A thesis presented for the degree of M.Sc. in Statistics

Clinical trial designs for detecting gene signatures

Angeliki Skandali

September 2019


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ACKNOWLEDGMENTS

Throughout the writing of this dissertation I have received a great deal of support and assistance. I would first like to thank my supervisor, Professor Dimitris Karlis of the Statistics Department at Athens University of Economics and Business, for his dedicated involvement in every step during this process. Without his support and expertise in organizing my methodology, this thesis would have never been accomplished.

At this point, I gratefully acknowledge the generosity to Professor of Biostatistics at the University of Athens and Director of Frontier Science Foundation-Hellas (FSF-H) Urania Dafni, for offering me a scholarship. Due to this collaboration, I had the opportunity to conduct my thesis in such an interesting field as Clinical Trial Research.

Finally I must express my gratitude to my family and Antonis for their love, inspiration and patience. Throughout my years of study, they always make sure of providing me with a great deal of support to overcome any difficulties. Without them the writing of this thesis would not have been possible. Thank you.


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ABSTRACT

Clinical trials have contributed in a great extent to the evolution of medical research, through the examination of treatment efficacy. Tumor biology has changed in a way that most tumors are heterogeneous and only a subset of patients with a particular cancer has the potential to benefit from a treatment. This phenomenon was created a shift towards the traditional way of comparing treatments. In recent years, most clinical trials are conducted in a 2-stage design, where first a sensitive subgroup is identified and then the treatment efficacy is calculated for both the sensitive subgroup and all population.

In this study we are interested in creating an adaptive design for detecting gene signatures and prospectively identify a target subpopulation when having time-to-event endpoints as a primary outcome. For this purpose, we will use the Cox proportional hazard model with only parameters the treatment effect and its interaction with the expression genes. This interaction has a crucial role in our design, as it determines which genes will be included in our signature. Moreover, as far as the signature is developed, we will set a classification rule based in the genes belonging in the signature and the hazard ratio of each patient, to identify the sensitive subpopulation.

Finally, we are focusing to the evaluation of the treatment by conducting one hypothesis test for all randomized patients and a second only for the sensitive subgroup. The performance of the design will be measured in terms of statistical power, regarding two different scenarios and by alternating several parameters such as the sample size.


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Contents

List of Figures

List of Tables

Chapter 1

Clinical Trials

1.1 Introduction to clinical trials

1.1.1 Definition and importance of clinical trials

Since early civilizations, people have been concerned about the quality and safety of their life. Consequently, in the pursuit of a better living standard, people have contributed in the advancement of science and medicine to a great extent. Every year among other medical innovations, there is an increasing number of new treatments designed to cure people from a simple flu to more serious illnesses such as cancer. All these upcoming treatments require proper investigation in order to figure out any beneficial effects on patients.

Throughout these years, clinical trials have become the most widespread method of obtaining reliable information about any positive aspects and risks of new therapeutic products or interventions. Besides that, clinical trials

highlight the benefits and safety of existing therapies too, providing doctors and patients with the option of choosing between alternative treatments.

Clinical as a term refers to the medical care of real patients. So, *clinical trials* can be described as an experiment testing the cause-and-effect relationship between a medical intervention and health outcome on human volunteers or research participants [Nellhaus and Davies, 2017a]. More specific, according to the U.S National Institutes of Health (NIH), clinical trial is a prospective biomedical research study of human subjects, designed to answer specific questions about biomedical or behavioral interventions [NIH,].

1.1.2 Types and Phases of clinical trials

It is widely known that medical research has evolved dramatically and nowadays covers a wide research spectrum. This phenomenon made clinical institutes to categorize the trials depending on what researchers are studying. Below are the descriptions of the types reported by the FDA [FDA,]:

- **Treatment trials** that test new treatments, new combinations of drugs, new techniques in surgery, radiation therapy or even medical devices.
- **Prevention trials** which are designed to identify ways to prevent a disease in people who never had the disease or to prevent the disease from returning. The approaches used can be the use of medicines, lifestyle changes or even dietary supplements such as vitamins.
- **Diagnostic trials** that aim to identify improvements in tests or methods used to diagnose disease.

- **Screening trials** that look for ways to detect specific conditions before the patient has any symptoms of the disease.

- **Quality of life** or supportive care trials that explore and measure ways to improve the comfort and quality of life for people suffering from chronic conditions or diseases.

In summary, it is obvious that clinical trials are not related only with drugs that most of the people believe, but also with treatments, medical devices, lifestyle changes and disease prevention. Furthermore, it is very important to highlight that even though clinical trials are divided into a variety of types, where different research topics are being investigated, all trials serve a common purpose and that is to study how safe, effective and helpful several modifications are in human lives.

Another important feature of clinical trials is that they often advance through four phases to test a new treatment, look for the appropriate dosage and search for side effects. In detail, when clinical research is used to evaluate medications, devices or treatments, the experiment is conducted in phases, where each phase has its purpose and help scientists answer different questions. If researchers, after the three phases, find an intervention to be safe and effective, the FDA approves it for clinical use and continue tracking its effects. Clinical trials of drugs are usually described based on their phase. The FDA typically requires Phase I, II, and III trials to be conducted to determine if the drug can be approved for use.

To further explain the purpose of each phase, the four phases are presented below, according to the book of Steven Piantadosi [Piantadosi, 2017]:

Phase I trials: Researchers test an experimental drug or treatment in a small group of people for the first time. In the process of testing, they attempt to find the best dose of a drug and identify side effects. The participants are 20 to 100 healthy volunteers or people with the disease and the length of the study is several months. Approximately 70% of the drugs move up to the following stage.

Phase II trials: The experimental drug or treatment is given to a larger group of people at a fixed dose in order to determine the efficacy and safety of the drug. This phase can last from several months to 2 years and the participants are up to hundred people with the disease. Only 33% of the drugs move up to the following stage.

Phase III trials: In this phase, there are comparative trials that evaluate the effectiveness of the new treatment relative to a standard one also known as control arm. The participants count up to 3,000 volunteers who have the disease or condition and the length of the study is 1 to 4 years. Only 25-30% of the drugs move up to the following stage.

Phase IV trials: After the FDA approves a medicine and it is made available to the public, researchers track its safety in the general population, seeking more information about a medicine or treatment's benefits and optimal use. The participants at this stage are several thousands volunteers who have the disease. That means significant difference at the amount of people participating in the last stage from the others.

Before describing the fundamentals and principles of clinical trials, a brief history of clinical trials is presented in the next section, which outlines how

clinical trials evolved throughout the years and the key moments that made medical research better morally and scientifically.

1.1.3 A brief history of clinical trials

As mentioned before, people's curiosity about quality and safety of their life actually began thousands of years ago. In fact, the first controlled clinical trial was conducted by James Lind in 1747 when he tried to cure twelve sailors suffered by scurvy [Dodgson, 2006]. In his book "A Treatise on the Scurvy" [Lind, 1980], is mentioned that he placed all sailors on the same diet, but fed one group with additional items such as cider and vinegar, while the other group with lemon juice. The group who had the lemon juice supplement recovered from scurvy in just six days.

Following the enactment and publication of Lind's work, an increasing number of clinical trials took place and especially studies comparing treatments for certain disease (comparative trials). A very important fact that demonstrates Lind's great contribution to clinical trials is that in 2003, the publicity and popularity of *James Lind Library* [1] has made 20 May to be designated International Clinical Trials Day [Chalmers et al., 2008]. In addition to Lind's work, in 1865 a French physiologist Claude Bernard published the book "Introduction to the Study of Experimental Medicine" [Bernard, 1957], whose goal was to urge medical professionals to consider applying scientific principles to their standards of care.

[1] a website established to commemorate 250th anniversary of Lind's book publication

Bernard wrote:

> *"To learn we must necessarily reason about what we have observed, compare the facts, and judge them by other facts as controls."*

His aim to use scientific methods in medicine ultimately became a cornerstone in modern clinical trial operations and therefore clinical trials follow sound scientific principles. Over the next years, the scientific benefits of conducting clinical trials and the related medical evolution continued to improve at a seemingly exponential rate, especially in the early 1900's [Nellhaus and Davies, 2017b].

Despite all the positive outcomes that clinical trials brought in medical research, there were studies that had negative impact on their participants and considered as inappropriate or unethical. For instance, the investigators of two widely condemned studies in the 1970s and 1980s, the Tuskegee study of men with untreated syphilis and the New Zealand study of women with carcinoma exploited morally and physically the research subjects [Paul and Brookes, 2015] . These actions took place without any regard to the standards of Nuremberg Code, a set of research ethics principles for human experimentation created as a result of the Nuremberg trials at the end of Second World War [Grodin, 1992]. At that moment was proved that a greater focus on the ethics of clinical trials and subjects protection was necessary.

The following years in the pursue of putting humanity over science, the World Medical Association developed an international manuscript called the

Declaration of Helsinki [Association et al., 2013], in order to expand the Nuremberg code. Since its adoption in 1964, the Declaration of Helsinki has undergone a total of 8 revisions, the most recent in 2000. As a result, clinical trials not only began to follow ethical guidelines, but also to be examined before their beginning by government authorities or national institutes including the Food and Drug Administration (FDA) of the United States, the European Medicines Agency (EMEA), and Japan's Pharmaceuticals and Medical Devices Agency (PMDA).

1.2 Fundamentals of clinical trials

1.2.1 Study Approval and Protocol

Starting a clinical trial is not a simple task. From the trial approval to study completion, the research group has to monitor a wide variety of complex tasks, complying with regulations all the way. To begin with, before the potential therapy is applied in human beings, scientists perform laboratory tests in animals to test the safety and efficacy of the intervention. If these studies show favorable results, governing bodies called **Institutional Review Boards** (IRB) in accordance with the FDA regulations have to approve that the therapy can be tested in humans, in a clinical trial.

IRB is a type of committee that serves an important role in the protection of rights and welfare of human research subjects. In order to accomplish this purpose, they conduct some form of analysis regarding the risk and benefits of the specific intervention, by reviewing in depth the research protocol (see next paragraph) and related materials such as participants' rights. It is of

great importance to mention that the approval of a therapy does not mean that it is safe or effective, but only that the trial can be legally conducted.

