Cov(\sum_{i=1}^{n_{A_{k_1}}} X_{Ai}, \sum_{i=1}^{n_{A_{k_1}}} X_{Ai}) = \frac{1}{n_{A_{k_1}}}\frac{1}{n_{A_{k_2}}} n_{A_{k_1}}\sigma_A^2 \tag{3.41}$$

Using the above relation (3.41):

$$Cov(Z_{k_1}, Z_{k_2}) = (\frac{1}{n_{A_{k_2}}}\sigma_A^2 + \frac{1}{n_{B_{k_2}}}\sigma_B^2)\sqrt{\mathcal{I}_{k_1}}\sqrt{\mathcal{I}_{k_2}} = (\mathcal{I}_{k_2}^{-1}\sqrt{\mathcal{I}_{k_1}}\sqrt{\mathcal{I}_{k_2}}) = \sqrt{\frac{\mathcal{I}_{k_1}}{\mathcal{I}_{k_2}}} \tag{3.42}$$

Thus the standardized statistics $(Z_1, ..., Z_k)$ satisfy all the conditions and they follow a canonical joint distribution with information levels $\mathcal{I}_{k_2}, ..., \mathcal{I}_{k_n}$ for the parameter $\mu_A - \mu_B$.

It can be shown with similar methodology that in the case of dichotomous outcomes using the sequence of statistics

$$Z_k = (\hat{p}_A^{(k_1)} - \hat{p}_B^{(k_2)})\hat{\mathcal{I}}_k = (\frac{1}{n_{Ak_1}}\sum_{i=1}^{n_{Ak_1}} Y_{Ai} - \frac{1}{n_{Bk_2}}\sum_{i=1}^{n_{Bk_2}} Y_{Bi})\hat{\mathcal{I}}_k$$

where the Y_{Ai}, Y_{Bi} are Bernoulli random variables and $\hat{\mathcal{I}}_k$ is the estimate of the Fisher information under the null.

For the case of the log-rank test:

$$L_R^{(k)} = \frac{\sum_{i=1}^k (d_{i1} - E(d_i|n_{i1}, n_{i2}, d_i))}{\sqrt{\sum_{i=1}^k Var(d_{i1}|n_{i1}, n_{i2}, d_i)}} = \frac{S^{(k)}}{\sqrt{\sum_{i=1}^k Var(d_{i1}|n_{i1}, n_{i2}, d_i)}}$$

under an alternative hypothesis and no ties where $\theta = -log\frac{h_2(t)}{h_1(t)}$ is close to zero one can use the approximation:

$$S^{(k)} \sim \mathcal{N}(\theta\mathcal{I}_k, \mathcal{I}_k)$$

where $\mathcal{I}_k = \sum_{i=1}^k V(d_{i1}|n_{i1}, n_{i2}, d_i)$ Thus, for $Z_k = L_R^{(k)} = \frac{S^{(k)}}{\sqrt{\mathcal{I}_k}}$ a group sequential test can be developed in the usual way.

3.4.1 Assumptions and Power in Group Sequential Designs

Consider a sequence of test statistics that belong to the canonical joint distribution. A useful assumption that simplifies the computations is that the interim analyses are performed at times that equal groups sizes have accrued. This means that the if $n_1, ..., n_K$ the total number of patients at each interim analysis then $n_j = jn_1$. If this assumption is made then it will also hold that $\mathcal{I}_j = j\mathcal{I}_1$. Thus condition (3) will become:

$$Cov(Z_{k_1}, Z_{k_2}) = \sqrt{\frac{\mathcal{I}_{k_1}}{\mathcal{I}_{k_2}}} = \sqrt{\frac{k_1}{k_2}}$$

Suppose that for a hypothesis of the form

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

a test statistic satisfying $Z_k \sim N(\theta\sqrt{I_k}, 1)$ will be used. For the alternative a proper sample size for power 1-β at $\theta = \delta$ is desired, so:

$$\begin{aligned} 1 - \beta &= P(|Z| > z_{a/2} | \theta = \delta) \\ &= P(Z > z_{a/2} \text{ or } Z < -z_{a/2} | \theta = \delta) \\ &\approx P(Z > z_{a/2} | \theta = \delta) \end{aligned} \tag{3.43}$$

The last approximation holds because if $\theta = \delta$ the event $\{Z < -z_{a/2}\}$ can be considered close to zero. This will eventually yield the relation

$$\mathcal{I}_0 = \frac{(z_{a/2} + z_\beta)^2}{\delta^2} \tag{3.44}$$

Now, in a group sequential trial the information level at the final stage K will be larger than the information of a trial that utilizes only one test. This means that:

$$\mathcal{I}_K = R\mathcal{I}_0, \text{ where } R > 1$$

And assuming equal group sizes, the information levels will be equidistant, thus the information level at a monitoring time k can be written as:

$$\mathcal{I}_k = \frac{k}{K}\mathcal{I}_K = \frac{k}{K}R\mathcal{I}_0 = \frac{k}{K}R\frac{(z_{a/2} + z_\beta)^2}{\delta^2}$$

One can actually show that the parameter R will not depend on the value of the alternative hypothesis or other parameters such as the variance.It will actually depend only on $\alpha, 1-\beta, K$. One can see why this is the case by considering quantitative outcomes with equal sample sizes and observing that the sample size is proportional to the ratio of δ^2, σ^2. Thus the sample size changes in the same way this ratio changes and it occurs that R will be independent of these parameters. This can then be generalized to other endpoints. This section presents these results because they are important for the next section where the probabilities of type I and type II error rates will be presented. As far as now, the information levels at each monitoring time are considered observed. But in the process of designing a group sequential trial the information levels are not known. Making some additional assumptions, such as the equal group sizes assumptions, the specification of the information levels depends only on the specification of the final information level,the number of interim analyses and the desired power and type I error, simplifying the computations. Of course, the assumption of equal group sizes might not hold. In this case more complex methods are used, but even in this case assumptions about the information levels are still made. This dissertation is based on the assumption of equal group sizes, which is also the most common assumption made.

3.4.2 Error Rates Probability Computation

Suppose we want to test the hypotheses of the form:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

with power for the alternative hypothesis at $\delta > 0$, using K interim analyses and disjoint continuation regions (a_k, b_k) for the statistic of paragraph (3.4.1). We will use this type of continuation and rejection regions because as Turnbull has commented they are the most commonly used ones. Thus we note that in any interim analysis $k \leq K$ the tests will be of the form:

- if $Z_k > b_k$ or $Z_k < a_k$ $(a_k < b_k)$ reject the null hypothesis and stop the trial
- otherwise proceed to the next stage (or stop the trial if at the last stage)

Of course the continuation regions might be different but this does not alter significantly the following computational process. Additionally the computations that will follow are not different for other statistics such as the log-rank or the risk difference.

For the boundaries that we have selected, the type I error rate r_k for a specific interim analysis k will be

$$P(a_1 < Z_1 < b_1, a_2 < Z_2 < b_2, ..., Z_k > b_k|H_0)+P(a_1 < Z_1 < b_1, a_2 < Z_2 < b_2, ..., Z_k < a_k|H_0)$$

Thus the total error will be:

$$\sum_{k=1}^{K} r_k$$

equivalently for the power of the test approximately for large δ:

$$\sum_{k=1}^{K} P(a_1 < Z_1 < b_1, a_2 < Z_2 < b_2, ..., Z_k > a_k|\mu_A = \mu_B + \delta)$$

Now the need for the exact joint distribution of $(Z_1, ...Z_k)$. Notice that the statistics that follow the common joint distribution are actually a markov sequence of statistics, so:

$$f(z_k|z_{k-1}, ..., z_1) = f(z_k|z_{k-1})$$

To see why this is a markov sequence notice that:

$$Z_k = (\sqrt{\mathcal{I}_k}\frac{n_{A(k-1)}}{n_{Ak}}\bar{X}_A^{(k-1)} - \sqrt{\mathcal{I}_k}\frac{n_{B(k-1)}}{n_{B(k)}}\bar{X}_B^{(k-1)}) + \frac{1}{n_{Ak}}\sqrt{\mathcal{I}_k}\sum_{i=1}^{n_{A1}} X_{Ai} + \frac{1}{n_{Bk}}\sqrt{\mathcal{I}_k}\sum_{i=1}^{n_{B1}} X_{Bi}$$

and by the assumption of equal increments this can be written as:

$$Z_k = \sqrt{\frac{k-1}{k}}Z_{k-1} + \frac{1}{n_{Ak}}\sqrt{\mathcal{I}_k}\sum_{i=1}^{n_{A1}} X_{Ai} + \frac{1}{n_{Bk}}\sqrt{\mathcal{I}_k}\sum_{i=1}^{n_{B1}} X_{Bi}$$

Writing Z_k at this form it is obvious that if Z_{k-1} is observed then observing $Z_{k-2}, ...Z_1$ does not offer any other information for the value of Z_k, so the markov property holds. Additionally writing,

$$Z_k\sqrt{\mathcal{I}}_k - Z_{k-1}\sqrt{\mathcal{I}}_{k-1} = \sqrt{\mathcal{I}}_k(\sqrt{\frac{k-1}{k}}Z_{k-1} + \frac{1}{n_{Ak}}\sqrt{\mathcal{I}_k}\sum_{i=1}^{n_{A1}} X_{Ai} + \frac{1}{n_{Bk}}\sqrt{\mathcal{I}}_k\sum_{i=1}^{n_{B1}} X_{Bi}) - \sqrt{\mathcal{I}_{k-1}}Z_{k-1}$$

it is straightforward that $Z_k\sqrt{\mathcal{I}}_k - Z_{k-1}\sqrt{\mathcal{I}}_{k-1}$ can be written as a linear combination of independent normal variables. Utilizing this fact and also defining $\delta_k = \mathcal{I}_k - \mathcal{I}_{k-1}$ we will then have:

$$Z_k\sqrt{\mathcal{I}}_k - Z_{k-1}\sqrt{\mathcal{I}}_{k-1} \sim N(\theta\delta_k, \delta_k)$$

The variance is δ_k because:

$$Var(Z_k\sqrt{\mathcal{I}_k} - Z_{k-1}\sqrt{\mathcal{I}_{k-1}}) = \mathcal{I}_k Var(Z_k) + \mathcal{I}_{k-1}Var(Z_{k-1}) - 2\sqrt{\mathcal{I}_k}\sqrt{\mathcal{I}_{k-1}}Cov(Z_k, Z_{k-1})$$

This is convenient because:

$$f(z_k, z_{k-1}, ..., z_1) = f(z_k|z_{k-1})f(z_{k-1}, ..., z_1) = f(z_1)f(z_2|z_1).....f(z_k|z_{k-1})$$

Thus only

$$f(z_k|z_{k-1}) = \frac{f(z_k, z_{k-1})}{f(z_{k-1})}$$

needs to be computed. For the numerator the correlation coefficient for the bivariate distribution of (z_k, z_{k-1}) is:

$$\rho = \frac{Cov(Z_k, Z_{k-1})}{1} = \sqrt{\frac{\mathcal{I}_{k-1}}{\mathcal{I}_k}}$$

thus

$$2(1-\rho^2) = 2\frac{\delta_k}{\mathcal{I}_k}$$

The exponential part of the bivariate normal distribution for (z_k, z_{k-1}) is:

$$\begin{aligned}
&-\frac{\mathcal{I}_k}{2\delta_k}(z_k - \theta\sqrt{\mathcal{I}_k})^2 + \frac{\mathcal{I}_k}{2\delta_k}\sqrt{\frac{\mathcal{I}_{k-1}}{\mathcal{I}_k}}(z_k - \theta\sqrt{\mathcal{I}_k})\times \\
&\times (z_{k-1} - \theta\sqrt{\mathcal{I}_{k-1}}) - \frac{\mathcal{I}_k}{2\delta_k}(z_{k-1} - \theta\sqrt{\mathcal{I}_{k-1}})^2 + \frac{1}{2}(z_{k-1} - \theta\sqrt{\mathcal{I}_{k-1}})^2 \\
&= -\frac{\mathcal{I}_k}{2\delta_k}(z_k - \theta\sqrt{\mathcal{I}_k})^2 + \frac{\mathcal{I}_k}{2\delta_k}\sqrt{\frac{\mathcal{I}_{k-1}}{\mathcal{I}_k}}(z_k - \theta\sqrt{\mathcal{I}_k}) + (z_{k-1} - \theta\sqrt{\mathcal{I}_{k-1}})^2(\frac{1}{2} - \frac{\mathcal{I}_k}{2\delta_k}) \\
&= -\frac{\mathcal{I}_k}{2\delta_k}(z_k - \theta\sqrt{\mathcal{I}_k})^2 + \frac{\mathcal{I}_k}{2\delta_k}\sqrt{\frac{\mathcal{I}_{k-1}}{\mathcal{I}_k}}(z_k - \theta\sqrt{\mathcal{I}_k}) - (z_{k-1} - \theta\sqrt{\mathcal{I}_{k-1}})^2\frac{\mathcal{I}_{k-1}}{2\delta_k}
\end{aligned} \tag{3.45}$$

This is the usual identinty yielding:

$$(\sqrt{\frac{\mathcal{I}_k}{2\delta_k}}(z_k - \theta\sqrt{\mathcal{I}_k}) - \sqrt{\frac{\mathcal{I}_{k-1}}{2\delta_k}}(z_{k-1} - \theta\sqrt{\mathcal{I}_{k-1}}))^2 = (\frac{\sqrt{\mathcal{I}_k}z_k - \theta\mathcal{I}_k - \sqrt{\mathcal{I}_{k-1}}z_{k-1} + \theta\mathcal{I}_{k-1}}{\sqrt{2\delta_k}})^2 \tag{3.46}$$