The **research protocol** mentioned above, is a written plan for a clinical trial, as it contains the precise study plan and addresses every aspect of the planned trial. It is the keystone of the planning process and deserves much of the investigators' attention before the study is submitted for approval [Piantadosi, 2017]. Moreover, the protocol must be carefully designed in order to balance the potential benefits of a trial with the risk to participants. The total length of a protocol will often be 50 pages on average and some of the most essential sections included are the following:

- **Objectives of the Study** is a section that describes the reasons of conducting the study as well as the goals.
- **Statistical Considerations** that address every statistical detail from the sample size selection to the final analysis.
- **Data Recording, Management and Monitoring** is a section where instructions for the collection and review of the data are provided.
- **Study Design** that includes details regarding the treatment plan, doses, expected side effects and actions to take.
- **Study Calendar** that highlights the expected duration of the study and specific instructions on how to manage a situation where a participant is leaving the trial for a specific reason.
- **Eligibility and Exclusion criteria** is a section that has the guideline for who can or cannot participate in the study. The criteria differ from

study to study. They may include age, gender, medical history, and current health status.

Despite the general importance of the study protocol, there is a chance that the trial plans are not followed precisely in all patients. This normally can happen as a consequence of unexpected events or wrong clinical judgment about what is good medical practice for a specific patient. Minor deviations from the protocol are very common and most of the times they do not have negative impact in the process of the trial, but major deviations may affect the validity of the trial [Piantadosi, 2017].

1.2.2 Aspects of participating in a clinical trial

Each research study is different, so the choice of the participants is absolutely dependent to the study goal. The trial may require healthy participants or participants who are affected by the specific disease being examined in the trial. The stage of illness, previous treatment history, age, and other factors can determine one's eligibility. Researchers weigh the study needs and decide who can participate in it, defining the Eligibility criteria listed in the study protocol.

Benefits

Healthy volunteers are talking part to clinical trials with a view to ameliorate medical care and help people in need. Moreover, people with an illness or disease decide to be a part of a clinical trial so as to take advantage of all the positive aspects that a well designed clinical study has to offer. Benefits of

participating in a such a study include free access to new treatments before they are widely available as well as careful medical attention from doctors and other healthcare professionals. Last benefit but not least is that participants may have their last chance to cure their illness with the application of new medicines as any of the previous standard treatments did not work well on them.

Risks

Despite the mentioned advantages, something very important to consider is that clinical trials undermine unavoidable risks. The reason is that even though trials are designed to minimize the risks to all participants, there is an inherent uncertainty in all medical research studies involving new treatments and in interventions not proven beneficial yet. The NIH lists two of the most common risks participants are subject to:

1. Most clinical trials have unpleasant side effects, which often last only a short time. Although it is extremely rare that participants experience life-threatening situations, some may be subject to complications that require medical attention.

2. The study may require more substantial time and attention than a standard treatment would. This may include often visits to the hospital, more blood tests or complex dosage schedules without knowing if the treatment will be effective at last.

Informed Consent

Clinical trials are regulated to ensure that participants in clinical trials are protected. An important safeguard measure to ensure safety of participants is through the Informed Consent process. It is a procedure which aims at answering all participants' questions existed regarding the benefits, risks and nature of the research study. Only participants who sign an Informed Consent document can enter the trial. With their signature, participants confirm that they are willing to be a part of the trial and all the important facts about the trial have been explained to them.

Signing the informed consent document is not binding process, so participants have the right to withdraw from the trial at any point in time.

1.2.3 Ethical Principles

In every clinical trial patients place a great deal of trust in the investigator's expertise and in many circumstances they participate in trials when the chance of personal benefit is low. Actually, they willingly provide information that cannot be obtained in any other way and that makes them primary factors for the conducting of a trial. As a consequence, research volunteers clearly deserve the gratitude of the world community and their dignity must be protected by the researcher of the study.

History has shown that in the first years of conducting clinical trials, investigators did not always respect the rights of the participants. There were evidence of inappropriate and unscientific behavior of physicians and absence of international standards for the ethics of experimentation of the human

subjects. Only after the establishment of Nuremberg Code in 1947 several principles of ethical conduct came into the surface. Nowadays in the modern medical practice there are three principles of ethics widely accepted that were first outlined in the Belmont report [Department of Health et al., 2014]. So following the Belmont report the ethical clinical research is guided by the principles of respect, beneficence and justice.

Respect for persons

Respect for persons is embodied in the Informed Consent implying that the subject's cooperation is voluntary and that all information pertaining to the subject is held in confidence. Moreover, the participant must be aware of the benefits as well as the possible risks of the study. This principle must be valid even for individuals that restrict their personal autonomy because of mental disability or a particular illness. These people need to be protected by all means and should be excluded from certain research activities.

Beneficence and Nonmaleficence

Beneficence is a principle that reflects the patient's right to be treated in a beneficial way and admire all the positive aspects a clinical has to offer. Nonmaleficence has its roots in the Hippocratic Oath and is the investigator's duty to avoid side effects. If not to avoid, at least to minimize them at a point that the individuals enrolled in the trial will not be subject to any harm. Both principles are very important, so in order to be satisfied, the physician must investigate an important question and place an appropriate risk-benefit ration for the participants.

Justice

Justice is the principle that takes into account all the processes by which populations are selected for study to ensure that the results benefit the community and as a result to choose individuals who may be likely to benefit from the study. Furthermore, justice is very important when dealing with vulnerable populations because investigators have to make sure that anyone will exploit them during the trial.

1.2.4 Defining the Research Question

Arguably translating the objective of a study in a targeted research question is the most important and challenging part of research, especially before the beginning of the study. The primary study question should integrate a subject that is of greatest interest to investigators and sponsors. It is the question around where all the study design and reporting of results will be emphasized. Alvan Feinstein wrote in his book *"Clinical Epidemiology, the Architecture of Clinical Research"* [Feinstein, 1995]:

> "When thinking of a new study, the most important task is to define its goals and to choose a design to match those goals."

Furthermore, most of the times, the research question is formulated into a specific hypothesis stating what the investigators expect to find [Giuffrida, 2016]. In a clinical trial, the hypothesis is typically how the primary outcome in the experimental group is expected to compare with the outcome in standard or control group. For instance, in cancer trials when new treatments are

compared with existing ones, the hypothesis is defined by the difference that participants had in overall survival. In general, the way the hypothesis is tested in a trial is by analyzing and interpreting the outcomes of participants in each group [Piantadosi, 2017].

In fact, by defining the question, researchers have to recruit the optimal target population, the experimental and control interventions, the specific details of comparison, the primary outcome and the time frame. These elements are commonly referred to as the PICOT (population, intervention, comparator, outcome, and time-frame) format [Giuffrida, 2016].

Population

The selection of population for a research study is a very challenging task because it has to fit properly in the scientific question. Researchers aim to obtain a representative sample of human subjects with a view to generalize the results of the study to the community. For this reason, the target groups most of the times share similar characteristics like the same age or stage of disease. In this way the results are more accurate for the people with same characteristics not participating in the trial. After recruiting the optimal population for a clinical study, another questionable matter arises and that is the particapants' allocation in the treatments groups. This choice determines in a great extent the final results, because in the case that the investigator selects the treatment assignment in his way of thinking (*non-randomized allocation*), the study results will be biased. More information about the ways to properly allocate participants in the treatments groups are in Section 1.3.1, where the principle of *randomization* is mentioned.

Intervention

Intervention is the specific treatment, medication or in general behavior being examined in a research study. For experimental studies the intervention is normally a treatment or action to which study subjects are assigned by investigators. While for observational studies, it could be an intervention that is applied to study subjects independent of the research study. There is a big deal of intervention examined in clinical research and it varies according to the type of the trial. The majority of the studies use new medicines or treatments. For instance, a phase II cancer trial listed in NIH site, examined how well *durvalumab* and *tremelimumab* work in treating patients with stage IV lung cancer. Plenty of other interventions are used like a medical device or even a surgery. Intervention's magnitude comes with the disease's severity. As a matter of fact, while in oncology trials heavy medicines are examined, there are trials were the intervention is a dietary supplement.

Comparator or Control

The alternative to the intervention being studied. More specific, a research study permits a comparison of subjects treated with the new agent with a suitable control population, so that the effect of the new agent can be determined. Experimental studies always have a comparison, whereas some observational studies do not. The comparator can be an existing treatment (*active control*), absence of the intervention or a *placebo*.

Placebo is a substance or treatment with any therapeutic value. Common placebos include inactive tablets, inactive injections and other procedures.

Outcome

Outcome is the variable that is evaluated during a study to provide information about the impact that a given intervention has on the health of a given population[Ferreira and Patino, 2017]. The outcome employed should be methodologically well-established so that the investigators can expect the results of the trial to be widely accepted. Ideally, the outcome has to be valid and precise in order to measure the indented effect and in parallel remain unaffected by external influences. There are several classes of outcomes (qualitative or quantitative) that are used in many types of trials. The most widely used are [Piantadosi, 2017]:

- *Ordered Categories* as an outcome are used in assessment of disease severity. For example the functional severity of illness might be described as mild, moderate or severe.

- *Dichotomies* are used when assessments have only two possible values. In the study population these outcomes can be summarized as a proportion of "successes" or "failures" in a treatment effect. This outcome can be modeled using logistic regression where odds ratio are measured.

- *Event times* is a measurement of the time from the beginning of a trial to important clinical events such as deaths which are the most common outcomes for chronic disease trials. Event time measurements deals with the possibility of **censoring** because some subjects being followed on the trial may not experience the event of interest (*e.g. death*) by the observation period. As a result, it is necessary to record two data elements for each individual: the **time at risk** that can be

days, months or years and the **censoring indicator** that designates if the time at risk represents an interval to the event or to a censoring point (1 or 0 respectively). One of the most reliable and helpful measurement for time at risk is the **survival time** because it is the subject's vital status and that is the reason why is widely used as an outcome in clinical trials.

- *Surrogate outcome* is one that is measured instead or in addition to the clinically most meaningful outcome. The surrogate outcome tracks the progress of the disease and can be used as predictors or disease markers. A good surrogate outcome must be measured easily and without cost effective procedures. An example of surrogate outcome is the *prostatic specific antigen (PSA)* which is a reliable marker indicating the potential of acquiring a prostatic tumor.

Time

The time frame over which the outcome will be measured. It depends on the disease examined and the purpose of the clinical trial. In particular, trials can last from months to several years like in oncology trials, where investigators want to ensure that there will not be any unobserved side effects. For example, the BIG 1-98 trial, which compared *femara* and *tamoxifen* after surgery to treat early-stage breast cancer, was started in 1998. The women took these medicines for 5 years and then were followed for several years after treatment, so the first results were available in 2005, 7 years after the trial began [Rabaglio et al., 2009].

1.3 Key elements of experimental design

1.3.1 Control of Random Error and Bias

Carefully selecting and defining the research question at the beginning of a trial as mentioned in the previous section is a very important stage in the organization of the trial, but in order to have a successful result this question must be followed by a good experimental design. Badly designed experiments can lead to incorrect conclusions, wasted time and scientific resources. One of the most important things that characterizes a good experimental design is the control of the major sources of variation [Festing, 2003]. Variations in the clinical trial results can arise from random variability (*i.e. random error*) and from deviations from the true value (*i.e. bias*) [Lewis, 1999].

Clinical research is always in a pursuit of avoiding bias and random error with regard to ensure that they have collected the most accurate results. Two of the most important design techniques for avoiding bias in clinical trials are *randomization* and *blinding* [Lewis, 1999], while random error can be controlled and reduced to acceptably low levels through *replication* [Piantadosi, 2017].

Randomization

Randomization introduces the significant element of chance into the allocation of treatments to subjects in a clinical trial [Lewis, 1999], since one of the major sources of bias is the process by which physicians and patients make treatment decisions. In particular, patients who satisfy the eligibility criteria of the trial are randomly assigned to treatment and this randomization

guarantees that there is no systematic selection bias in treatment allocation. Both observed and unobserved baseline differences between the treatment groups are attributable to chance and so their effects can be quantified by the statistician [Piantadosi, 2017]. For all these reasons *randomized clinical trials* are often considered as the *gold standard method* for comparing treatment effects [Gonzalez et al., 2009].

Blinding

Blinding or *masking* also reduces bias with a different manner: the identity of the treatment each patient has received is unknown. In a single - blind trial, the patients in the study are unaware of which treatment they receive, while in a double - blind trial neither the subject nor any of the investigators are aware of the treatment received [Lewis, 1999]. In many drug trials when the patients know that the new drug is applied to them, there is a great chance that they will overstate its effect. So in these cases, the use of blinding can improve the objectivity and even more in double - blind trials where the outcomes will not be influenced by investigators' expectations[Piantadosi, 2017].

Replication

Replication is one of the most basic features of experimental design. It is the only way along with the increase of the sample size, that reduces random error and its magnitude by repeating the experiment until the estimators' variability is measured [Piantadosi, 2017]. In fact, the procedure is that the experiment is repeated multiple times and an average value of the estimator is calculated.

1.3.2 Sample Size, Power and Hypothesis Testing

In comparative clinical trials the primary goal is to show that the experimental drug or treatment is statistically significant compared to a control group. In addition to this goal it is important the trial to have a high probability of detecting a clinically meaningful difference [Chow and Liu, 2008]. This probability is known as the *statistical power* of the trial. In detail, the approach of determining the power of a trial is based on a planned hypothesis test, where the null hypothesis usually implies similarity between the treatments. The alternative hypothesis is satisfied if difference of clinical importance is existed between the treatments.

In the process of making inferences from hypothesis tests, common random errors known as type I and II are affecting the study:

- Type I error known as α, is the probability of rejecting the null hypothesis when in fact the null hypothesis is true. This error can be controlled by choosing the *level of significance a* that rejects the null hypothesis and through this test the results of the study will be labeled as statistically significant or not [Green, 2000].

- Type II error known as β, is the probability of accepting the null hypothesis when in fact the null hypothesis is false. This kind of error is more difficult to control because it depends on the sample size of the experiment [Piantadosi, 2017]. A bigger sample size implies more precise estimates and this is the key of controlling type II errors.

After the random errors occurring from hypothesis testing were explained, it is important to underline the connection between type II error and power

starting with the following equation:

$$power = 1 - \beta \tag{1.1}$$

From equation 1.1 arises that the power of the study is the probability of not making a Type II error or in other words to detect the treatment effect. For a certain significance level (normally $a = 0.05$), the statistical power can be augmented by increasing the sample size [Chow and Liu, 2008].

To sum up, sample size is affecting a sequence of important facts in a clinical study and for this reason choosing the sample size is one of the most important decisions an investigator has to make during the early stages of study design. It is reasonable to be large enough to provide a reliable answer to the questions addressed and the study to be properly powered [Lewis, 1999]. Otherwise, an underpowered study could result in waste of money, misinterpreted results and patient discomfort [Giuffrida, 2016].

1.3.3 Validation of the study

In the process of designing a clinical trial, accuracy and precision are two very important features that should prevail in all clinical investigations. *Model validation* is the key of confirming that the outputs of a statistical model are valid and that the objectives of the investigation can be achieved. Along with the validity of the results, it is equally important to show also that on repeated measurements the results can be measured precisely [Sanchez and Binkowitz, 1999].

The majority of the models used in clinical research have prognostic nature. In fact, prognostic models are applied in order to assist investigators

with decisions regarding treatment choices or for example to gather information about patients' survival. As a result, it even more crucial to validate research models involving humans, before application in clinical practice. In that way, their generalizability will be judged and so the study findings can be incorporated in other populations or not [Dijkland et al., 2018].

There are several ways to establish how well a prognostic model performs for further patients, but two of the main ways reported by *Douglas Altman et al.* in his article "Prognosis and prognostic research:validating a prognostic model" [Altman et al., 2009] are:

1. ***Internal validation*** where a model's performance is assessed using new data from the same source as the derivation sample. A common approach is to split the dataset into two parts, develop the model using the first portion (i.e. training set), and assess its predictive accuracy on the second portion. The dataset can be splitted in the two sets equally, by choosing a proportion and by techniques of data re-use, such as cross validation and bootstrapping. Internal validation is very helpful, but it cannot provide information about the model's performance elsewhere.

2. ***External validation*** where a model's performance is assessed using new data from elsewhere, not from the simulated data used in the model development. This strategy is the only way to examine the generalizability of the model.

All in all, the goal of this section was to describe the key elements of experimental designs. Controlling random error and bias, choosing the sample size of the study, securing high probability of detecting a treatment effect

and at last validating the study results are the most crucial considerations in the process of designing a clinical study.

1.4 Clinical Cancer Trial Designs

1.4.1 Enrichment Designs

In recent years, there is a remarkable progress in the areas of tumor biology combining with genomics technology. More and more medicines are being discovered and this phenomenon have motivated great advances in clinical cancer research. This evolution happened because, tumors have been found to be biologically heterogeneous with regard to their causal mutations. As a result, molecularly targeted drugs are not always capable to benefit patients of broad diagnostic categories in clinical trials [Simon and Simon, 2013a]. For this cause, cancer researchers are innovating the standard clinical trial paradigm of treatment development. Their goal is to accelerate the time it takes to test whether a treatment is efficacious and also to identify patient subpopulations that will benefit the most from the new treatment. This pioneering idea created a huge amount of *Enrichment designs* widely used in various therapeutic areas.

Enrichment is the prospective use of any patient characteristic, either historical, genetic or other with a view to select a study population in which detection of a drug effect is more likely. Most of the times these characteristics are binary or continuous biomarkers like for instance the absence of a protein or a cholesterol indicator respectively. Once a predictive biomarker is identified, it is used to restrict the entry of patients to the trial comparing

the new drug with a suitable control [Simon and Simon, 2013a].

Maitournam and Simon wrote in their article "On the efficiency of targeted clinical trials"[Maitournam and Simon, 2005] that "Such enrichment designs can serve to magnify the treatment effect and thereby improve the efficiency of the clinical trial."

Enrichment designs usually consist of two or three stages, as seen in the Figure 1.1 [Miranta Antoniou,]. The first stage is the screening process, where based on the predictive biomarker, a certain subpopulation is selected. The patients that can not meet the objectives of the trial are excluded from the study. In the followings stages, the patients belonging in the enriched subpopulation are prospectively randomized between the two treatment groups(i.e the new drug and control) and the efficacy of the new drug is examined. At the final stage, statistical analysis makes use of the study outcomes, in order to reach to a conclusion for the selected population [Fedorov and Liu, 2014].

1.4.2 Master Protocol designs

The enrichment design has been used extensively over the past 15 years for phase III registration studies. As mentioned above, this design identifies a sensitive subgroup that is more likely to present positive outcome given the new treatment and then test its efficacy on this subpopulation. Patients whose tumors do not contain the genomic alteration (e.g. biomarker) which would render them possibly sensitive to the treatment are excluded from the randomization [Simon, 2017]. With this design, targeted treatments are being established for certain tumor types based on genetic mutations. Imagine

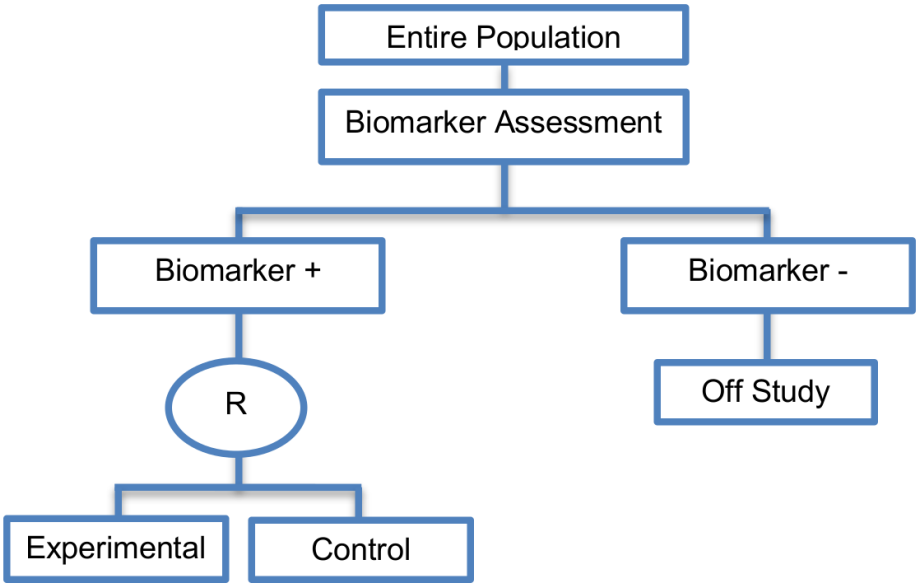

Figure 1.1: Enrichment Design (*BiGTeD, Antoniou et al.)*

how many different types of tumors have been discovered, along with their molecular subtypes. It is not realistic to conduct phase I–III trials for each subpopulation, when funding and time for cancer research has been limited [Hirakawa et al., 2018].