For the non-exponential part:

$$\sqrt{\frac{\mathcal{I}_k}{\delta_k}}\frac{\sqrt{(2\pi)}}{2\pi}$$

So combining these results will yield

$$f(z_k|z_{k-1}) = \sqrt{\frac{\mathcal{I}_k}{\delta_k}}\phi(\frac{\sqrt{\mathcal{I}_k}z_k - \theta\delta_k - \sqrt{\mathcal{I}_{k-1}}z_{k-1}}{\sqrt{\delta_k}})$$

where ϕ denotes the density function for a standard normal distribution. With the conditional distributions found it remains to calculate:

$$\begin{aligned}
&P(a_1 < Z_1 < b_1, a_2 < Z_2 < b_2, ..., Z_k > a_k|H_1) = \\
&= \int_{a_1}^{b_1} .... \int_{b_k}^{\infty} f(z_1, ..., z_k)dz_k...dz_1 \\
&= \int_{a_1}^{b_1} .... \int_{b_k}^{\infty} f(z_k|z_{k-1})....f(z_1)dz_k...dz_1 \\
&= \int_{a_1}^{b_1} .... \int_{a_{k-1}}^{b_{k-1}} f(z_{k-1}|z_{k-2})....f(z_1)[\int_{b_k}^{\infty} f(z_k|z_{k-1})dz_k]dz_{k-1}...dz_1 \\
&= \int_{a_1}^{b_1} .... \int_{a_{k-1}}^{b_{k-1}} f(z_{k-1}|z_{k-2})....f(z_1)[P(Z_k > b_k)]dz_{k-1}...dz_1
\end{aligned} \tag{3.47}$$

For the probability inside the integral it can be noticed that

$$P(Z_k > b_k) = P(\frac{Z_k\sqrt{\mathcal{I}_k} - z_{k-1}\sqrt{\mathcal{I}_{k-1}} - \theta\delta_k}{\sqrt{\delta_k}} > \frac{b_k\sqrt{\mathcal{I}_k} - z_{k-1}\sqrt{\mathcal{I}_{k-1}} - \theta\delta_k}{\sqrt{\delta_k}}) \tag{3.48}$$

which is actually a standard normal variable thus the final integral can be written as:

$$\int_{a_1}^{b_1} .... \int_{a_{k-1}}^{b_{k-1}} f(z_{k-1}|z_{k-2}) .... f(z_1)(1 - \Phi(\frac{b_k\sqrt{\mathcal{I}}_k - z_{k-1}\sqrt{\mathcal{I}_{k-1}} - \theta\delta_k}{\sqrt{\delta_k}}))dz_{k-1}...dz_1 \tag{3.49}$$

The intergral in (3.49) is evaluated through the use of numerical methods. For a single integral:

$$\int_a^b q(z)dz \approx \sum_{i=1}^{m} w(i)q(z(i))$$

Where the number of points at which q is evaluated and their locations $z(i)$ and weights $w(i)$ are chosen to ensure a sufficiently accurate approximation. Similarly for the multiple integral of (3.49)we will have

$$\sum_{i_1=1}^{m_1} ... \sum_{i_{m-1}=1}^{m_{K-1}} w_1(i)f(z_1(i))w_2(i_2)f(z_2(i)).....$$

The process and details of selecting proper grid points for the weights and the locations is omitted. For information on this part one can see [46]. Notice that all these computations are for the case of $\theta \neq 0$. In order to compute the type I error first notice that under H_0:

$$Z_k \sim N(0, 1)$$

Thus there is no need to define the previous distribution of the difference of two statistics.Under the assumption of equal group sizes, we will have that $\sqrt{\frac{\mathcal{I}_{k-1}}{\mathcal{I}_k}} = \sqrt{\frac{k-1}{k}}$. Using the markov identity and the previous framework one can compute the multiple integral with similar methodology.

3.4.3 Choice of Boundaries

In the previous section the theoretical framework of how to compute the boundary values for specific type I type II errors and number of stages was presented. Based, on these frameworks there are standard boundary types that are used in group sequential trials. The type of boundaries used also differ based on whether the two-sided or one sided test are to be used. For a two sided test:

$$H_0 : \theta = 0$$

$$H_1 : \theta \neq 0$$

some boundary types are presented below:

- Pocock boundaries: Supposing that we have intervals (a_k, b_k) defined as before the Pocock boundaries are defined as $b_k = c = c(K, \alpha), a_k = -b_K$. Thus, Pocock boundaries stay the same for any stage. We write $c = c_{K,\alpha}$ because the selection of the boundaries are dependent on the number of stages and the type I error, which means that they calculated in order to control the type I error below the prespecified level for a prespecified number of stages.

- O' Brien and Fleming boundaries: In this case we define $b_k = c(k) = c(K, \alpha)\sqrt{\frac{k}{K}}, a_k = -b_k, k = 1, , , K$. Note that in this case the boundaries will change and depend on the stage of the group sequential trial. The boundaries are computed in the same way as before.

- Wang and Tsiatis boundaries: In this case $b_k = c = c(K, \alpha, \delta) = c(K, \alpha, \delta)(\sqrt{\frac{K}{k}})^{\delta - \frac{1}{2}}$ where δ takes values between $[0, 0.5]$ .This is a 'mix' of the other two boundaries since for $\delta = 0.5$ and $\delta = 0$ the Pocock-type boundaries and the O'Brien Fleming boundaries are acquired respectively.

For the case of one sided tests:

$$H_0 : \theta = 0$$
$$H_1 : \theta > 0$$

The boundaries for stages $1, ..., K - 1$ are of the type:

$$a_k = c_1(K, \alpha, \beta, \delta)(\frac{k}{K})^{\delta - \frac{1}{2}}$$
$$b_k = \delta\sqrt{\mathcal{I}_k} - c_2(K, \alpha, \beta, \delta)(\sqrt{\frac{k}{K}})^{\delta - \frac{1}{2}}$$

For the Kth stage tge two boundaries must be equal, since this way acceptance or rejection of the null hypothesis is ensured. As a result equating $a_K = b_K$ will yield:

$$\mathcal{I}_K = \frac{[c_1(K, \alpha, \beta, \delta) + c_2(K, \alpha, \beta, \delta)]^2}{\delta^2}$$

As Turnbull and Jenisson have noted, the upper boundary can be viewed as a repeated significance test of $\theta = 0$ with critical values for the Z_k proportional to

$k - 0.5, \mathrm{k} = 1, \ldots, \mathrm{K}$ while the lower boundary arises as a repeated significance test of $\theta = \delta$, this hypothesis being rejected if $Z_k - \delta\sqrt{\mathcal{I}_k}$, the standardized statistic with mean zero if $\theta = \delta$, is negative and less than a critical value proportional to $k - 0.5$.

Thus using the above formula the inflation factor and following some algebraic calculations we will have:

$$R = \frac{[c_1(K, \alpha, \beta, \delta) + c_2(K, \alpha, \beta, \delta)]^2}{(z_\alpha + z_\beta)^2}$$

From the above it seems that a general methodology for a two sided test would be as follows:

- First decide the type I error, type II error and number of stages K.
- Choose the boundary type, and decide the test statistic and the expression for the information levels.
- Find the values of the boundary types that control the type I error.
- After selecting the proper boundaries for specified (K, a) define the maximum information level (see section 2) by changing the inflation factor properly and assuming equal information spacings. Perform power Computations for the selected boundary and the information levels and select the inflation factor for which the power is increased to the specified level.

Through this framework a significant advantage occurs. The computation of the boundary values and the sample size are two different processes, with the selection of the sample size following the computation of the boundary values. This is not true for the case of one sided tests however. Sprecifically, in order to compute the constants $c_1(K, \alpha, \beta, \delta), c_2(K, \alpha, \beta, \delta)]$ one needs to perform a two dimensional seacrch. That is, for different values of $c_1(K, \alpha, \beta, \delta), c_2(K, \alpha, \beta, \delta)]$ replace them in the boundary formulas for one sided tests and calculate the power and type I error. Evaluate whether the error constraints are satisfied (note here that another type of error constraint might be used. For example Turnbull advises to choose the constants for which the squared sum of differences between the type I error and the target value and between the power and the target power is close to zero). R is then selected based on the specified constants and the power constraints. Of course, for a one sided test one can simply set $b_i = -\infty$ for each stage and proceed with specifying the upper boundary values.

Chapter 4

Simulations and Data Analysis

4.1 Designing Simon's Two Stage Designs

In this section we present results for Simon's two-stage designs for different null and alternative hypotheses. To acquire these results the RStudio enviroment and the R function "ph2simon" from the package "clinfun" were used. For prespecified power $\beta = 0.9$ and type I error $\alpha = 0.05$ The results are presented in matrices 1 and 2. Using the same notation as in chapter 2, the first matrix presents the quadruple (r_1, r, n_1, n) when the optimal criterion is used, whilst the second one presents the quadruple (r_1, r, n_1, n) when the minimax criterion is used.

Simon's Two Stage Design (Optimal Design)						
$H_0 : p = p_0$	$H_1 : p = p_1$	r_1	n_1	r	n	EN
0.05	0.25	0	9	2	17	11.96
0.10	0.30	1	10	5	29	15
0.20	0.40	3	13	12	43	20.58
0.30	0.50	5	15	18	46	23.63
0.40	0.60	7	16	23	46	24.52
0.60	0.80	7	11	30	43	20.48
0.70	0.90	4	6	22	27	14.82

Table 4.1: Results for Simon's two stage design for $\beta = 0.9$ and $\alpha = 0.05$, using the optimal criterion

Simon's Two Stage Design (Minimax)						
$H_0 : p = p_0$	$H_0 : p = p_0$	r_1	n_1	r	n	EN
0.05	0.25	0	12	2	16	13.84
0.10	0.30	1	15	5	25	19.51
0.20	0.40	4	18	10	33	22.25
0.30	0.50	6	19	16	39	25.69
0.40	0.60	17	34	20	39	34.44
0.60	0.80	8	13	25	35	20.77
0.70	0.90	19	23	21	26	23.16

Table 4.2: Results for Simon's two stage design for $\beta = 0.9$ and $\alpha = 0.05$, using the minimax criterion

From the two tables one can also compare the two criteria. For example in the case of $H_0 : p = 0.05$ the minimax and optimal designs seem not to differ a lot in terms of expected sample sizes and total sample size. For $H_0 : p = 0.05$ the total sample size is 26 and the expected number of patients is approximately 12 while using the minimax design the total sample size is 28 and and the expected sampe size approximately 14. Choosing between these two criteria is easy in this case since the optimal design has both smaller total sample size and expected number of patients.This can be also seen to be the case for $H_0 : p = 0.10$

For the case of $H_0 : p = 0.20$ this is not the case however. The design for $H_0 : p = 0.10$ has a total sample size of 56 using the optimal criterion versus a total sample size of 51 using the minimax criterion. There is also a difference in expected number of patients specifically 21 (optimal) vs 22 (minimax). It seems that in this case the minimax design should be preffered. This is because we expect one more patient to enroll compared to the optimal design but the total sample size is 5 patients less. Thus it seems better to select the minimax design.

In the case of $H_0 : p = 0.30$ it seems that, in contrast to the previous designs, there is not a specific justification that favors one criterion compared to the other. Indeed, the minimax design has 3 less patients while the optimal design has 2 less expected patients. In such it is not clear how to proceed and which criterion should be used. This is also the case for $H_0 : p = 0.40$.

For the hypotheses $H_0 : p = 0.60$ and $H_0 : p = 0.70$ it is clear that the minimax and optimal designs should be preferred respectively. However, these designs are of

computational interest only, since it is unreasonable for the null response rates to be that large.

We also used the ph2simon function to design trials with prespecified type I error 0.05 and power 0.9. The results for the optimal criterion are presented in table 3 and the results for the minimax criterion in table 4.

Simon's Two Stage Design (Minimax)						
$H_0 : p = p_0$	$H_0 : p = p_0$	r_1	r	n_1	n	EN
0.05	0.25	0	9	3	30	16.76
0.10	0.30	2	18	6	35	22.53
0.20	0.40	4	19	15	54	30.43
0.30	0.50	8	24	24	63	34.72
0.40	0.60	11	25	32	66	35.98
0.60	0.80	12	19	37	53	29.47
0.70	0.90	11	15	29	36	21.23

Table 4.3: Results for Simon's two stage design for $\beta = 0.9$ and $\alpha = 0.05$, using the optimal criterion

Simon's Two Stage Design (Minimax)						
$H_0 : p = p_0$	$H_0 : p = p_0$	r_1	r	n_1	n	EN
0.05	0.25	0	15	3	25	20.37
0.10	0.30	2	22	6	33	26.18
0.20	0.40	5	24	13	45	31.23
0.30	0.50	7	24	21	53	36.62
0.40	0.60	12	29	27	54	38.06
0.60	0.80	15	26	32	45	35.90
0.70	0.90	13	18	26	32	22.62

Table 4.4: Results for Simon's two stage design for $\beta = 0.9$ and $\alpha = 0.05$, using the minimax criterion

It is straighforward that as the power constraint increases both the minimax and optimal criteria will yield designs that have both higher maximum and expected sample sizes. Increasing the power constraint does not seem to affect which of the two

criteria are more preferrable.