This phenomenon has created a shift towards development of biomarker targeted agents and in recent years new approaches have arisen in clinical cancer research driven from the need of studying many agent and target combinations in parallel [Renfro and Sargent, 2016]. In order to assess the combination of molecular markers and their targeted treatments for single or multiple tumor types, common protocols are required. These protocols are called "*master protocols*" and are considered as the next generation in clinical trial design [Hirakawa et al., 2018]. *Umbrella* and *basket* trials are

the most known master protocol trials.

Umbrella trials

Umbrella trials are focusing their attention on a single tumor type, but evaluates multiple targeted therapies for different molecular marker. As it is described in the Figure 1.2 [Hirakawa et al., 2018], there are different sub-studies for each marker, where the patients are assigned regarding their targeted biomarkers. These sub-studies can be single arm, phase II, or phase II/III trials that are randomized and compared to placebo or a standard therapy [Hirakawa et al., 2018]. An example is the ALHEMIST trial, which uses the umbrella design to identify and screen patients with the *EGFR* and *ALK* mutations in early-stage non-small cell lung cancer [Govindan et al., 2015].

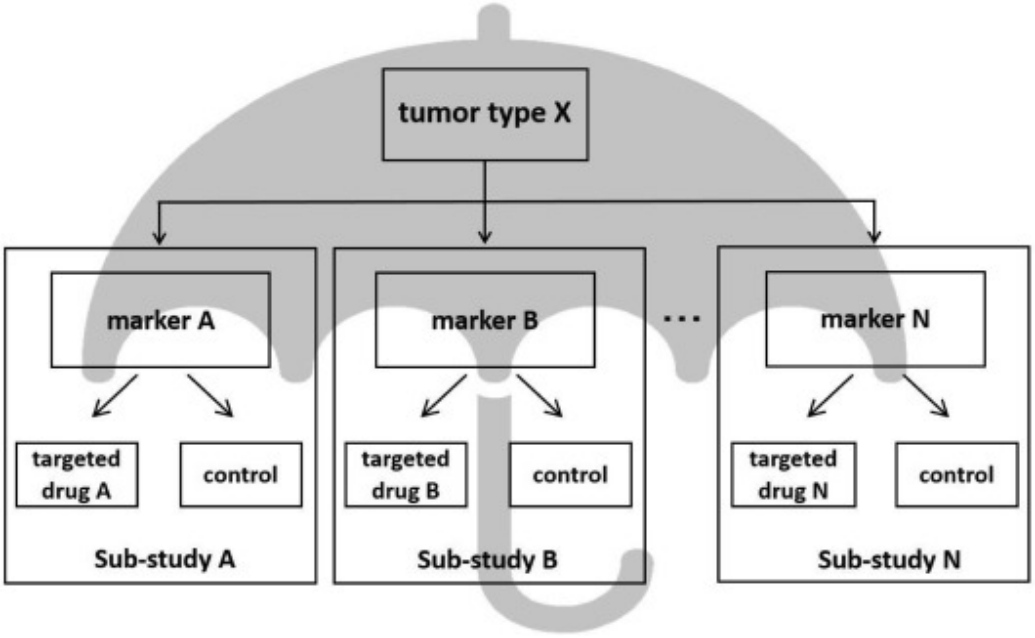

Figure 1.2: Umbrella trial (*Akihiro et al. 2018*)

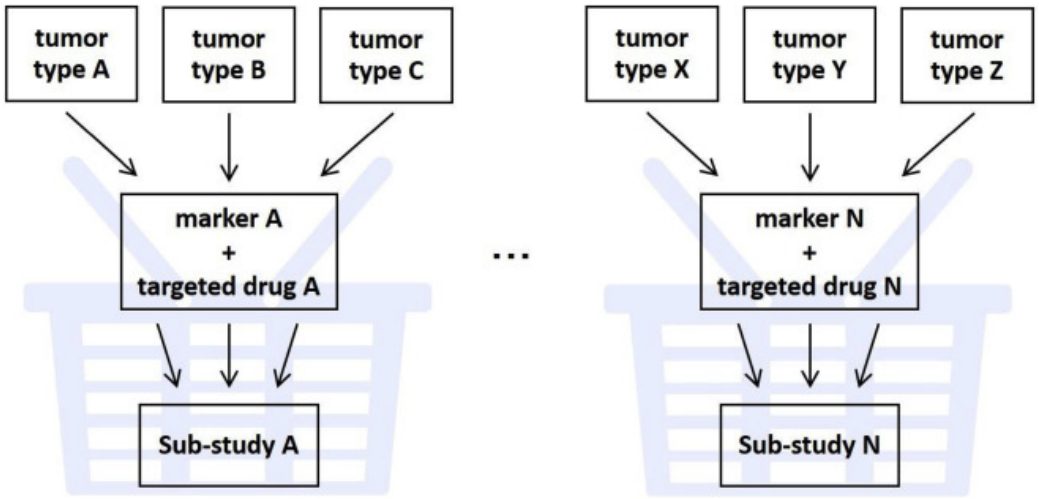

Figure 1.3: Basket trial (*Akihiro et al. 2018*)

Basket trials

Basket trials evaluate one targeted therapy on multiple diseases or multiple disease subtypes. This gives the potential of including in the same trial multiple subpopulations with a common singular molecular marker, but across different tumor types. In this scenario as shown in the Figure 1.3, the grouped tumor types create a basket, and sub-studies are conducted by tumor groups within it. Basket trials are often conducted as single-arm or phase II trials. The absence of a control group is a limitation because the therapeutic effect cannot be evaluated. NCI-MATCH is a phase 2 clinical trial following a basket design. This trial seeks to determine whether targeted therapies for people whose tumors have certain gene mutations will be effective regardless of their cancer type. Researchers use a DNA sequencing test to identify gene mutations in patients' tumors. The test looks for mutations in 143 genes as-

sociated with cancer that can be targeted by one of the drugs being studied in the trial.

1.5 Description of the current study

Due to the heterogeneity of most human cancers, only a subset of patients can get a positive impact from a given treatment. There are lots of technologies, such as microarrays [Rosenwald et al., 2002] that contribute in a great extent to the identification of a genetic signature for patients, who are most likely to benefit from the treatment. Genetics can account for 20 to 95 percent of variability in drug disposition [Kalow et al., 1998] and that proves the strong connection between a possible treatment and the human genes. Finding genes or more specific a signature that modify treatment response has the potential to significantly improve the performance of clinical trials [He and Allen, 2010] and prospectively ameliorate people's life. This signature, most of the times, is not available in the beginning of a trial because of the huge number of genes needing analysis. Solution to this problem came from the clinical research scientists that recently begin to establish the biomarker adaptive designs [2]. The main goal of these designs is the identification of patients sensitivity to a treatment based on *prognostic covariates* (e.g. biomarker indicators or genes expressions) and their *interaction* with the treatment. Then, these biomarkers or gene signatures are used to classify the sensitive patients into a target subpopulation. After identifying the most suitable subpopulation, statistical tests are conducted in order to evaluate the treatment effectiveness.

[2] see section 2 for a review of adaptive designs

A review of the most recent adaptive designs is in the second chapter of the study. Following this review, in the third chapter, we will propose an adaptive design that identifies gene signatures. Its performance in terms of statistical power is evaluated by alternating several important parameters, such as the sample size. All the methods used, the simulation study and results are available in the sections 3.3 and 3.4. In the final chapter, we do a discussion over the results of the study and we give some suggestions for further research in our design.


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Chapter 2

A Review of Biomarker Adaptive Designs

Over the last years, a big variety of papers have been written introducing different perspectives for biomarker adaptive designs. Freidlin and Simon were the first that introduced the Adaptive Signature Design (ASD) [Freidlin and Simon, 2005]. The following designs tried to enhance the ASD by changing their classification rule and the outcome variable. Some other significant features considered in the proposed adaptive designs, with a view to ameliorate the results of ASD, was the sample size, the power, the level of significance and the treatment allocation. This chapter will be an extended review of the most recent adaptive designs proposed including biomarkers. Several designs with binary or continuous outcome, their improvements or alterations are described by yearly order below. Starting from 2005, *Freidlin and Simon* [Freidlin and Simon, 2005] established a new adaptive design for randomized clinical trials of molecularly targeted agents in circumstances

where a biomarker that identifies sensitive patients is not available. The reason for doing so, is that the tumor biology has changed in a way that most tumors are heterogeneous and only a subset of patients with a particular cancer has the potential to admire the benefits from a treatment. Their approach is proposed for **binary** outcome (response) and include three basic components:

1. Development of a gene expression-based classifier in order to ensure a valid identification of the sensitive subsets of patients.

2. Test of overall treatment effect for all randomized patients with 4% level of significance, instead of 5% used in the traditional approach.

3. Test of treatment effect only for the sensitive subgroup with 1% level of significance.

More specifically, Freidlin and Simon choose to randomly assign patients to the combination of the new and standard treatment (arm E) or to the standard treatment alone (arm C) with both arms to include equal number of patients. They assumed that among L genes, there is a subset of K "sensitivity" genes and the responsiveness to the treatment is influenced by these K genes through the following logistic model.

For the ith patient, p_i is the probability of response to the treatment, t_i is the treatment the ith patient received (0 for arm C and 1 for arm E) and $x_{i1}, \ldots, x_{iK}$ are the levels of expression for the K unknown genes.

$$\log\left(\frac{p_i}{1-p_i}\right) = \mu + \lambda t_i + \gamma_1 t_i x_{i1} + \ldots + \gamma_K t_i x_{iK} \tag{1}$$

In the model, λ is the *treatment effect* and γ is *interaction* of treatment and expression genes. It is very important to understand that this model defines the patients probabilities of response and plays a vital role in the identification of the sensitive subset.

Further information about ASD is that the total number N of patients were divided in two stages equally. In the first stage, stage 1 patients where used to find the K sensitive genes, while in the second stage with the K genes identified, stage 2 patients were able to be classified as sensitive or not.

The design used multivariate normal distribution for the data simulation and the study was conducted on 400 patients, 10000 genes with a number of 3, 10 or 20 sensitive genes. Results shown that:

- The performance of the ASD relative to the more traditional design shown that when the number of sensitive patients is low, the ASD reduces the chance of false rejection of the new treatment.
- In the case that the new treatment is generally effective the ASD has similar power to detect the overall effect with the traditional design.
- When the difference between arms is moderate to low, the total sample size required to develop the selection procedure have to be much larger.

After the extensive work of *Freidlin* and *Simon* with the ASD design, two years later in 2008, the same authors along with *Jiang* proposed a different biomarker adaptive threshold design. This time, the biomarker that identify the sensitive patients, was measured on a **continuous** scale and has a threshold value as a cutpoint to distinguish the sensitive subgroup. The design,

based on hypothesis tests, provides information about the new treatment and its efficacy on the subpopulation defined by the biomarker. Furthermore, in this study the patients are randomized in two treatment arms, one with the experimental treatment and one with standard care. These arms are compared with respect to time(t) to a clinical event and modeled using **proportional hazard model** (see section 3.2.2). The model is written as

$$\log \frac{b_E(t)}{b_C(t)} = \begin{cases} 0 & \text{for biomarker values below } c_0 \\ \gamma & \text{for biomarker values above } c_0 \end{cases} \tag{2}$$

where c_0 is the cutoff value. The model assumes that above this value the patients benefit from the treatment, while below this value do not. This design have a lot of things in common with the ASD design, as the same strategy with the two hypothesis test is conducted. One test in all patients with level of significance α_1 and a second test for the sensitive subset with level of significance α_2.

The study is divided into procedure A and procedure B:

1. Procedure A is in fact the idea of separating the test of treatment effect in the broad population from the subset selection. Thus, at the end of the study, one of the three objectives is fulfilled: *treatment benefit in broad population* or *treatment benefit in sensitive subgroup* or *no treatment effect*.

2. Procedure B is a generalization of procedure A in which the overall and subset tests are combined by incorporating the correlation structure of the test statistics. For each biomarker cutoff value, model 2 is fitted for

the subset of patients exceeding that certain value and a log likelihood ratio statistic is calculated.

For the simulation biomarker values were generated from uniform distribution. Moreover there were 7 scenarios applied: one with all patients benefited from the treatment, four with patients that exceed a certain cutoff value and two with linear increase in hazard ratio. Results shown that:

- The highest power for the overall effect was in the scenario where the new therapy is beneficial for all patients and it decreases as the proportion of sensitive patients decreases.
- Under the linear and delayed trend model (last two scenarios), the power was better than the standard design.
- In general procedure B preserved higher power that procedure A.

Hoering et al. in 2008 [Hoering et al., 2008] introduced once more the complexity of cancer therapies and the difficulty to distinguish which patients truly benefit from them. As a matter of fact, their article highlights an issue regarding the identification of sensitive or marker-positive patients. It is mentioned that there are cases where targeted agents (medicine or treatment) benefit patients who were not considered sensitive to this treatment in the first place. For example, *imatinib* seems to destroy gastrointestinal tumors but it was developed to target chronic myelogenous leukemia. Also there are extremely toxic agents designed to kill tumors cells but with the cost of damaging other organs. These examples show that there is a chance patients classified as sensitive to a treatment, do not have beneficial outcome

while non-sensitive patients to show improvement. For this reason there are different designs that examine the behavior of marker-positive (M+) and marker-negative (M+) patients regarding their response to the treatment. In this paper three designs for targeted agents are mentioned:

1. In **Randomized-all design**, the marker status of the patients is evaluated and all patients are randomized to one of the two treatments. This a good design to examine the responsiveness of the treatment for both M+ and M- subsets.

2. In **Targeted trial design** after the marker status of the patients is known, only M+ patients are enrolled in the trial and randomized to the two treatments. This design proves to be good if the biology of the disease is well understood and it is sure that only M+ patients can benefit.

3. **Strategy design** suggested by *Hayes*[Barlow and Hayes, 1979], where patients are randomized in the treatments based on their marker status (M+ patients new therapy, M- patients standard one) and in every patient independent of their status take standard of care.

These designs test different hypothesis as *randomized-all design* examines the possibility that the new treatment is effective in all patients, the *targeted trial design* test this possibility only for marker-positive patients and the *strategy design* addresses the question of whether the marker based treatment is better that the standard of care. So in this article the main purpose is to evaluate the effectiveness of the previous designs based on some scenarios. First in the absence of a valid marker, following in the presence of a *prognostic*

marker[1] and also several scenarios for *predictive marker*[2]. Possible scenarios for predictive markers is for example a case that the new treatment is valid only for M+ patients or a case that the new treatment is beneficial for M+ patients and harmful for M- patients.

The analysis of study focuses on binary outcomes, the distribution of the biomarker is continuous and a cutpoint is used to distinguish patients with marker status above or below a threshold. The performance for the designs in the various scenarios are evaluated as a function of the cut point, the sample size and number of patients in order to obtain a certain power. Results showed that:

- Moving the cutpoint did not had an effect on power in the randomized-all design, but had a large impact on the other two designs because the cutpoint defines the classification of the subjects.

- Targeted design performs the best in scenarios with a predictive marker, in contrast with the other designs where the benefit of M+ patients is reduced.

- The randomized-all design performs as well or better than strategy design except for the scenario that M+ patients benefit and the treatment in M- patients is harmful. That is the scenario when the overall treatment effect is null.

- The strategy design seems to be inefficient in comparing the difference

[1] a marker that gives information about a likely disease outcome independent of a treatment

[2] a marker that gives information about a disease outcome based on treatment

of two treatments as patients in different arms are treated with same therapy.

After considering various scenarios with different designs the article proposes to use targeted design only if it is known that the treatment is not beneficial for all patients to some degree and that the cutpoint is well established. Also, it is better to use randomized-all design if there is a suspicion that except M+ patients, M- patients are also benefited.

Following the work of *Freidlin and Simon*, who introduced the adaptive signature design, in this paper the same authors along with Wenyu Jiang proposed a *cross-validated* extension of ASD in 2010 [Freidlin et al., 2010].

Their consideration was based in the fact that in ASD only half proportion of the patients was used to evaluate genes and continuously identify them as sensitives. In fact, this strategy can not have the best results as more often signature development in high dimensional data requires big sample sizes. Furthermore, when the number of sensitive patients is low, larger number of patients is required to achieve proper power to detect the overall effect. For these reasons, they thought that it is better to use the entire population in the signature development and validation steps.

The cross-validated adaptive signature design (CVASD) used K-fold cross-validation approach for the signature development and subset effect testing. First, the trial population is divided into an M-patient validation cohort ($M = N/K$) and (N-M)-patient development cohort. As a result, K different patients cohorts are created. For the k-set of patients, D_k is denoted the patients in development stage and V_k the ones in validation stage. Contin-

uously, for each D_k a predictive signature is developed and it is applied to the V_k to identify the sensitive patient subset S_k. The procedure is repeated K times and at the end of the study the sensitive population identified as $S = \cup_{k=1}^{K} S_k$. The outcomes for the sensitive patients received the experimental therapy can be compared with the outcomes from the ones received the standard therapy. Moreover, in CVASD, the p-value is calculated by a permutation method due to the fact that the standard asymptotic theory is not applied when the subset is obtained by cross-validation. Results in this study showed that:

- Cross-validation can preserve the type-I error while increasing the power to detect a significant treatment effect in a subset of patients that benefits from the experimental treatment.
- CVASD is better than ASD because when the treatment effect is low the power of detection in CVASD is quite bigger that the one in ASD.
- For higher fractions of sensitive patients CVASD, ASD and the traditional design tend to have similar results.
- When all patients benefit equally from the treatment both CVASD and ASD correctly indicate the absence of subpopulation.

The importance of identifying a subgroup of patients that benefit from a new treatment is also discussed in the article of *Cambon et al. 2015* [Cambon et al., 2015a]. Also, the purpose of this study was to examine non-parametric classification methods in order to further understand and best

use them for treatment subset prediction. The article begins explaining fundamentals features of classifications methods. To start with, it is mentioned that they are divided into two categories, supervised and unsupervised learning. It is not difficult distinguish the difference between them as the outcome variable is not present in unsupervised (only features are observed), while in supervised the outcome variable guide the whole process. Another fact mentioned is that in classification methods the covariates (features) play a very important role and can decide which method will be used. For example, in high-throughput data (HTD) like genomic data that there are hundreds or more features and always features exceed sample size ($p >> n$). These data need dimension reduction (DR) methods and for that reason, methods like Random Forests, Support Vector Machines (SVM), Boosting have built-in DR techniques.