4.2 Designing Simon's Randomized Phase II Designs

We the proceeded with Simon's randomized phase II designs. In his original paper it is mentioned that for three arms and a supposed 15 per cent difference one can calibrate the probability of selecting the most effective treatment to 0.9 by enrolling 44 patients to each arm. We used the r function pselect from the R package "clinfun" to confirm this and to also study the probability of correct selection by changing the supposed response rates while controlling the difference in response rates between the most effective and other treatments to 15 percent.

At this point we mention some differences between the pselect function and Simon's original paper regarding methodology. First, Simon mentions that for specified response rates, response rate difference and probability of correct selection we continuously change the sample size per arm and find which design satisfies the probability of correct selection. In contrast pselect asks first for the sample size and the response rates of all the inteventions (thus we indirectly specify the response rate difference) and then returns the probability of correct selection. This is not significantly different but it is mentioned for clarity. The most important difference between the methodology of the pselect function and Simon's randomized phase II design regarding methodology is that the probability of correct selection when ties are present is not calculated by the pselect function. Even though this probability is relatively small one should still compute it since it represents a possible scenario.

As a result we created a function named `simon_rand` which also takes into account the probability of correct selection when ties are present. The code for this function is presented in the appendix. The resuts of the above methodoogy are presented in table 5 below.

$\delta = 0.15$ **and** $n = 44$		
R.R. of O.T.	**R.R. of M.E.**	**P.C.S**
0.2	0.35	0.90
0.3	0.45	0.87
0.4	0.55	0.86
0.5	0.65	0.86
0.6	0.75	0.88
0.7	0.85	0.92
0.8	0.95	0.98

Table 4.5: Results for Simon's Randomized design for three interventions when we want to detect a 15 per cent difference and enroll 44 patients at each arm.First column is the response rate of the less effective treatments, second column is the response rate of the most effective treatment and the last column is the probability of correct selection (P.C.S.)

For a difference of 15 percent and a sample size of 44 patients per arm or a total sample size of 132 patients table 5 shows that the probability of correct selection remains high for any selection of response rates. Additionally, phase II trials must generally have low sample sizes and enrolling only 132 patients to compare three experimental treatments is a satisfactory sample size for a phase II trial.

4.3 Distribution of UMVUE vs Distribution of MLE

in this section we will compare the distribution of the UMVUE of the response rates computed in chapter 2 with another common estimator of the response rates, the MLE.

The methodology for this comparison is as follows. For different null and alternative hypotheses the respective Simon two-stage designs and their quadruples (n_1, n, r_1, r) have already been computed in the previous section. We will restrict our interest for the case of $H_0 : p = 0.3$ and $H_1 : p = 0.5$ since these are common null and alternative hypothesis. Additionally, the results are not different for other hypotheses such as $H_0 : p = 0.2$ and $H_1 : p = 0.4$. We must also assume what the true response rate when receiving the treatment actually is. We restricted our interest for the cases in which the true response rates when receiving the treatment is $p_i = 0.1, 0.2, 0.3, 0.4, 0.5$. Under these parameters we simulated 1000 times (for each estimator) two binomial random variables of the form:

$$X \sim \mathcal{B}(15, p_i), Y \sim \mathcal{B}(46, p_i)$$

$$X \sim \mathcal{B}(19, p_i), Y \sim \mathcal{B}(39, p_i)$$

where 15 and 46 are the sample sizes for the first and second stage using the optimal criterion for testing $H_0 : p = 0.3$ and $H_1 : p = 0.5$ and 19 and 39 are the sample sizes for the first and second stage using the minimax criterion based on Simon's two stage design (see table 1). At each iteration we estimated the response rates using the UMVUE and MLE, for both criteria (optimal and minimax). The results are presented on plots below:

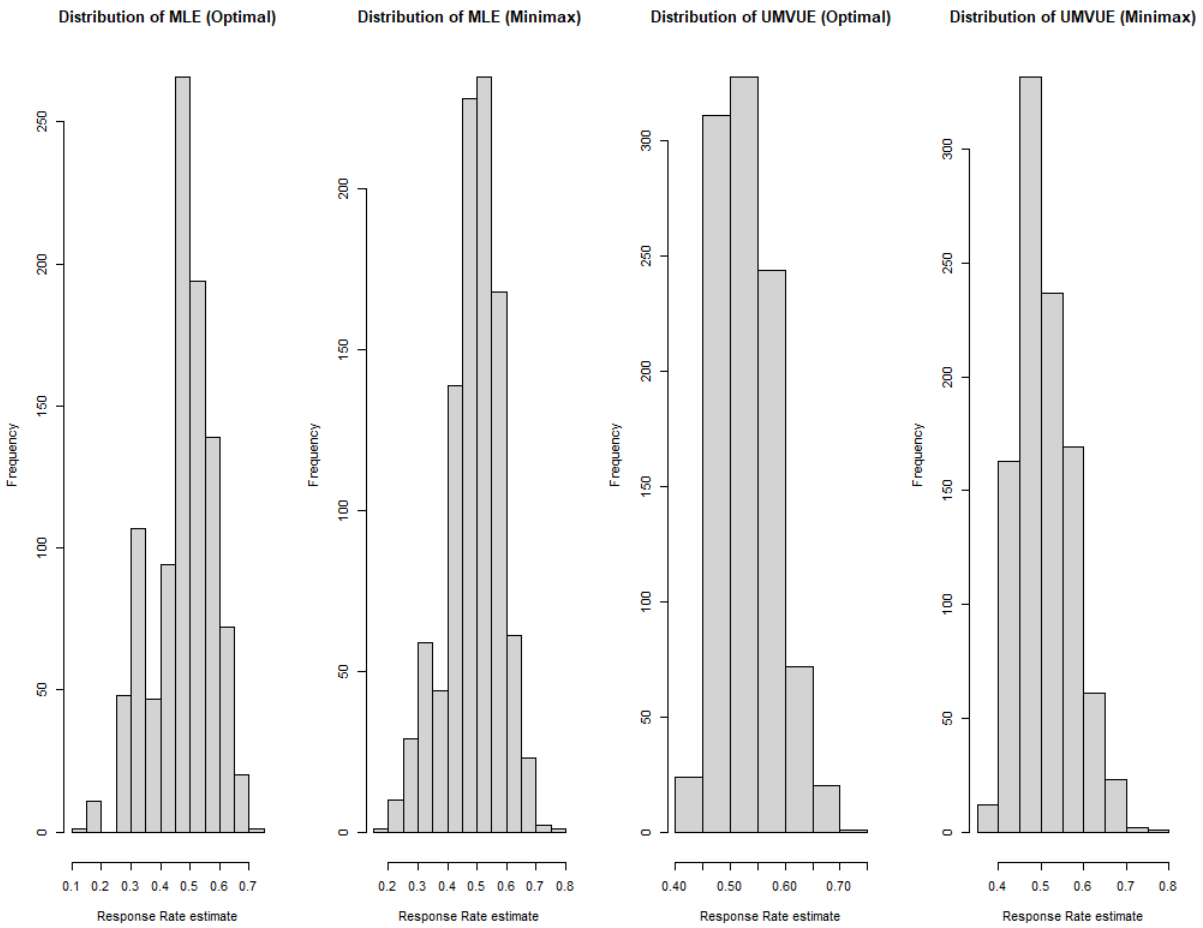

Figure 4.1: Comparison of estimators for true p=0.5

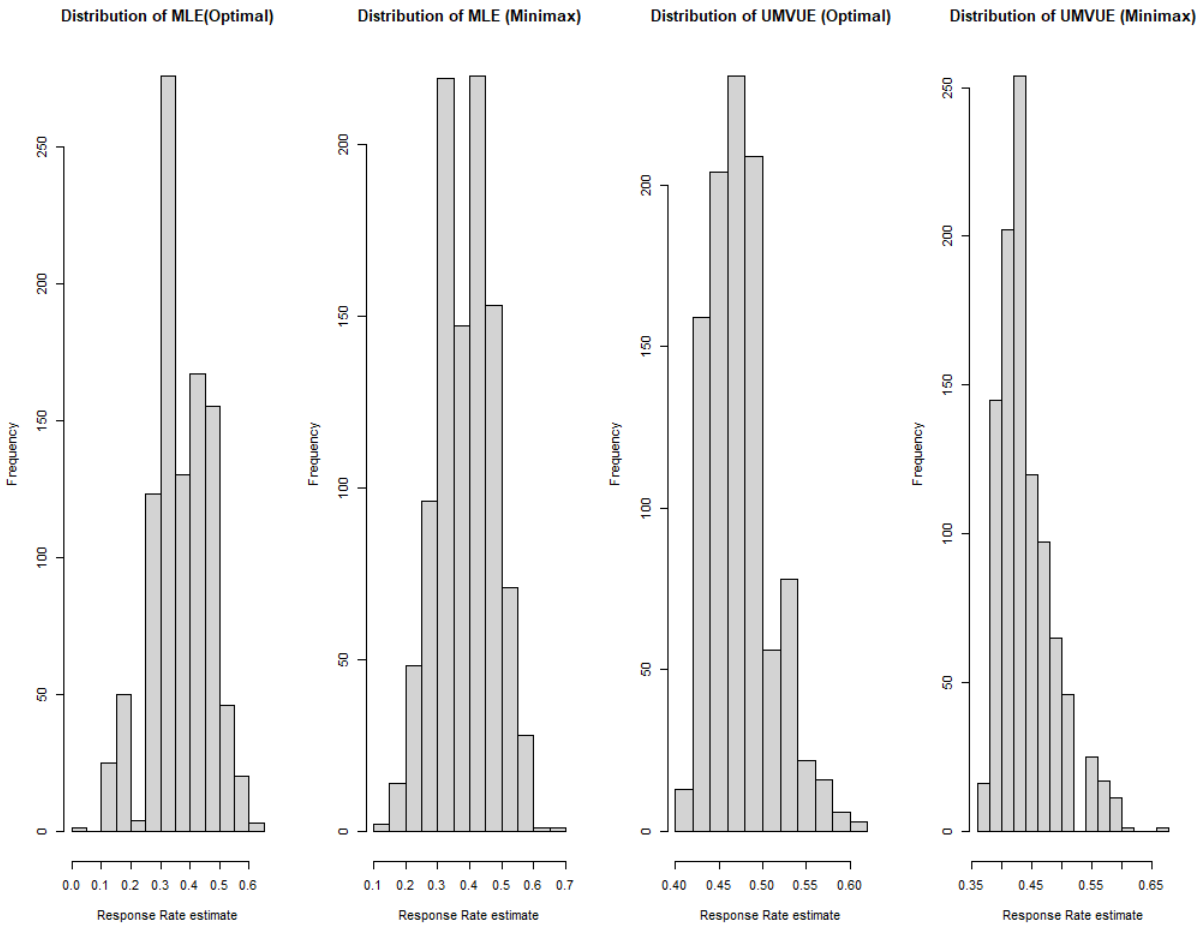

Figure 4.2: Comparison of estimators for true p=0.4

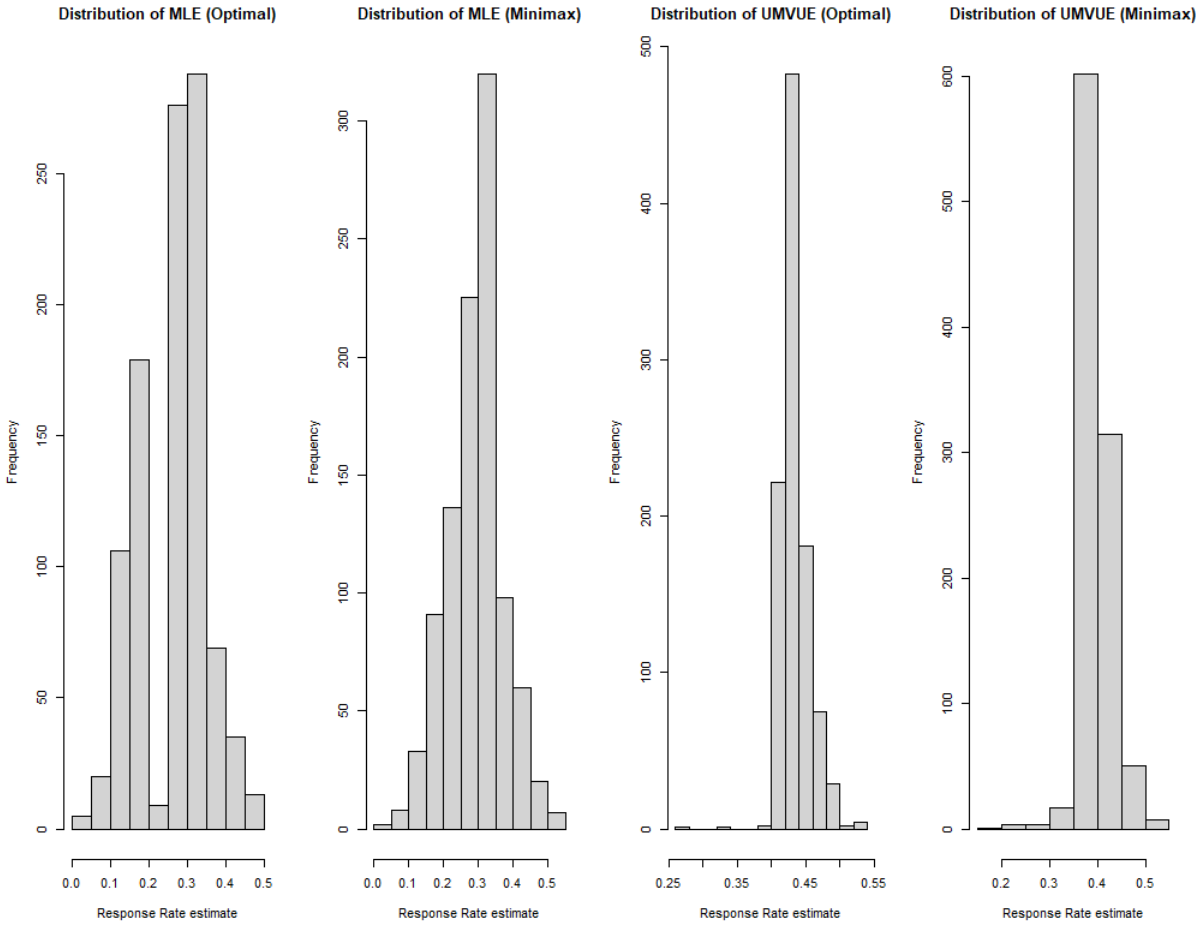

Figure 4.3: Comparison of estimators for true p=0.3

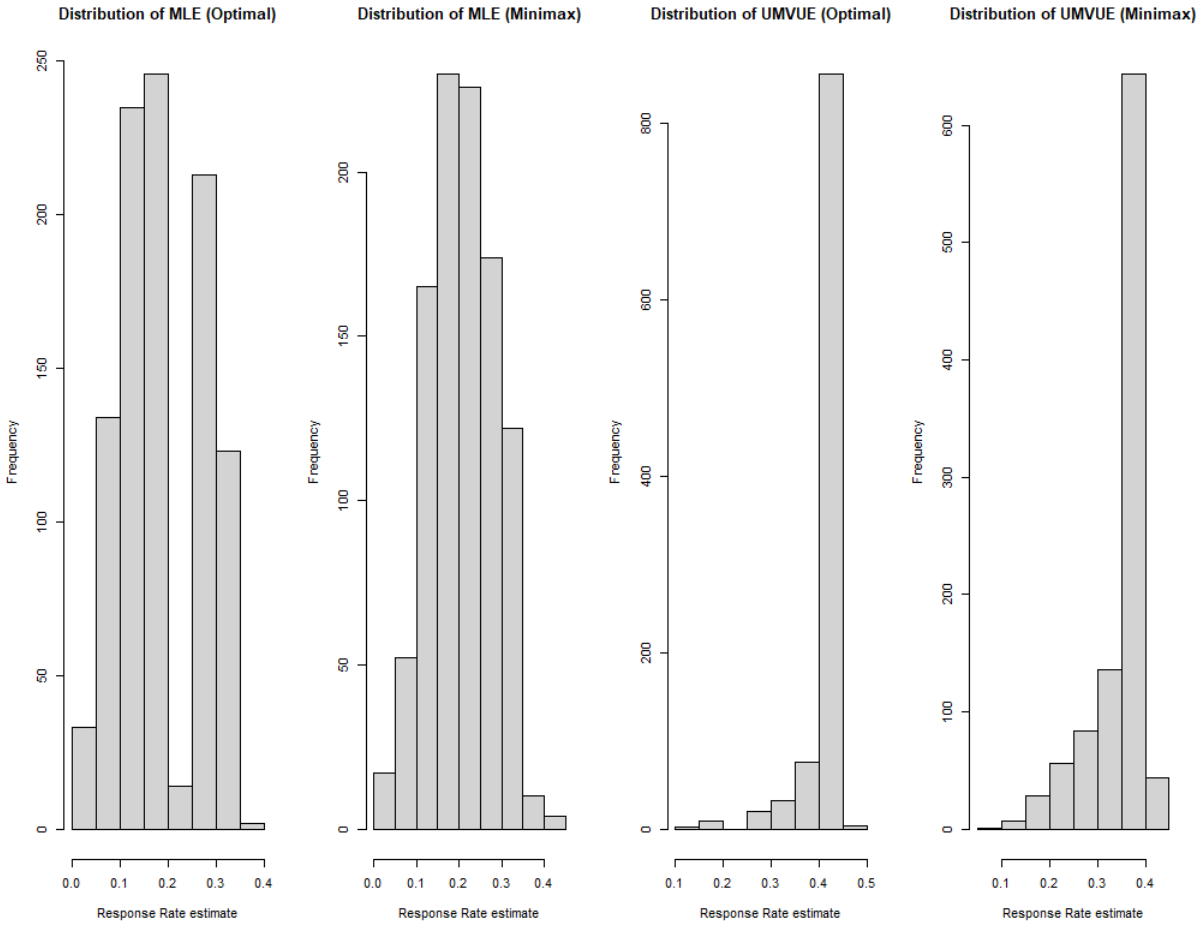

Figure 4.4: Comparison of estimators for true p=0.2

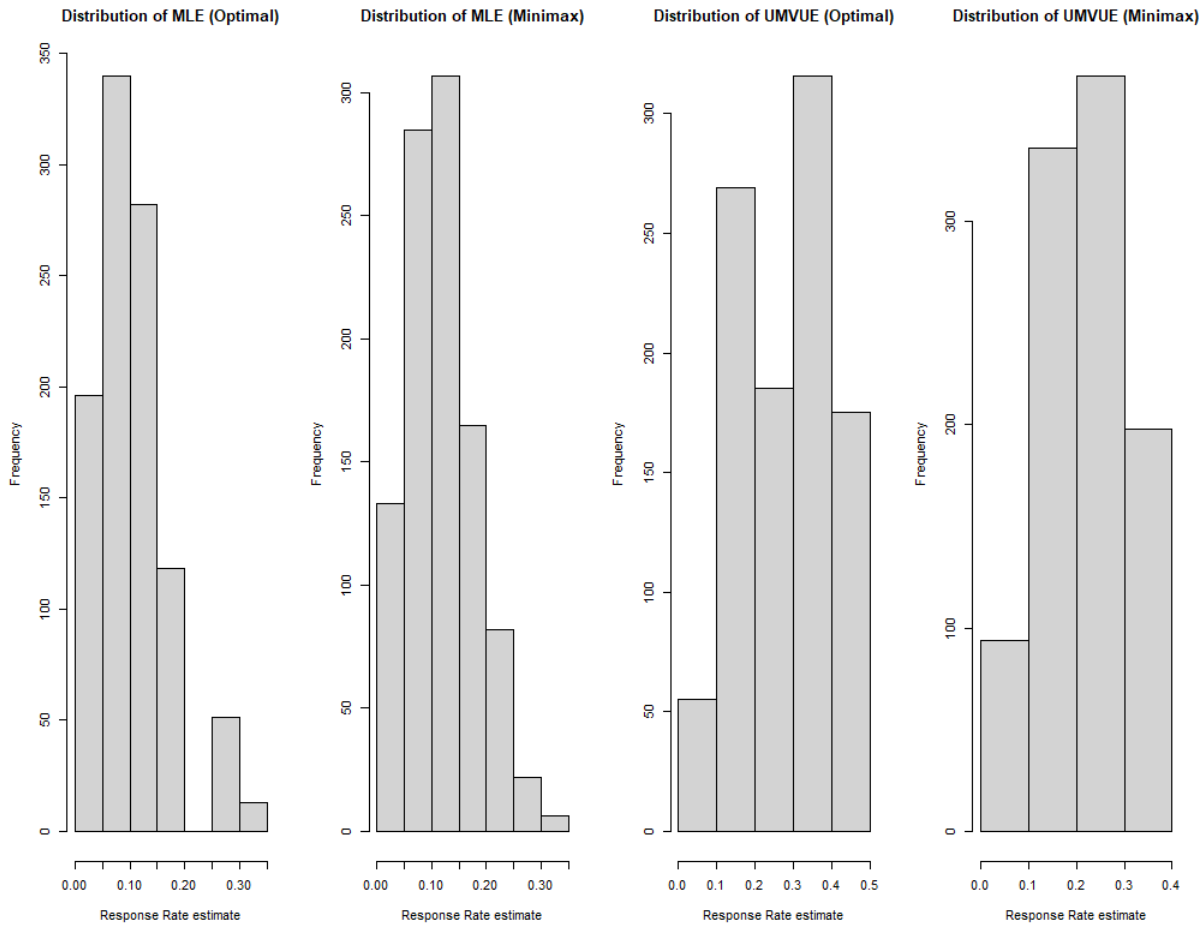

Figure 4.5: Comparison of estimators for true p=0.1

From the plots the minimum variance identity of the UMVUE is obvious. This is an obvious advantage of the UMVUE compared to the MLE which seems to have,in contrast, large variance. In fact, the MLE values seem to not give a lot of information about the value of the true response. In the plots where the actual response rate is 0.4 or 0.3 or 0.2 the MLE has a very wide range, since there are many values ranging from 0 to 0.5 regardless of the actual response rate. Thus looking at the plots of the MLE the actual response rate could be anything between In contrast the UMVUE seems to mostly overestimate the true response rate but since it has minimum variance one can get a more clear picture of the actual response rate. For example for $p = 0.4$ the MLE seems to have, in high frequency, values anywhere between 0.25-0.5 while the plots of the UMVUE for both the optimal and minimax criteria imply that the response is probably at leat 0.4 (which is correct) in this case.

4.4 Power Analysis for Phase III Trials

In this section the sample sizes needed to have a specific degree of power for different classifications of clinical trials (Superiority, Non-Inferiority, Equivalence) were compared (this is called power analysis). In figure (6) below the results are presented for a clinical trial testing the hypotheses:

$$H_0 : \mu_1 \leq \mu_2$$
$$H_1 : \mu_1 > \mu_2$$

with power at $\mu_1 = \mu_2 + 20$ and expected variance $\sigma = 60$.

$$H_0 : \mu_1 \leq \mu_2 - 10$$
$$H_1 : \mu_1 > \mu_2 - 10$$

where we consider that the first superiority trial compares an experimental treatment with placebo and the other two designs compare this treatment to a second experimental treatment. 10 is the non-inferiority margin with power at $\mu_1 = \mu_2 + 0$ and expected variance $\sigma = 60$ (as noted in chapter 3 the non-inferiority margin is always smaller than the possible clinically significant difference, that is N-I margin must be smaller than 20 in this case).

$$H_0 : |\mu_1 - \mu_2| > 10$$
$$H_1 : |\mu_1 - \mu_2| \leq 10$$

where 10 is the equivalence margin with power at $\mu_1 = \mu_2 + 0$ and expected variance $\sigma = 60$.

The results of these three clinical trials are presented in figure 6. The differences in sample sizes needed is obvious. A superiority trial will reach the prespecified power level by recruiting fewer patients than the non-inferiority trial or an equivalence trial. In fact, it can be seen that in order to have a power of 80 percent a non-inferiority trial will approximately have to accrue two times the patients needed in a superiority trial.

We also run a power analysis for the hypotheses

$$H_0 : \mu_1 \leq \mu_2$$
$$H_1 : \mu_1 > \mu_2$$

with power at $\mu_1 = \mu_2 + 10$ and expected variance $\sigma = 60$.

$$H_0 : \mu_1 \leq \mu_2 - 5$$
$$H_1 : \mu_1 > \mu_2 - 5$$

where 5 is the non-inferiority margin with power at $\mu_1 = \mu_2 + 0$ and expected variance $\sigma = 60$.

$$H_0 : |\mu_1 - \mu_2| > 10$$
$$H_1 : |\mu_1 - \mu_2| \leq 10$$

That is, in that second power analysis we assumed a smaller difference that would be clinically meaningful. The resuts are in figure (7).

With these specifications the curves for non-inferiority and equivalency almost become a straight line. This means that the power increases steadily and very slowly wrt sample size. Essentially, since the power must commonly be above 80 percent using

a non-inferiority or equivalency design would practically be infeasible since very large sample sizes and a long period of accrual time would be needed.