The classification model is built using data on a training set and then this rule is applied in the validation set in order to avoid bias of prediction error. For that reason, well organized internal validation is need. Data splitting is a major issue in classification and contributes to minimize the prediction error. For example, Adaptive Signature Design incorporates this process by optimizing a set of tuning parameters using nested cross validation. Bootstrap (sampling with replacement) can also be used in classification methods to avoid prediction error as well. External validation challenges also the adaptive designs because of the high dimensionality of the genomic data. Until 2000s HDT involved mRNA expressions but it was shown that other expressions like proteins and miRNA's play more important role in cancer. Next Generation Sequencing (NGS) platforms are replacing traditional microarray

platforms and that is the reason why classification methods that are robust to changes have to be generalized.

Further in the article there is a section describing nonparametric dimension reduction in classification. The article defines dimension reduction methods as a feature selection or feature extraction (to distinguish the most promising features). Along with DR methods, there is a big section analyzing most of the non-parametric methods including K-nearest neighbor (KNN), Kernel density analysis (KDA), Support Vector Machines and a lot more. The advantages, disadvantages and the situation when is most suitable to use one of these methods are highlighted too.

At the end, a great deal of importance is given in the adaptive signature design of Freidlin and Simon. In fact, it is mentioned that ASD uses weighted voting and single gene logistic models with treatment-gene interaction. A tuning parameter η is used to select predictive genes, R to determine the cutoff for the magnitude of odds ratio(OR) and H to determine the number of selected genes. In this specific design the treatment arm OR behaves actually as a distance measure to compare response to the treatment.

This paper gives three suggestions for further evaluation in the ASD design and those are to replace the classification rule with new distance measures employed in:

1. Kernel Density Analysis (KDA)
2. K nearest neighbor (KNN)
3. Support Vector Machines (SVM)

At last the article proposes to calculate an average score for all genes instead of weighted voting method with a view to eliminate the need for tuning parameter H.

Cambon et al. continued their work by focusing on how to make a classification decision for patient population using parametric distance measures [Cambon et al., 2015b]. The beginning of the article is dedicated to the scientists contribute the most in the evaluation of discriminant analysis. Fisher, Welch, Neyman and Pearson are some of the scientists that outlined very important methods or elements in discriminant analysis. Fisher for the Fisher's Discriminant Analysis (FDA), Welch for Linear Discriminant Analysis (LDA) and Neyman - Pearson for introducing likelihood ratio test (LRT). All these methods mentioned where developed before high throughput data (HTD) became prevalent in the clinical research. As already mentioned in previous paper, many dimension reduction (DR) methods were introduced to address the situation. In fact, these early parametric methods where used in conjunction with DR techniques and as a consequence, modified versions of these methods where developed. Following in the paper these new methods are briefly analyzed and examined focusing on treatment subset prediction. In particular, the dimension reduction methods mentioned are Diagonal Linear Discriminant Analysis (DLDA), Principal Component Analysis (PCA), Feature selection (FT) methods, Partial Least Squares (PLS), L1 and L2 regularization analysis.

Apart from the dimension reduction methods, the most important parametric classification methods and their statistical background are outlined in the next section of the article including FDA, LRT, Linear regression and

more.

After considering a lot of parametric classification methods the article give recommendations of other ways to incorporate the distance measure in the adaptive signature design of Freidlin and Simon. ASD uses the treatment arm OR as a distance in weighted voting method to determine sensitivity to treatment. Methods like LDA and LCA are also constructed using distance metrics. For example, log likelihood ratio (LLR) measures the distance of a subject from the border of two classes. As a result, sensitive subjects can be predicted using methods that incorporate this distance. The equation for the difference in the LLR between treatment arms for gene k and subject i is:

$$\widehat{DLLR}_{ik} = \widehat{LLR}_{ik,t=1} - \widehat{LLR}_{ik,t=0} = \log \frac{\hat{\pi}_{11} f(z_{ik}|\hat{\theta}_{11k})}{\hat{\pi}_{01} f(z_{ik}|\hat{\theta}_{01k})} - \log \frac{\hat{\pi}_{10} f(z_{ik}|\hat{\theta}_{10k})}{\hat{\pi}_{00} f(z_{ik}|\hat{\theta}_{00k})}$$

where $f(z_{ik}|\hat{\theta}_{gtk})$ is the density for gene g with estimated parameter vector $\hat{\theta}_{gtk}$ and evaluated at z_{ik}. Distances such that can be used in ASD incorporating similar tuning parameters to η, R and H.

Instead of $\widehat{OR}_{ik}$ in an ASD, also single posterior odds ratios $\widehat{OR}_{ik\ post}$ can be used in the same way and it may also be a more intuitive. Furthermore a new distance measure can be created by the LDA mentioned before and the FDA using the Mahalanobis distance:

$$DLLR_i = \log \frac{\hat{\pi}_{11} exp\Big\{ -\frac{1}{2} m_{i11} \Big\}}{\hat{\pi}_{01} exp\Big\{ -\frac{1}{2} m_{i01} \Big\}} - \log \frac{\hat{\pi}_{10} exp\Big\{ -\frac{1}{2} m_{i10} \Big\}}{\hat{\pi}_{00} exp\Big\{ -\frac{1}{2} m_{i00} \Big\}}$$

where m_{igt} is the square of Mahalanobis distance for subject i, using parameters for class g and treatment arm t. This method can have extension to DLDA, weighted voting and posterior OR's methods.

In 2017, *Zhiwei Zhang et al.* [Zhang et al., 2017] proposed new procedures for subgroup selection and treatment effect estimation under an adaptive signature design, as most of the times this subgroup population is unavailable. Besides dealing with subgroup selection and estimation, this study also tested possible problems in previous ASDs. In fact, Zhiwei Zhang considered finding the optimal subpopulation on the basis of an arbitrary set of baseline covariates, with a view to maximize the power of detecting a positive treatment effect.

First, a simple characterization of the optimal subgroup that maximizes the power of the design is provided based on a specified utility function. For binary outcome this characterization takes the form of a half-space in terms of covariate response rates in both treatment arms and utility if specified. The procedure consists of three steps:

Step 1: estimate the covariate-specific response in each treatment group.

Step 2: estimate the expected gain for each candidate half-space defined by a vector of coefficients and estimates from Step 1.

Step 3: choose the half-space with the largest estimate of the expected gain.

Moreover, this paper restricts its attention to adaptive signature designs but performing only one test for treatment efficacy for the selected subpopulation. The procedure of estimating the treatment effect is based on cross validation approach in order to reduce the resubstitution bias. Another positive characteristic of this approach is that it is not vulnerable to any collinearity in the baseline covariates. In conclusion, the augmented inverse probability weighting to extract information from the covariates and

improve precision is also a very advantageous point in contrast with the more traditional approaches in adaptive signature design. This approach can incorporate variable selection to choose between a large number of covariates(gene expression data for example like in ASD).

To further continue his work, *Cambon et al.* in 2017 [Cambon et al., 2017] comes with a new article, in which they calculate the empirical power of the parametric and non parametric classification methods mentioned in Cambon et al. (2015a) and Cambon et al. (2015b) in addition with a method similar to the ASD of Freidlin & Simon (2005). All these methods were evaluated under different simulation scenarios which include the presence of:

1. Expression-treatment interaction only.
2. Expression main effect and expression-treatment interaction.
3. Expression and treatment main effects together with interaction.
4. Both equal and unequal gene expression variance between sensitive and nonsensitive patients.

Also the methods used to make predictions under the four mentioned scenarios were evaluated in the case that the gene expression main effects are included (TG models) and excluded (T models).

The article consists of a section describing in detail the background of adaptive signature design. Something important mentioned is that a test of proportions such Pearson chi-squared can be used to compare the two treatments arms with all patients and Fisher's exact for comparing the arms in

the subgroup of sensitive patients. Further in the study, the article summarize all the proposed classification methods in the two papers of Cambon mentioned. All these methods have in common that they use densities to estimate gene expression densities for each subject. Linear, quadratic or kernel density analysis (LDA, QDA, KDA respectively) are the different ways to calculate the "plug-in" densities. As the logistic regression model does not include gene expression main effect in order to facilitate the comparison with the other methods, the posterior OR's for LDA, QDA and KDA are derived without the presence of gene expression main effect. Results shown that:

- Models with fewer terms (T models) outperformed the more complex ones (TG models) even under scenarios with large gene expression main effects.
- The QDA method in scenarios without gene expression main effect (QDA_T) often had the highest empirical power, while KDA had the worst performance.

Gu Mi posted an article in 2017 with the title "Enhancement of the adaptive signature design for learning and confirming in a single pivotal trial" [Mi, 2017]. In his work, he discussed some practical aspects of 2-stage designs. Firstly, how many patients should be allocated to learn versus confirm stage, how to allocate patients in each stage and what is the impact of different splits of α. All these elements have to be deeply considered in a clinical study as they may affect the power in a great extent. The most important characteristics of Gu Mi's study are:

- All patients were randomized in the two treatment arms, treatment or placebo. In every simulation the proportion of sensitive patients was about 40%, indicating the M+ subgroup. The patients that received no benefit, were referred as M- subgroup. The marker subgroups were defined by a step function with cutoff value 0.40.

- There were 3 scenarios considered. A **moderate** with small difference in hazard ratio (HR) between M+ and M- subgroups, a **strong** with bigger difference and a **strongest**.

- Two different cases of biomarkers were evaluated. One with 3 biomarkers and one with 10 biomarkers. At last, only one biomarker was true predictive and was used to identify sensitive patients.

- The maximum follow up was 547.5 days and the censoring rate was 20%. So 20% of the patients enrolled in the study have not experienced an event because they left the study or the study ended.

- Cox proportional hazard model (see section 3.2.2) was fitted including the treatment, the biomarker and their interaction, as the outcome was survival time. The model was used to determine the biomarker in the learn stage, identify sensitive subgroup and further to test the hypothesis based on both populations.

Results of this extensive study shown that:

- The learm/confirm allocation of 30/70 or 40/60 (%) offers the greatest power advantage.

- Equal split of α (i.e. 0.025/0.025) was better in strongest scenario in contrast with the other two scenarios.
- Power decreases as the number of biomarkers increase, but in general there was a power gain in the 2-staged design over the 1-stage even with an increased number of biomarkers.
- 2-stage powers dominate 1-stage power in the strongest scenario.

A new proposal to address some issues in subgroup selection with survival endpoint, comes from *Yu-Chuan Chen et al.* in 2018 with their work "Development of predictive signatures for treatment selection in precision medicine with survival outcomes" [Chen et al., 2018]. In their study they examine two issues regarding a score conversion model. This is a model originally proposed by *Matsui et al* [Matsui et al., 2012] that actually converts a set of biomarkers for each patient, into a univariate score and using the median of the univariate scores to divide the patients into biomarker-positive and biomarker-negative subgroups. This procedure may lead to bias in patient subgroup identification regarding two issues:

1. The treatment is equally effective for all patients and/or there is no subgroup difference.
2. The median value of the univariate scores as a cutoff may be inappropriate if the sizes of the 2 subgroups differ in a great extent.

To adress those issues *Yu-Chuan Chen et al.* divide the subgroup selection into 2 steps:

1. The **first step** involves a scoring model to convert a patient's biomarkers into a univariate score. To bypass the issue of many predictive variables, they use multiple univariate models in which, each model is fitted with a single candidate predictive biomarker. For each predictive biomarker, a standardized test statistic z is calculated from the interaction term of the model. For each patient, the composite score s_i is calculated by

$$s_i = \sum_{j \in U} z_i x_{ij},$$

where U is the set of significant covariates and z_i is the standardized test statistic from the interaction coefficient of the j-th covariate.

2. The **second step** is to find a threshold cutoff to divide patients into sensitive and non-sensitive. Usually the median value is used but because of the second issue mentioned, in this paper it is proposed to use change point method (see [Jandhyala et al., 2013, Diao et al., 2018] for further information) to find the threshold cutoff. The change-point method can be applied to the voting-based classifier (VBC) method proposed by Freidlin and Simon and thus the need of specifying the tuning parameters R and G will be eliminated [Freidlin and Simon, 2005]. For a given biomarker j, the score for the i-th patient with x_{ij} is calculated as $s_i = \lambda_j x_{ij}$, where λ_j is the interaction coefficient. Then, a change point threshold is estimated for the set of patients scores and

the patients are classified as sensitive or not based on this threshold for each biomarker. After this procedure, they apply likelihood ratio test (LRT) for testing homogeneity among the patient population. The bootstrapping method is applied to obtain the P-value for the test. If the test is significant, they proceed to the next step of subgroup selection. Otherwise, they stop the procedure and claim that the sampled patients are homogeneous.

As a result, by applying the likelihood ratio test (LRT), they accomplish to examine the homogeneity of the sampled patients. Furthermore, in the context of identification of the subgroup in adaptive design and in the improvement of adaptive power, they suggested that subgroup selection is carried out if the LRT is significant. As far as the second issue is concerned, they utilize a likelihood-based change-point algorithm to find an optimal cutoff. The simulation study showed that type I error generally is controlled, while the overall adaptive power to detect treatment effects sacrifices approximately 4.5% by performing the LRT. Last but not least, it was shown that the change-point algorithm outperforms the median cutoff considerably when the subgroup sizes differ in a great extent.

The last article to include in the review of the adaptive designs is an article written by *Diao et al.* in 2018. The study is addressed to situations when limited data is available to identify a single biomarker or to determine a cutpoint that defines a sensitive subset before the beginning of the trial. Numerous methods have been proposed like the ASD from **Freidlin** and **Simon** [Freidlin and Simon, 2005] where a cutpoint is developed, using data from stage- 1 patients only, and the classifier is not used to restrict enrollment

in stage 2 but to define a subset of sensitive patients. Recently, **Simon** and **Simon** [Simon and Simon, 2013b] established a new design (AED) that enrolls all patients from the beginning and gradually restricts entry in an adaptive way. As a matter of fact, this design essentially reduces to a non-adaptive design when all patients respond to the treatment, even though the treatment effect may be different on subjects with different biomarker values. In this study, *Diao et al* wanted to enhance the performance of AED design particularly for survival endpoints and so they proposed the biomarker threshold adaptive design (BTAD). The design is divided into two stages:

1. In the first stage, they determine subgroups for one or more biomarkers such that patients in these subgroups benefit the most from the new treatment. The analysis in this stage can be based on historical or pilot studies.

2. In the second stage, they sample subjects from the subgroups determined in the first stage and randomly allocate them to the treatment or control group.

The proposed methodology is based on sampling subjects who will benefit the most from the treatment. Even if the treatment is better than the control in every biomarker group, our proposed design aims to select the group who respond to the treatment the best. Therefore, our proposed methods tend to be more powerful than the method of *Simon* and *Simon* [Simon and Simon, 2013b] that samples all subjects from only the positive biomarker group.


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Chapter 3

Adaptive Design for Detecting Gene Signatures

3.1 The idea

As already mentioned in the description of the study (see section 1.5) the goal of this study is to develop an adaptive design for detecting gene signatures. Its structure is in a great extent influenced by two remarkable studies described in the previous chapter. More specific, the design will combine characteristics from the Adaptive Signature Design (ASD) and the Enhancement of the Adaptive Signature Design (EASD) (see [Freidlin and Simon, 2005] and [Mi, 2017] respectively). In fact, the design will use gene expression levels to develop a signature that will be able to identify a sensitive subgroup like in ASD, but it will not be based on a binary outcome. The primary outcome will be time-to-event endpoints and the performance of the design will be evaluated regarding several parameters like in Gu Mi's design. Fol-

lowing are the characteristics of the designs that influenced this study, so as to better understand its elements and purpose.

Characteristics of ASD and EASD

ASD and EASD seek to fullfill two purposes that will be also primary goals in the current study:

1. Development of a signature classifier that identifies a specific subgroup of patients for whom the new treatment is beneficial.

2. Testing the overall efficiency of the new treatment in the all population and in the sensitive subgroup.

In the pursuit of achieving the previous goals, these two designs end up having a lot of similarities. Actually, Gu Mi's design maintain several elements from Freidlin and Simon's in his willingness to enhance their design. The major similarities are listed below:

- Both designs consist of two stages. In first stage, also known as "learn" stage, a proportion of patients (e.g., learn set) is used to develop the classifier, which is applied in the rest of the patients(e.g., confirm set) to identify the sensitive ones. In the second stage are performed 2 hypothesis tests for evaluating the power of the design to detect the treatment effect: one for all the population and one for the sensitive group.

- In the process of developing the classifier, both designs use the same strategy to determine which genes or biomarkers (in ASD or EASD

respectively) to include in the signature. In fact, their rule is that if the marker and treatment have **statistically significant interaction**, the marker is a possible predictive element and it is included in the signature.

- The empirical power is the benchmark for evaluating the performance in both designs and also to compare the final results.

While sharing common purposes, these designs use different methods and implementation details. The major differences of these designs are:

- ASD uses gene expression levels to create the signature that identifies patients as good or poor candidates for the new treatment, while EASD use possible biomarkers to make the identification.

- In one hand, the study of ASD is based on binary outcome and so logistic regression to accomplish its goals. On the other hand, the study of EASD is developed for time-to-event endpoints (e.g., survival) and thus the proportional hazards model (see section 3.2.2) is used.

- The study of EASD examined several parameters that could increase the power of the design like different learn-confirm proportions or different split of level of significance. In contrast, ASD without any further examination, divided equally patients in the learn and confirm set. Also, it assumed that 80-20 (%) allocation in the level of significance (e.g., $\alpha_1 = 0.04$ and $\alpha_2 = 0.01$ correspond to α level of 0.05) was capable of detecting an overall treatment effect.

3.2 Modeling Survival Data

3.2.1 Survival and Hazard function

As mentioned before, the design of this study is addressed for clinical studies with time-to-event endpoints, which in our case will be survival time. Before explaining the modeling procedure, it is important to examine in depth the connection between survival time and hazard function.

Let T be a random variable representing survival time. As a random variable, T has a *cumulative distribution function* (i.e, cdf)

$$F(t) = Pr(T \leq t)$$

and *probability density function* $f(t) = dF(t)/dt$. The survival function is expressed as the complement of cdf and so

$$S(t) = P(t) = Pr(T > t) = 1 - F(x),$$

which gives the probability of being alive just before the time t or more generally that the event of interest has not occurred at time t [Fox, 2002].

Modeling survival data usually requires the hazard function or the log hazard. The *hazard function* h(t) assesses the instantaneous risk at time t, conditional on survival to that time [Fox, 2002]:

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{Pr[(t \leq T < t + \Delta_t) | T \geq t]}{\Delta t} \\ &= \frac{f(x)}{S(t)} = \frac{f(x)}{1 - F(x)} \\ &= -\frac{d}{dt} \log S(t) \end{aligned}$$

If we integrate from 0 to t and define S(0)=1 (since the event has not occured at time 0), we can create the formula for the probability of surviving to duration t as a function of hazard :

$$S(t) = \exp^{-\int_0^t h(x)dx} \tag{3.1}$$

The simplest possible survival distribution is obtained by assuming a constant risk over time, so the hazard is h(t)=λ for all t. With this choice the survival function is formed as

$$S(t) = \exp^{-\lambda t},$$

This formula reminds of the *exponential distribution* with parameter λ. This distribution plays a central role in survival analysis as it is possible to simulate survival time with constant hazard and this is the type of data used by the Cox proportional hazards model.

3.2.2 The Cox proportional hazards model

As already mentioned, the design of this study is addressed for clinical studies with time-to-event endpoints and the Cox proportional hazard model is the most suitable way to manipulate this type of data. As a matter of fact, the Cox model is a frequently used tool that allows to do survival analysis with respect to several explanatory variables simultaneously. More specifically, it gives the opportunity to evaluate how these specific variables influence the rate of a particular event happening at a particular point in time [Cox, 1972]. For instance, like in this study, imagine that 2 groups of people are compared

regarding specific genes and their interaction with the treatment. If one of the groups have positive impact in several genes, any difference in their survival will be attributable to these genes.