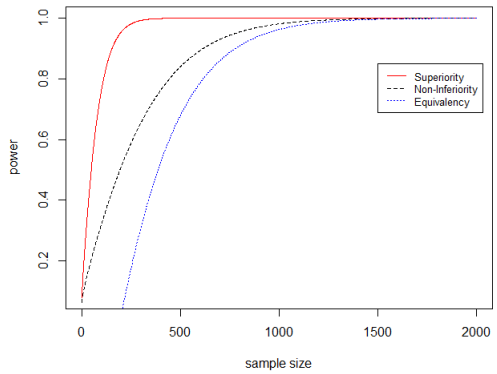

Figure 4.6: Power plot for Superiority,Non-Inferiority and Equivalence trials for a mean clinical difference of 20

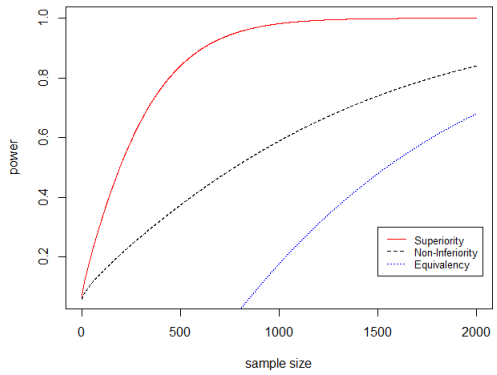

Figure 4.7: Power plot for Superiority,Non-Inferiority and Equivalence trials for a mean clinical difference of 10

For a survival trial, considering that we want to have power at a hazard ratio of 0.75 (that is, the new intervention has a hazard rate 25 percent less) and a type I error of 0.05 the power plot is given below :

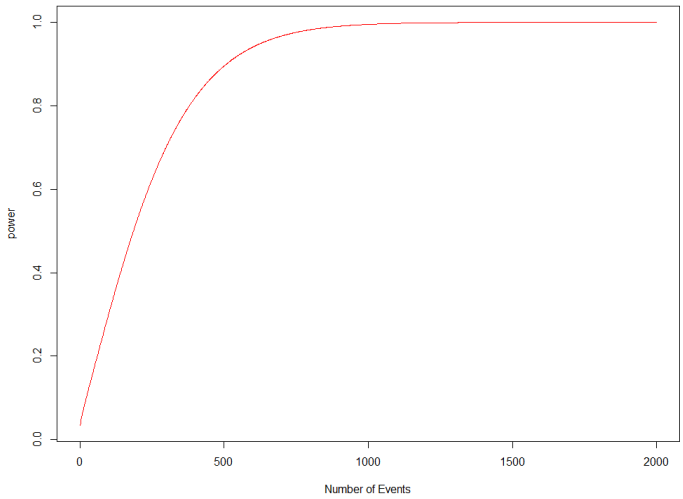

Figure 4.8: Power plot for the alternative $\frac{h_1(t)}{h_1(t)} = 0.75$ with a type I error of 0.05

Note that even though Figure 8 looks similar to a superiority clinical trial with quantitative outcomes this plot visualized the power increase as a function of the number of events (e.g. number of deaths). This implies that the actual sample size needed to observe this number of deaths will be much higher.

4.5 Group Sequential Trials

In this section we analyze and present results for group sequential designs. We center our attention for hypotheses of the form:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

with power at $\mu_1 = \mu_2 + 20$ and expected variance $\sigma = 60$.

In order to test the above hypothesis we must also prespecify the parameters below:

- The type I error rate is below 0.05

- The power at $\mu_1 = \mu_2 + 20$ must be above 0.8
- The number of interim analyses (stages) are 5
- The information or sample sizes accrued at each stage are equal (this means that the sample sizes at each stage will be 0.2,0.4,0.6,0.8,1 of the final sample size)

Having specified these parameters the function gsDesign Package was used to compare the boundary values between Pocock's and O' Brien and Fleming's boundaries as well as compute the sample sizes needed in each case. A graphical illustration of the boundary values is given in figure 9.

Looking at figure 9, as the formula of O' Brien and Fleming's boundaries' suggests, the boundaries are decreasing as we reach the final stage. The area between the two curves (or straight lines in the case of Pocock's boundaries) is the continuation region, meaning that if the computed statistic is inside the area for a specific stage value we will proceed to the next stage (or we will accept the Null hypothesis if we are already at the final stage). Thus from this plot it is easy to see that O' Brien and Fleming's boundaries' have a larger continuation region at the beginning stages compared to Pocock and a slightly smaller continuation region at the final stages. Thus if O' Brien and Fleming's boundaries' are used for a group sequential design we would expect that more patient would be accrued until reaching a conclusion since it is more difficult to reject the null hypothesis at the beginning.

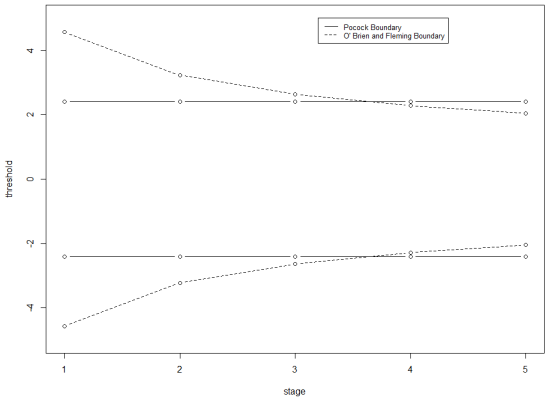

Figure 4.9: Graphical comparison of Pocock and O' Brien Fleming boundaries. The area between the upper and lower boundaries is the continuation area

Afterwards, for the same hypothesis, we studied the sample size needed as well as the boundary values for O' Brien and Fleming's and Pocock's boundaries separately. For the latter, the results are presented below:

O'Brien and Fleming Boundaries			
Number of stages	Sample Size (Per Arm)	Min Threshold	Max Threshold
2	288	1.97	2.79
5	294	2.04	4.56
8	296	2.071	5.85
10	297	2.086	6.59

Table 4.6: Table for Fleming-O' Brien presenting the sample sizes and the values of the first (stage 1) and last (stage 2,5,8 or 10) boundaries

We have computed the first and final boundary to study how the boundaries "behave" for each stage. As the stages increase the sample size needed to achieve power at 0.8 and control the type I error at 0.05 increases as well. Additionally, the thresholds also increase as the stages increase (as it can be seen by the columns Min Threshold and Max Threshold) thus as the stages increases the probability of terminating early decreases.

The following table presents the results for the Pocock Boundaries

Pocock Boundaries		
Number of stages	Sample Size (Per Arm)	Threshold
2	317	2.414
5	351	2.178
8	365	2.512
10	371	2.556

Table 4.7: Table for Fleming-O' Brien presenting the sample sizes and the values of the boundaries boundaries for different stages

4.6 Survival Trial Design Using Historical Data

In this section we present a design for simple two-arm randomized survival trials using data from a previous trial. We will aso perform data analysis on the historical data using the cox regression model as well as the Kaplan Meier estimand to evaluate the survival rates of different subgroups.

The supposed historical dataset will be the dataset 'larynx' from the 'KMsurv' R package. This dataset constains data on 90 males diagnosed with cancer of the larynx during the period 1970–1978 at a Dutch hospital. Times recorded are the intervals (in years) between first treatment and either death or the end of the study (January 1, 1983). Also recorded are the patient's age at the time of diagnosis, the year of diagnosis, and the stage of the patient's cancer.The dataset larynx contains 5 variables:

- stage: The stage of disease (1=stage 1, 2=stage 2, 3=stage 3, 4=stage 4)
- time : Time to death or on-study time, months
- age : Age at diagnosis of larynx cancer
- diagyr : Year of diagnosis of larynx cancer
- delta : Death indicator (0=alive, 1=dead)

For the analysis of the dataset the R packages 'survival' and 'survminer' were used. The variable diagyr which describes the year of diagnosis was deemed to not offer any information on the survival times of the patients and was omitted from the analysis. A Kaplan-Meier survival estimate was fitted on the whole dataset to assess survival rates. Afterwards, a cox model was fitted with all the variables initially to assess which variables affected the survival rates significantly. Graphical tests for the assessment of the proportionality of hazards and the goodness of fit were performed using Schoenfeld and Cox-Snell residuals respectively. For the final model, analyses based on the signicant variables using hazard ratios and Kaplan-Meier curves were performed.

The Kaplan Meier survival function for the whole dataset is given in Figure 4.10:

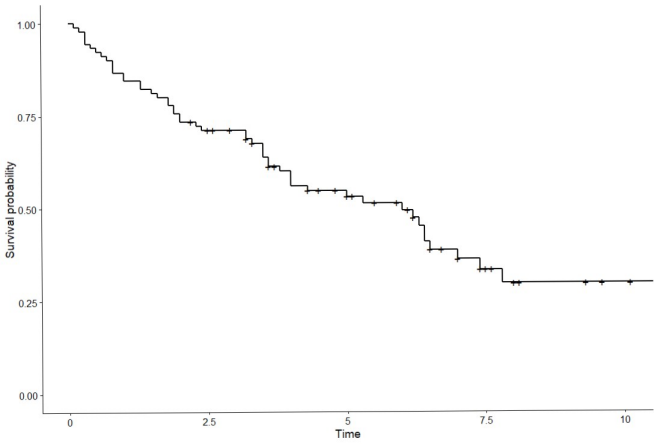

Figure 4.10: Kaplan Meier Estimate of the survival function. The small crossed on the lines represent censoring times.

Here, the small crosses on the line represent the censoring times. The initial fitted Cox model was:

$$h(t|X) = h_0(t)e^{b_2X_{stage_2}+b_3X_{stage_3}+b_4X_{stage_4}+b_5X_{age}}$$

After fitting the model a graphical test of proportional hazards using the Schoenfeld residuals was performed. The results are in Figure 11 below.

These plots were generated using the function ggcoxzph as shown in the appendix. This function automatically produces a smooth curve using the data-points (dots) to help make the results easier to understand. Similarly to the theory of scaled Schoenfeld residuals if the smoothed curve is a straight line parallel to the x-axis the proportionality of hazards will hold. This functions also performs the Schoenfeld individual and global test. From the results shown in figure 11 it was concluded that the proportionality of hazards hold.

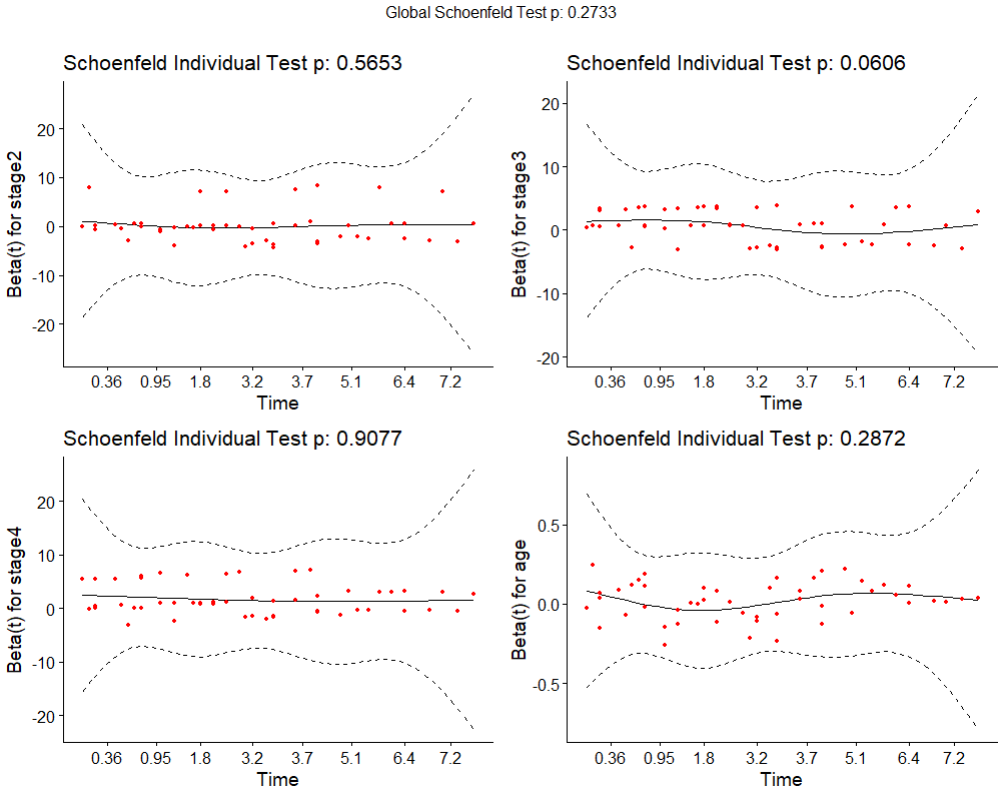

Figure 4.11: Schoenfeld Residuals plot for each one of the variables used in the model $h(t|X) = h_0(t)e^{b_2X_{stage_2}+b_3X_{stage_3}+b_4X_{stage_4}+b_5X_{age}}$.

The results of this fitted model are given in the table below:

Results		
Variable	Coefficient Value	P-value
Stage 2	0.14	0.76
Stage 3	0.64	0.07
Stage 4	1.70	0.0005
Age	0.01	0.1820

Table 4.8: Table of coefficient values and p-values of wald tests for each variable for the model $h(t|X) = h_0(t)e^{b_2X_{stage_2}+b_3X_{stage_3}+b_4X_{stage_4}+b_5X_{age}}$

Age is thus insignificant and it was removed from the model. A cox model with only the stage variable was refitted. The proportional hazards plot is presented in Figure 12 below.

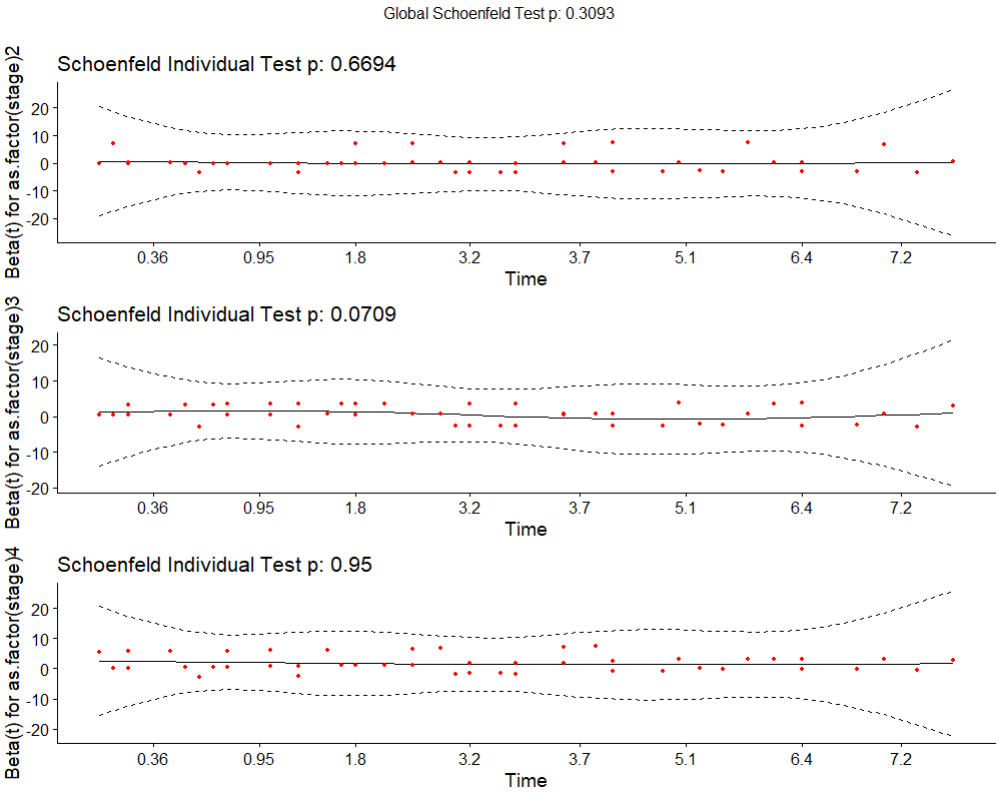

Figure 4.12: Schoenfeld Residuals plot for each one of the variables used in the model $h(t|X) = h_0(t)e^{b_2X_{stage_2}+b_3X_{stage_3}+b_4X_{stage_4}}$

The proportionality of hazards assumption is satisfied. The results of the fitted model are given below:

Results		
Variable	Coefficient Value	P-value
Stage 2	0.064	0.88
Stage 3	0.61	0.08
Stage 4	1.73	0.00035

Table 4.9: Table of coefficient values and p-values of wald tests for each variable for the model $h(t|X) = h_0(t)e^{b_2X_{stage_2}+b_3X_{stage_3}+b_4X_{stage_4}}$

We also compare the two models (the one including age and the one with only the stage variables). The maximized log-likelihood of the model using age is $log\hat{L}_1 = -187.7074$ and the maximized log-likelihood of the model without age is $log\hat{L}_2 = -188.6208$. Comparing the result of the difference $2(log\hat{L}_1 - log\hat{L}_2) = 1.826852$ to a chi square distribution with one degree of freedom it is obvious that the variable age is not significant for the model.

Additionally we graphically chech the goodness of fit of the Cox model using the Cox-Snell residuals. The plot is given below:

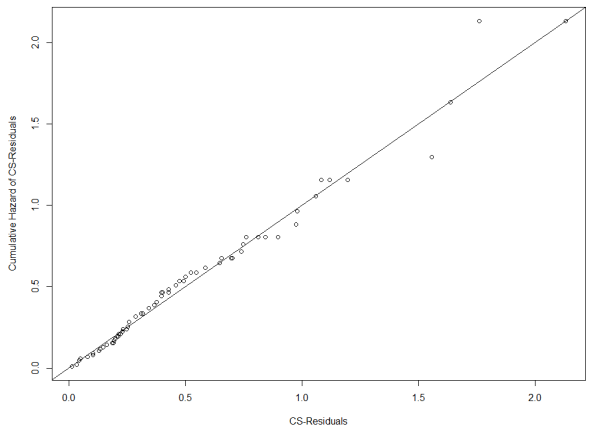

Figure 4.13: Plot of Cox-Snell residuals for the $h(t|X) = h_0(t)e^{b_2X_{stage_2}+b_3X_{stage_3}+b_4X_{stage_4}}$ to assess the goodness of fit.

From figure 13 it seems that the model fits the data relatively well.

Finally we visualize the four different survival functions that occur for the four different stages. The results are presented in the plot below:

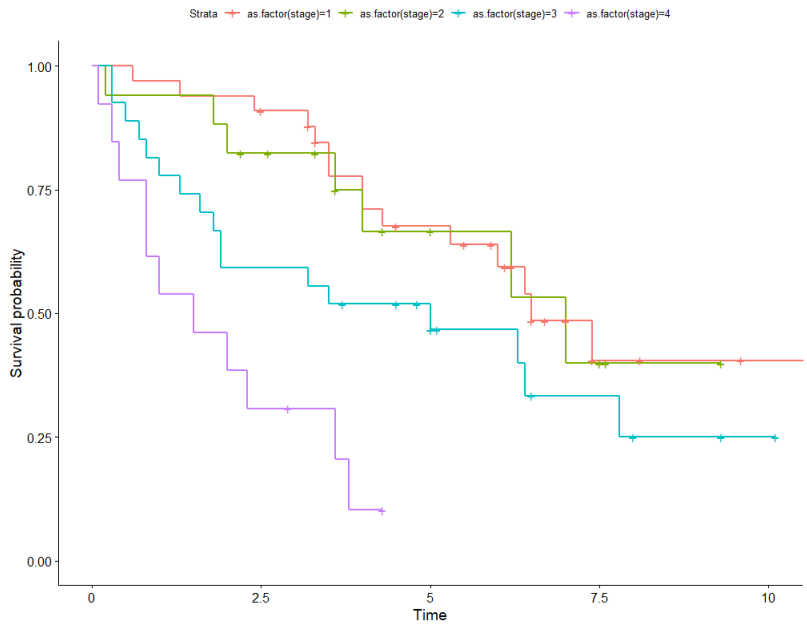

Figure 4.14: Kaplan Meier estimates for each stage group

From this plot it seems that the stage 4 patients have very low survival rates compared to the other stages. In fact the sample size corresponding to the stage 4 patients becomes zero before 5 years. Stage 1 and Stage 2 patients seem to have approximately the same survival rates. Stage 3 patients seem to be in the middle of stage 1 or stage 2 patients and stage 4 patients in terms of survival rates. This exploratory analysis combined with the cox model can help shape the entry criteria of an upcoming clinical trial. For example, suppose we wanted to design a clinical trial for an intervention targeted on stage 1 and stage 2 patients. If these two subgroups had significantly different survival rates than more complex methods of randomization would be needed to ensure that the two groups would have an equal number of stage 1 and stage 2 patients. Otherwise there would be a possibility of biased results. With this analysis it seems that stage 1 and stage 2 patients do not have different survival rates and one would not need to account for a different randomization scheme.

Now suppose that based on this data we want to design a clinical trial for an intervention suitable for stage 4 patients and compare it to the intervention given in the historical data. Based on the historical data the probability of surviving after 4 years is estimated as 0.103 (see appendix[A]). We want to improve the survival rate with the new medicine by increasing the 4-year survival rate to 0.3. We suppose that we want to design a trial with 2 years accrual and 2 years follow-up and we also make the assumption that censoring occurs only in the end of the trial. Additionally we want to achieve 90 percent power and have $\alpha = 0.5$. Based on the historical data of the dataset larynx and the appendix (A) we can calculate from the historical data that:

$$\hat{S}_1(2) = 0.385, \hat{S}_1(3) = 0.308, \hat{S}_1(4) = 0.103$$

Here the index 1 denotes the control group (which uses the therapy from the historical data). Thus for a future clinical trial one can estimate the probability of having an event in the control group as:

$$\hat{p}_1 = 1 - \frac{1}{6}\{\hat{S}_1(2) + 4\hat{S}_1(3) + \hat{S}_1(4)\} = 0.7134$$

Additionally, since we actually want to triple the survival rates with the new medicine we can first use the relation that

$$S_2(t) = [S_1(t)]^{\delta} \iff H_2(t) = \delta H_1(t) \iff h_2(t) = \delta h_1(t)$$

where the second equaltiy occurs by taking logs and the third by differentiating.

The desirable hazard ratio can then be estimated as:

$$\frac{h_2(4)}{h_1(4)} = \frac{ln(S_2(4))}{ln(S_1(4))} = \frac{ln(0.3)}{ln(0.103)} \approx 0.5 = \delta$$

Note that assuming proportionality of hazards this ratio will be the same for any timepoint. Addtionally, the above relation implies that

$$S_2(t) = [S_1(t)]^{0.5}$$

Thus,

$$\hat{p}_2 = 1 - \frac{1}{6}\{(\hat{S}_1(2))^{0.5} + 4(\hat{S}_1(3))^{0.5} + (\hat{S}_1(4))^{0.5}\} = 0.487$$

and the probability of observing a death in either group with randomization 1:1 is

$$P = 0.50 \times 0.7134 + 0.50 \times 0.487 = 0.6$$

Using the power formula for the number of deaths in (3.37) we will get that $d = 275$ to achieve the specified power of 0.9. Then the total sample size will be $n = 275/0.6 = 458$ people. Thus in order to design a trial with a duration of 4 years (2 years follow up and 2 years accrual) a sample size of 458 patients is required to observe 275 events which will give a total power of 90 per cent.

Now assumptions about the accrual rate need to be made. For example, assuming that we expect 15 patients to enroll per month in two years which is the accrual rate we will have a total of $15 \times 24 = 360$ patients. But obviously in such a case the study would be under powered and it should be designed differently. This means that one could extend the trial duration for more months since this would adjust for the low accrual rate. However other parameters come into play when making a trial longer, in terms of budget and human resources.

If we expected 30 patients than we would have $30 \times 24 = 720$ patients, that is many more patients than what is needed to accomplish the required power. Thus under these expectations one could decrease the duration of the trial (for example decrease the accrual duration) since we would still have the required power.

This demonstrates that using in order to design a proper clinical trial experimentation with the trial duration is necessary. Additionally, it demonstrates how non-statistical factors also play a role in determining statistical parameters.

Appendix A

R Code

```
R-CODE FOR THE DESIGN OF TWO STAGE SIMON DESIGNS SECTION 4.1

#Simon two stage design for 20% difference power 0.8 type I error 0.05

ph2simon(0.05,0.25,0.05,0.2)
ph2simon(0.10,0.30,0.05,0.2)
ph2simon(0.20,0.40,0.05,0.2)
ph2simon(0.3,0.5,0.05,0.2)
ph2simon(0.4,0.6,0.05,0.2)
ph2simon(0.6,0.8,0.05,0.2)
ph2simon(0.7,0.9,0.05,0.2)

#Simon two stage design for 20% diff power 0.9 type I error 0.05

ph2simon(0.05,0.25,0.05,0.1)
ph2simon(0.10,0.30,0.05,0.1)
ph2simon(0.20,0.40,0.05,0.1)
ph2simon(0.3,0.5,0.05,0.1)
ph2simon(0.4,0.6,0.05,0.1)
ph2simon(0.6,0.8,0.05,0.1)
ph2simon(0.7,0.9,0.05,0.1)


#Simon two stage design for 15% diff power 0.8 type I error 0.05
```

```
ph2simon(0.05,0.2,0.05,0.2)
ph2simon(0.10,0.25,0.05,0.2)
ph2simon(0.20,0.35,0.05,0.2)
ph2simon(0.30,0.45,0.05,0.2)
ph2simon(0.40,0.55,0.05,0.2)
ph2simon(0.50,0.65,0.05,0.2)
ph2simon(0.6,0.75,0.05,0.2)
ph2simon(0.7,0.85,0.05,0.2)
ph2simon(0.8,0.95,0.05,0.2)


#Simon two stage design for 15% diff power 0.9 type I error 0.05

ph2simon(0.05,0.2,0.05,0.1)
ph2simon(0.10,0.25,0.05,0.1)
ph2simon(0.20,0.35,0.05,0.1)
ph2simon(0.30,0.45,0.05,0.1)
ph2simon(0.40,0.55,0.05,0.1)
ph2simon(0.50,0.65,0.05,0.1)
ph2simon(0.6,0.75,0.05,0.1)
ph2simon(0.7,0.85,0.05,0.1)
ph2simon(0.8,0.95,0.05,0.1)
```

```
#PICK THE WINNER DESIGN\ SIMON'S RANDOMIZED RESIGN SECTION 4.2

#Simon Pick the Winner Design Sample Size n=44 per arm K=3 arms 15% diff

pselect(44,c(0.2,0.2,0.35))
pselect(44,c(0.3,0.3,0.45))
pselect(44,c(0.4,0.2,0.35))
pselect(44,c(0.4,0.2,0.35))
pselect(44,c(0.4,0.2,0.35))
pselect(44,c(0.4,0.2,0.35))


#This is the probability of selecting the correct treatment without ties
#With the code below we create a new function that adds to the previous result the
#of selecting the correct treatment even when there are ties.
i<-0
j<-1:2
x<-0

sum(dbinom(i,44,0.35))

length(i)


#probability for ties can be written as
#sum(choose(2,j)*dbinom(i,44,0.20)**j*pbinom(i-1,44,0.20)**(2-j)/(j+1)) for specifi

for (i in 0:44 ){
x<-x+dbinom(i,44,0.35)*sum(choose(2,j)*(dbinom(i,44,0.20)**j)
*(pbinom(i-1,44,0.20)**(2-j))/(j+1))
}

x

simon_rand<-function(n,prob){
result<-pselect(n,prob)
prob_tot<-as.numeric(result$prob.selection[length(prob),length(prob)])
```

```
j<-1:(length(prob)-1)
for (i in 0:44 ){
prob_tot<-prob_tot+dbinom(i,44,0.35)*sum(choose(2,j)*(dbinom(i,44,0.20)**j)
*(pbinom(i-1,44,0.20)**(2-j))/(j+1))
}
return(prob_tot)
}



#Thus the total probability of correct selection is :

simon_rand(44,c(0.2,0.2,0.35))
simon_rand(44,c(0.3,0.3,0.45))
simon_rand(44,c(0.4,0.4,0.55))
simon_rand(44,c(0.5,0.5,0.65))
simon_rand(44,c(0.6,0.6,0.75))
simon_rand(44,c(0.7,0.7,0.85))
simon_rand(44,c(0.8,0.8,0.95))
```

```
# MLE VS UMVUE SECTION 4.3

#Distribution of MLE and UMVUE null p=0.3 alternative p=0.5 actual p=0.5

z<-rep(0,1000)
c<-rep(0,1000)
r<-rep(0,1000)
t<-rep(0,1000)



for (i in 1:1000){

x<-rbinom(n=1,size=15,prob=0.50)
y<-rbinom(n=1,size=31,prob=0.50)


c[i]<-if(x<=5){
c[i]<-x/15} else{
c[i]<-(x+y)/46
}


z[i]<-twostage.inference(x+y,5,15,46,0.30,0.05)[1]

x<-rbinom(n=1,size=19,prob=0.50)
y<-rbinom(n=1,size=20,prob=0.50)

t[i]<-if(x<=6){
t[i]<-x/19} else{
t[i]<-(x+y)/39
}


r[i]<-twostage.inference(x+y,6,19,39,0.30,0.05)[1]
```

```
}


par(mfrow=c(1,4))
hist(c,main='Distribution of MLE (Optimal)',xlab = 'Response Rate estimate')
hist(t,main='Distribution of MLE (Minimax)',xlab = 'Response Rate estimate')
hist(z,main='Distribution of UMVUE (Optimal)',xlab = 'Response Rate estimate')
hist(r,main='Distribution of UMVUE (Minimax)',xlab = 'Response Rate estimate')



#Distribution of MLE and UMVUE null p=0.3 alternative p=0.5 actual p=0.4


z<-rep(0,1000)
c<-rep(0,1000)
r<-rep(0,1000)
t<-rep(0,1000)




for (i in 1:1000){

x<-rbinom(n=1,size=15,prob=0.40)
y<-rbinom(n=1,size=31,prob=0.40)


c[i]<-if(x<=5){
c[i]<-x/15} else{
c[i]<-(x+y)/46
}


z[i]<-twostage.inference(x+y,5,15,46,0.30,0.05)[1]

x<-rbinom(n=1,size=19,prob=0.40)
y<-rbinom(n=1,size=20,prob=0.40)
```

```
t[i]<-if(x<=6){
t[i]<-x/19} else{
t[i]<-(x+y)/39
}


r[i]<-twostage.inference(x+y,6,19,39,0.30,0.05)[1]

}



par(mfrow=c(1,4))
hist(c,main='Distribution of MLE(Optimal)',xlab = 'Response Rate estimate')
hist(t,main='Distribution of MLE (Minimax)',xlab = 'Response Rate estimate')
hist(z,main='Distribution of UMVUE (Optimal)',xlab = 'Response Rate estimate')
hist(r,main='Distribution of UMVUE (Minimax)',xlab = 'Response Rate estimate')


#Distribution of MLE and UMVUE null p=0.3 alternative p=0.5 actual p=0.3


z<-rep(0,1000)
c<-rep(0,1000)
r<-rep(0,1000)
t<-rep(0,1000)




for (i in 1:1000){

x<-rbinom(n=1,size=15,prob=0.30)
y<-rbinom(n=1,size=31,prob=0.30)


c[i]<-if(x<=5){
```

```
c[i]<-x/15} else{
c[i]<-(x+y)/46
}


z[i]<-twostage.inference(x+y,5,15,46,0.30,0.05)[1]

x<-rbinom(n=1,size=19,prob=0.30)
y<-rbinom(n=1,size=20,prob=0.30)

t[i]<-if(x<=6){
t[i]<-x/19} else{
t[i]<-(x+y)/39
}


r[i]<-twostage.inference(x+y,6,19,39,0.30,0.05)[1]

}




par(mfrow=c(1,4))
hist(c,main='Distribution of MLE (Optimal)',xlab = 'Response Rate estimate')
hist(t,main='Distribution of MLE (Minimax)',xlab = 'Response Rate estimate')
hist(z,main='Distribution of UMVUE (Optimal)',xlab = 'Response Rate estimate')
hist(r,main='Distribution of UMVUE (Minimax)',xlab = 'Response Rate estimate')



#Distribution of MLE and UMVUE null p=0.3 alternative p=0.5 actual p=0.2

z<-rep(0,1000)
c<-rep(0,1000)
r<-rep(0,1000)
```

```
t<-rep(0,1000)


for (i in 1:1000){

x<-rbinom(n=1,size=15,prob=0.20)
y<-rbinom(n=1,size=31,prob=0.20)


c[i]<-if(x<=5){
c[i]<-x/15} else{
c[i]<-(x+y)/46
}


z[i]<-twostage.inference(x+y,5,15,46,0.30,0.05)[1]

x<-rbinom(n=1,size=19,prob=0.20)
y<-rbinom(n=1,size=20,prob=0.20)

t[i]<-if(x<=6){
t[i]<-x/19} else{
t[i]<-(x+y)/39
}


r[i]<-twostage.inference(x+y,6,19,39,0.30,0.05)[1]

}

par(mfrow=c(1,4))
hist(c,main='Distribution of MLE (Optimal)',xlab = 'Response Rate estimate')
hist(t,main='Distribution of MLE (Minimax)',xlab = 'Response Rate estimate')
hist(z,main='Distribution of UMVUE (Optimal)',xlab = 'Response Rate estimate')
hist(r,main='Distribution of UMVUE (Minimax)',xlab = 'Response Rate estimate')

#Distribution of MLE and UMVUE null p=0.3 alternative p=0.5 actual p=0.1
```

```
z<-rep(0,1000)
c<-rep(0,1000)
r<-rep(0,1000)
t<-rep(0,1000)




for (i in 1:1000){

x<-rbinom(n=1,size=15,prob=0.10)
y<-rbinom(n=1,size=31,prob=0.10)


c[i]<-if(x<=5){
c[i]<-x/15} else{
c[i]<-(x+y)/46
}


z[i]<-twostage.inference(x+y,5,15,46,0.30,0.05)[1]

x<-rbinom(n=1,size=19,prob=0.10)
y<-rbinom(n=1,size=20,prob=0.10)

t[i]<-if(x<=6){
t[i]<-x/19} else{
t[i]<-(x+y)/39
}


r[i]<-twostage.inference(x+y,6,19,39,0.30,0.05)[1]

}


par(mfrow=c(1,4))
```

```
hist(c,breaks=5,main='Distribution of MLE (Optimal)',
xlab = 'Response Rate estimate')
hist(t,breaks=5,main='Distribution of MLE (Minimax)',
xlab = 'Response Rate estimate')
hist(z,breaks=5,main='Distribution of UMVUE (Optimal)',
xlab = 'Response Rate estimate')
hist(r,breaks=5,main='Distribution of UMVUE (Minimax)',
xlab = 'Response Rate estimate')
```

```
#POWER IN SUPERIORITY, NON-INFERIORITY AND EQUIVALENCE TRIALS. POWER ANALYSIS
#FOR SURVIVAL OUTCOMES BASED ON NUMBER OF DEATHS.
#SECTION 4.4


#Mean difference is 20

pwr_sup<-rep(0,2000)
pwr_ninf<-rep(0,2000)
pwr_equiv<-rep(0,2000)
i<-0
q<-0
r<-0

for (n in 0:2000){
pwr_sup[n]<-1-pnorm(1.64-(20*sqrt(n))/sqrt(7200))
i<-i+1
}

index<-1:2000


for (n in 0:2000){
pwr_ninf[n]<-pnorm(-1.64+(10*sqrt(n))/sqrt(7200))
q<-q+1
}

for (n in 0:2000){
pwr_equiv[n]<-pnorm(-1.64 +(10*sqrt(n))/sqrt(7200))
-pnorm(1.64 -(10*sqrt(n))/sqrt(7200))
}

pnorm(3)

pwr_equiv

par(mfrow=c(1,1))
```

```
plot(index,pwr_sup,type="l",col="red",xlab = 'sample size',ylab = 'power')
lines(index,pwr_ninf,type="l",lty=2)
lines(index,pwr_equiv,col="blue",type="l",lty=3)
legend(1500,0.85, legend=c("Superiority", "Non-Inferiority","Equivalency"),
col=c("red", "black","blue"),lty=1:3,  cex=0.8)

#Mean difference is 10
pwr_sup<-rep(0,2000)
pwr_ninf<-rep(0,2000)
pwr_equiv<-rep(0,2000)
i<-0
q<-0
r<-0

for (n in 0:2000){
pwr_sup[n]<-pnorm(-1.64+(10*sqrt(n))/sqrt(7200))
}

index<-1:2000


for (n in 0:2000){
pwr_ninf[n]<-pnorm(-1.64+(5*sqrt(n))/sqrt(7200))
}

for (n in 0:2000){
pwr_equiv[n]<-pnorm(-1.64 +(5*sqrt(n))/sqrt(7200))
-pnorm(1.64-(5*sqrt(n))/sqrt(7200))
}


plot(index,pwr_sup,type="l",col="red",xlab = 'sample size',ylab = "power")
lines(index,pwr_ninf,lty=2)
lines(index,pwr_equiv,col="blue",lty=3)
legend(1500,0.30, legend=c("Superiority", "Non-Inferiority","Equivalency"),
col=c("red", "black","blue"),lty=1:3,  cex=0.8)
```

```
#Survival analysis
pwr_surv<-rep(0,2000)
r<-0


for (d in 0:2000){
pwr_surv[d]<-pnorm(-1.96-0.5*log(0.75)*sqrt(d))
}

pwr_surv


index<-1:2000

par(mfrow=c(1,1))
plot(index,pwr_surv,type="l",col="red",xlab = 'Number of Events',ylab = "power")
```

```
#Group Sequential designs SECTION 4.5

install.packages("gsDesign")
library(gsDesign)
install.packages
library(ggpubr)

a<-gsDesign(k=5,test.type=2,alpha=0.025,beta=0.2,
timing = c(0.2,0.4,0.6,0.8,1),sfu="Pocock")
b<-gsDesign(k=5,test.type=2,alpha=0.025,beta=0.2,
timing = c(0.2,0.4,0.6,0.8,1),sfu="OF")

a$upper

boundary_pocock<-rep(2.413176,5)
boundary_fleming<-c(4.561743, 3.225639, 2.633723, 2.280871, 2.040073)


index<-1:5

plot(index,boundary_pocock,type="b",lty=1,xlab = 'stage',
ylab = "threshold",ylim=c(-5,5))
lines(index,boundary_fleming,type="b",lty=2)
lines(index,-boundary_pocock,type="b",lty=1)
lines(index,-boundary_fleming,type="b",lty=2)
legend(3.2,5, legend=c("Pocock Boundary","O' Brien and Fleming Boundary"),
lty=1:2,  cex=0.8)



c<-gsDesign(k=2,test.type=2,alpha=0.025,beta=0.2,timing = c(0.5,1),sfu="Pocock")
d<-gsDesign(k=2,test.type=2,alpha=0.025,beta=0.2,timing = c(0.5,1),sfu="OF")

c$upper
```

```
e<-gsDesign(k=8,test.type=2,alpha=0.025,beta=0.2,
timing = c(0.125,0.25,0.375,0.500,0.625,0.75,0.875,1),sfu="Pocock")
f<-gsDesign(k=8,test.type=2,alpha=0.025,beta=0.2,
timing = c(0.125,0.25,0.375,0.500,0.625,0.75,0.875,1),sfu="OF")


e$upper

g<-gsDesign(k=10,test.type=2,alpha=0.025,beta=0.2,
timing =c(0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1),sfu="Pocock")
h<-gsDesign(k=10,test.type=2,alpha=0.025,beta=0.2,
timing = c(0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1),sfu="OF")

par

inflation_factor_ob_stage_0.025<-c(1.008,1.028,1.037,1.040)

inflation_factor_po_stage_0.025<-c(1.110,1.229,1.279,1.301)


index<-1:4

par(mfrow=c(1,1))
plot(index,inflation_factor_ob_stage,type="b",lty=1,xlab = 'Number of Stages',
ylab = "",ylim=c(0,2))
lines(index,inflation_factor_po_stage,type="b",lty=2)
legend(3.2,5,
legend=c("O' Brien Fleming Inflation Factor","Pocock Inflation Factor"),
lty=1:2,  cex=0.8)
```

```
SURVIVAL TRIAL DESIGN SECTION 4.6

install.packages("asaur")
library(asaur)

install.packages("survival")
library(survival)

install.packages("survminer")
library(survminer)

install.packages("rlang")
library(rlang)


install.packages("dplyr")
library(dplyr)

install.packages("KMsurv")
library(KMsurv)

data(larynx)

str(larynx)

ggsurvplot(survfit(Surv(time,delta) ~1,larynx),conf.int = F,palette="black",
legend.title="")

larynx$stage<-as.factor(larynx$stage)



cox_mod_larynx_2<-coxph(Surv(time,delta)~stage+age,data=larynx)

cox_mod_larynx_2

logLik(cox_mod_larynx_2)
```

```
par(mfrow=c(1,1))
#SCHOENFELD
test.ph = cox.zph(cox_mod_larynx_2,terms=FALSE, singledf=FALSE, global=TRUE)
test.ph



ggcoxzph(test.ph)


cox_mod_larynx<-coxph(Surv(time,delta)~as.factor(stage),data=larynx)

logLik(cox_mod_larynx)

-2*(-logLik(cox_mod_larynx_2)+logLik(cox_mod_larynx))

pchisq(1.826852, df=1, lower.tail=F)



#SCHOENFELD
test.ph =  cox.zph(cox_mod_larynx,terms=FALSE, singledf=FALSE, global=TRUE)
test.ph

ggcoxzph(test.ph)


#COX-SNELL RES
larynx$res3<- residuals(cox_mod_larynx, type="martingale",data=larynx)
##see martingale section in [50]
larynx$csres<-larynx$delta-larynx$res3
surv3<-survfit(Surv(csres,delta)~1,type="fleming-harrington",data=larynx)
summary(surv3)

plot(surv3$time,-log(surv3$surv),
ylab="Cumulative Hazard of CS-Residuals",xlab="CS-Residuals")
```

```
abline(a=0,b=1)

ggsurvplot(survfit(Surv(time,delta) ~as.factor(stage),larynx))

survival<-survfit(Surv(time,delta) ~as.factor(stage),larynx)

summary(survival)
```

Appendix B

UMVUE for Two Stages

In section 2.1.4 a general proof of the UMVUE was presented. However this proof is very complex since it accounts for any stage and also for early termination. In a real setting the stages are mostly two and early termination is allowed only for futility. In this appendix a simpler proof for this case is presented. The joint distribution of $M = 1, 2$ and $S = \sum_{i=1}^{M} X_i$ can then be simplified into:

$$P(M = m, \sum_{i=1}^{M} X_i = s) = \begin{cases} P(X_1 = s) & , M = 1 \\ \sum_{r_1 < x_1 \leq s, x_2 = s - x_1} P(X_1 = x_1) P(X_2 = x_2) & , M = 2 \end{cases}$$

The sum for $M = 2$ follows because $\{M = 2\} = \{r_1 < X_1 \leq n_1\}$, so replacing $M = 2$ with this equality easily yields the sum above. Using the definition of sufficiency we derive in the same way:

$$P(X_1 \dots X_m | M = m, S = s) = \dots = \frac{\prod_{i=1}^{m} P(X_i = x_i)}{P(M = m, S = s)}$$

Expanding the formulas for each value of M in the above fraction will yield a function that does not depend on p, thus sufficiency is proved.

Regarding completeness, the support of the distribution of (M,S) will be the union of the sets:

$$\mathcal{R}_1 = \{(1, s) : s \leq r_1\}$$
$$\mathcal{R}_2 = \{(2, s) : r_1 < s \leq n_1 + n_2\}$$

Thus,

$$E_p(g(M,S)) = \sum_{s=0}^{r_1} g(1,s)c_{1,s}p^s(1-p)^{n_1-s} + \sum_{s=r_1+1}^{n_1+n_2} g(2,s)c_{2,s}p^s(1-p)^{n_1+n_2-s}$$

where $c_{m,s}$ denote the binomial coefficients as in 2.1.4. As in 2.1.4 this is a polynomial of p of finite order and with the same arguments completeness follows.

Using the Rao-Blackwell theorem and the unbiased estimator $\hat{p} = \frac{x_1}{n_1}$:

$$\begin{aligned}\frac{1}{n_1}E(X_1|M=m,S=s) &= \frac{1}{n_1}\sum_{x_1}^{n_1} x_1 P(X_1|M,S=s) \\ &= \dots \begin{cases} \frac{s}{n_1} & ,M=1 \\ \frac{1}{n_1}\sum_{r_1<x_1\leq s, x_2=s-x_1} \frac{x_1 P(X_1=x_1)P(X_2=x_2)}{P(M=2,S=s)} & ,M=2 \end{cases}\end{aligned} \tag{B.1}$$

For $M = 1$ it is obvious that the values inside the sums in (B.1) will be zero except for the value s. For $M = 2$ one needs to use the Bayes formula and notice that $\{M = 2\} = \{r_1 < X_1 \leq n_1\}$. The sum for $M = 2$ follows immediately then.

Bibliography

[1] National Institutes of Health (NIH) Agency (2017). NIH's Definition of a Clinical Trial.

[2] National Institutes of Health (NIH) Agency (2018). NIH Clinical Trials Definition-NIDCD

[3] Smith PG, Morrow RH, Ross DA, editors. Field Trials of Health Interventions: A Toolbox.3rd edition. Oxford (UK): OUP Oxford; 2015 Jun 1. Chapter 2, Types of intervention and their development.

[4] Godby, Mary Earick. "control group". Encyclopedia Britannica, 14 May. 2020, https://www.britannica.com/science/control-group. Accessed September 12, 2022

[5] VCCC Alliance (2021). Clinical Trials Team — Roles and Responsibilities — VCCC Alliance.

[6] European Medicines Agency (EMA) (1996). ICH E3 Guideline for Industry Structure and Con- tent of Clinical Study Reports .

[7] European Medicines Agency (EMA) (2005): Guideline on data monitoring committees.

[8] Yin, Guosheng. (2012). Clinical Trial Design: Bayesian and Frequentist Adaptive Methods.

[9] Evans S. R. (2010). Fundamentals of clinical trial design. Journal of experimental stroke And translational medicine, 3(1), 1927. https://doi.org/10.6030/1939-067x-3.1.19 Accessed: 18/11/2021

[10] Poulopoulou, S. (2013). Adaptive designs in phase II clinical trials, PhD thesis, Athens University of Economics and Business.

[11] Suvarna V. (2010). Phase IV of Drug Development. Perspectives in clinical research, 1(2), 5760.

[12] Cipriani A, Barbui C. What is a clinical trial protocol? Epidemiol Psichiatr Soc. 2010 Apr- Jun;19(2):116-7. PMID: 20815294.Accessed: September 12, 2022

[13] Piantadosi, Steven. (2005). Clinical Trials: A Methodologic Perspective, Second Edition. 10.1002/0471740136. Accessed September 12, 2022

[14] Friedman, Lawrence And Furberg, Curt And DeMets, David And Reboussin, David And Granger, Christopher. (2015). Fundamentals of Clinical Trials. 10.1007/978-3-319-18539-2.Accessed September 12, 2022

[15] Corrigan, Paul. (2010). Mosby's Pocket Dictionary of Medicine, Nursing and Health Professions, Mosby. Mosby Elsevier, Missouri (2009). ISBN: 978-0-323-05291-7. Nurse Education in Practice. 10. 10.1016/j.nepr.2009.10.003. ACCESSED September 12, 2022

[16] Department of Health, Education and Welfare (1979). Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research.

[17] Gupta S. K. (2012). Use of Bayesian statistics in drug development: Advantages and challenges. International journal of applied And basic medical research, 2(1), 36. https://doi.org/10.4103/2229-516X.96789. Accessed September 12, 2022

[18] Pallmann, P., Bedding, A.W., Choodari-Oskooei, B. et al. Adaptive designs in clinical trials: why use them, and how to run and report them.BMC Med 16, 29 (2018). https://doi.org/10.1186/s12916-018-1017-7 Accessed September 12, 2022

[19] Mahajan, R., Gupta, K. (2010). Adaptive design clinical trials: Methodology, challenges and prospect. Indian journal of pharmacology, 42(4), 201207. https://doi.org/10.4103/0253-7613.68417 Accessed September 12, 2022

[20] Wason, J.M.S., Brocklehurst, P. Yap, C. When to keep it simple adaptive designs are not always useful. BMC Med 17, 152 (2019).https://doi.org/10.1186/s12916-019-1391-9 Accessed September 12, 2022

[21] European Medicines Agency (EMA) (2005): GUIDELINE ON DATA MONITORING COMMITTEES.

[22] European Medicines Agency (EMA) (2005): GUIDELINE ON DATA MONITORING COMMITTEES.

[23] Thomas, N. (2008). Historical Control. In Wiley Encyclopedia of Clinical Trials (eds R.B. D'Agostino, L. Sullivan and J. Massaro). https://doi.org/10.1002/9780471462422.eoct347 Accessed September 12, 2022

[24] Simon R. Optimal two-stage designs for phase II clinical trials. Control Clin Trials. 1989 Mar;10(1):1-10. doi: 10.1016/0197-2456(89)90015-9. PMID: 2702835 Accessed September 12, 2022

[25] Halabi, S., Michiels, S. (Eds.). (2019). Textbook of Clinical Trials in Oncology: A Statistical Perspective (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315112084 Accessed September 12, 2022

[26] Jung SH, Chang MN, Kang SJ. Phase II cancer clinical trials with heterogeneous patient populations. J Biopharm Stat. 2012;22(2):312-28. doi: 10.1080/10543406.2010.536873. PMID: 22251176; PMCID: PMC3324125. Accessed September 12, 2022

[27] Jung SH, Randomized Phase II Cancer Clinical Trials (2013). Chapman and Hall/CRC.

[28] Jung SH, Kim KM. On the estimation of the binomial probability in multistage clinical trials. Stat Med. 2004 Mar 30;23(6):881-96. doi: 10.1002/sim.1653. PMID: 15027078. Accessed September 12, 2022

[29] Simon R, Wittes RE, Ellenberg SS. Randomized phase II clinical trials. Cancer Treat Rep. 1985 Dec;69(12):1375-81. PMID: 4075313. PMID: 4075313 Accessed September 12, 2022

[30] Jung SH, Sargent DJ. Randomized phase II clinical trials. J Biopharm Stat. 2014;24(4):802-16. doi: 10.1080/10543406.2014.901343. PMID: 24697589; PMCID: PMC4024090. Accessed September 12, 2022

[31] Everson-Stewart S, Emerson SS. Bio-creep in non-inferiority clinical trials. Stat Med. 2010 Nov 30;29(27):2769-80. doi: 10.1002/sim.4053. PMID: 20809482. Accessed September 12, 2022

[32] Greene CJ, Morland LA, Durkalski VL, Frueh BC. Noninferiority and equivalence designs: issues and implications for mental health research. J Trauma Stress. 2008 Oct;21(5):433-9. doi: 10.1002/jts.20367. PMID: 18956449; PMCID: PMC2696315. Accessed September 12, 2022

[33] Hayes, R. J., Moulton, L. H. (2017). Cluster randomised trials, second edition. CRC Press. https://doi.org/10.4324/9781315370286 Accessed September 12, 2022

[34] Delgado A, Guddati AK. Clinical endpoints in oncology - a primer. Am J Cancer Res. 2021 Apr 15;11(4):1121-1131. PMID: 33948349; PMCID: PMC8085844. Accessed September 12, 2022

[35] Kaplan, E.L. and Meier, P. (1958) Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association, 53, 457-481. http://dx.doi.org/10.1080/01621459.1958.10501452 Accessed September 12, 2022

[36] D.R. Cox, David Oakes (1984). Analysis of Survival Data. Chapman and Hall/CRC.

[37] D. R. COX, Partial likelihood, Biometrika, Volume 62, Issue 2, August 1975, Pages 269–276, https://doi.org/10.1093/biomet/62.2.269 Accessed September 12, 2022

[38] Cox, D. R. (1972). Regression Models and Life-Tables. Journal of the Royal Statistical Society. Series B (Methodological), 34(2), 187–220. http://www.jstor.org/stable/2985181 Accessed September 12, 2022

[39] Tsiatis, A. A. (1981). A Large Sample Study of Cox's Regression Model. The Annals of Statistics, 9(1), 93–108. http://www.jstor.org/stable/2240872 Accessed September 12, 2022

[40] Grambsch, P. M. and Therneau, T. M. (1994). Proportional Hazards Tests and Diagnostics Based on Weighted Residuals. Biometrika, 81(3), 515–526. https://doi.org/10.2307/2337123 Accessed September 12, 2022

[41] Schoenfeld, D. (1982). Partial Residuals for The Proportional Hazards Regression Model. Biometrika, 69(1), 239–241. https://doi.org/10.2307/2335876 Accessed September 12, 2022

[42] DeMets D. and Cook T. (2008). Introduction to Statistical Methods for Clinical Trials. Chapman and Hall/CRC.

[43] Wu, J. (2018). Statistical Methods for Survival Trial Design: With Applications to Cancer Clinical Trials Using R (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780429470172 Accessed September 12, 2022

[44] Schoenfeld, D. (1981). The Asymptotic Properties of Nonparametric Tests for Comparing Survival Distributions. Biometrika, 68(1), 316–319. https://doi.org/10.2307/2335833 Accessed September 12, 2022

[45] Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. J Pharmacokinet Biopharm. 1987 Dec;15(6):657-80. doi: 10.1007/BF01068419. PMID: 3450848. Accessed September 12, 2022

[46] Jennison, C. and Turnbull, B. W. (2000). Group sequential methods with applications to clinical trials. Chapman and Hall/CRC.

[47] Proschan M.A., Lan G.K.K., Wittes J.T. (2006). Statistical Monitoring of Clinical Trials A Unified Approach. Springer New York, NY.https://doi.org/10.1007/978-0-387-44970-8 Accessed September 12, 2022

[48] Kalbfleisch, J.D. and Prentice, R.L. (2002) The Statistical Analysis of Failure Time Data. 2nd Edition, John Wiley and Sons, New York.

[49] Lehmann, E. L., Romano, J. P. (2005). Testing statistical hypotheses. New York: Springer. ISBN: 0-387-98864-5 Accessed September 12, 2022

[50] Collett, D. (1993): Modelling survival data in medical research. Chapman and Hall/CRC

[51] Chang MN, Therneau TM, Wieand HS, Cha SS: Designs for group sequential phase II clinical trials. Biometrics 43:865-874, 1987