The Cox model is expressed by the *hazard function* denoted by h(t). In other words, the hazard function can be explained as the risk of having the event (e.g., death or metastasis) at time t. In detail, it can be estimated as follow:

$$h(t) = h_0(t) \times \exp(b_1 x_1 + b_2 x_2 + \cdots + b_p x_p) \tag{3.2}$$

where,

- t presents the survival time
- $h(t)$ is the hazard function
- $b_1, b_2, \ldots, b_p$ are the coefficients measuring the impact of covariates $x_1, x_2, \ldots, x_p$.
- h_0 is the baseline hazard and corresponds to the value of the hazard if all the x_i are equal to zero.

Furthermore, the Cox model can be written as a multiple linear regression of the logarithm of the hazard on the variables x_i with h_0 being an *intercept* term. The quantities $\exp(b_i)$ are called *hazard ratios* (HR). The interpretation of this term is that a positive value of b_i will bring a HR greater that one and therefore the event hazard increases. Increase in hazard implies to a decrease of survival time. In terms of treatment efficacy, when the hazards

ratio are less than one, that undoubtedly means a prolonging survival for the sensitive to treatment patients.

In summary, three are the possible scenarios when dealing with hazard ratios:

- $HR = 1$: No effect
- $HR < 1$: Reduction in hazard
- $HR > 1$: Increase in hazard

As far as the data simulation is concerned, a Cox model need three different kind of inputs to be implemented. At first, the *time-to-event data* which is survival time for each subject in the trial. Secondly, the event or censor for each subject, which is a binary indicator stating if the subject experienced the event or not. Last but not least, is the predictors that are related to the purpose of the study. In our study will be the treatment and the treatment interaction with the genes.

3.3 Materials and Methods

3.3.1 Introduction to the design

The goal of this study is to develop a predictive gene signature, capable of identifying a subpopulation, which benefits most from the new treatment. To achieve this goal, following the ASD, this study will do a simulation over a phase III clinical trial that *randomly* assigns patients in two different treatments (i.e., arms). Apart from randomly assigned in the two arms,

the patients also will be equally distributed, meaning that both arms will have the same number of patients. Due to this *randomization* the results of the study will not be biased. Furthermore, the first arm (A) will receive a combination of the experimental and standard treatment, while the second arm (B) will receive just a standard treatment. The arm B is also known as control group.

As the study is focusing on building a gene signature, it will use DNA microarrays expression profiling like Freidlin and Simon did. In detail, among L evaluated genes, there is a subset of K *sensitivity* genes that tend to ameliorate the performance of the new treatment[Freidlin and Simon, 2005].

The development of the predictive classifier consists of three components: the gene identification, classifier development and performance assessment [Chen et al., 2014]. Also, it is important to mention that if N patients are participating in the trial, N_1 patients will be used for the gene identification (i.e., learn set)and the rest $N_2 = N - N_1$ for the classifier development and performance assessment (i.e., confirm set).

3.3.2 Gene identification

The gene identification will be based on Gu Mi's design [Mi, 2017] to identify possible biomarkers. The difference in this design is that instead of biomarkers, there are gene expressions as potential candidates for the signature [Freidlin and Simon, 2005].

For a given patient i and for the j gene, let x_{ij} denote the expression level,

and t the kind of treatment a patient is assigned

$$t = \begin{cases} 1, & \text{new and standard treatment} \\ 0, & \text{standard treatment} \end{cases}.$$

For the gene identification, a Cox proportional hazard model is fitted, including the treatment t and the gene-by-treatment interaction term $(x_{ij}t)$. Let this model be named as "M" in order to facilitate the description of the current study. In detail, for every $g \in 1, ..., L$ and only for the patients included in the learn set, the "M" is fitted. From now on, the rule that determines if the g gene will be included in the signature is whether the gene-by-treatment interaction term was statistically significant. The level of significance α is set at 0.01. The setting is that strict, with a view to reduce the chance of falsely include a *non-sensitivity* gene in the signature. Therefore, at the end of this stage a gene signature is identified and its contribution is essential for the classifier development. The number of the genes identified from the model can be more than the sensitivity genes K or less. From now on, in order to facilitate the description of the design it is better to include all these significant genes to a subgroup $\mathcal{S}$.

3.3.3 Classifier development

In the process of the classifier development, two groups are considered. A group of patients, which benefit most of the treatment and correspond positively to the predictive signature. This group from now on will be named as the "*sensitive subgroup*". Therefore, the other group of patients will be the one that corresponds negatively to the treatment. At this stage, only

the patients from the confirm set will be classified as sensitive or not. In detail, this study will use the same classification rule as the one described in the Adaptive Signature Design [Freidlin and Simon, 2005], only with a bit of alteration for survival outcome. The procedure of classification is organized in several steps as follows:

Step 1: For every significant gene $j \in \mathcal{S}$, fit the "M" model and save the coefficients from the treatment effect (b_{1j}) and interaction term (b_{2j}).

Step 2: For each patient in confirm set i and for every gene $j \in \mathcal{S}$ calculate the hazard ratio HR_{ij} using the b_{1j} and b_{2j} coefficients, with the formula

$$HR_{ij} = \exp^{b_{1j}+b_{2j}x_{ij}}$$

Step 3: Classify the i-th patient as sensitive if following condition is satisfied

$$\Big(\sum_{j\in\mathcal{S}} HR_{ij} \leq R\Big) \geq G$$

In fact, the interpretation of Stage 3 is that a patients is classified as sensitive if the predicted arm A versus arm B hazard ratio exceeds a specified threshold R for at least G of the significant genes. Therefore, it is obvious that the design requires two tuning parameters R and G that determine in a great extent if a patient will be classified as sensitive. More information on how these tuning parameters where selected is at section 3.4.1.

3.3.4 Performance assessment and subgroup analysis

After patient's classification it is very essential to ascertain whether the classifier has categorized accurately the patients in the sensitive subgroup or not.

This issue can be easily solved by calculating the accuracy of the model, by creating the confusion matrix of the actual and predicted values. In detail, the form of a confusion matrix is

		Predicted	
		Sensitive	Non-sensitive
Actual	Sensitive	a	b
	Non-sensitive	c	d

Table 3.1: The form of a confusion matrix

As a result, the values in the *diagonal* of the confusion matrix values, α and d are the number of patients correctly classified as sensitive and non-sensitive respectively. Therefore, in order to examine the accuracy of the model it is essential to calculate the following fraction

$$\text{Accuracy} = \frac{\alpha + d}{\alpha + b + c + d}$$

As far as the accuracy of the model is on average more than 90%, it is meaningful to proceed to the evaluation of the treatment efficacy in the subpopulation, which is one of the primary goals in this study. At the time of final analysis there two comparisons (i.e., hypothesis tests) that determine the results of the study

1. A test of hypothesis for the overall comparison of the new treatment with the control.

2. A test of hypothesis for the comparison of the new treatment versus control in the identified as sensitive subpopulation.

A p-value is calculated for each test, so the two p-values α_1 and α_2 have to be adjusted in a way that the overall type I error is no more than α [Chen et al., 2014]. Following this rule, there are multiple combinations for splitting α in order to preserve an overall 5% type I error. The most widely recommended allocation is the Bonferroni adjustment, which uses 2.5% significance level for each test [Aickin and Gensler, 1996]. In this study, more α allocations will be considered. The study is considered positive, if either of the two test is significant. In this way, it is sure that the new treatment is broadly effective and that it will not be withdrawn from the list of possible cancer treatments.

3.4 Simulation study

3.4.1 Data generation

To evaluate the performance of the adaptive design, simulated data were generated and so in this section is taking place all the generating procedure.

To begin with, assume $s = p \times N$ to be the number of sensitive patients generated, where p is the percentage of sensitive patients and N the sample size of the study. These s patients were determined to be the **first** patients. For instance, if the sample size of the study is 2000 and the percentage of sensitive patients is 40%, in this study the first 800 patients are sensitive. Also, in the simulation from the L genes the **last** K were generated as *sensitivity* genes.

The gene expression levels x_{ij} $i \in 1, \ldots, N$, $j \in 1, ..., L$ were generated using a multivariate normal distribution and the generation was divided in

three steps:

1. One $s \times K$ matrix with values generated by **N(0,0.25)** for the **predictive genes** in **sensitive patients** (up-right in the Figure 3.1).

2. One $(N - s) \times K$ matrix with values generated by **N(0,0.01)** for the **predictive genes** in **non-sensitive patients** (down-right in the Figure 3.1).

3. One $N \times (L - K)$ matrix with values generated by **N(0,0.1)** for the **non-predictive genes** in all patients (left in the Figure 3.1).

After connecting these matrices, the gene expression levels x_{ij} for $i \in 1, \ldots, N$ and $j \in 1, ..., L$, are formulated into a $N \times L$ matrix as figured in the Figure 3.1.

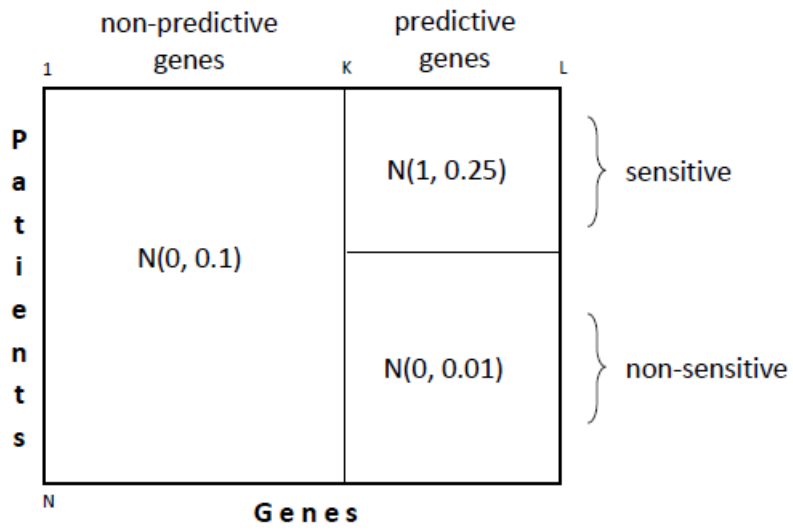

Figure 3.1: Generation of expression genes levels

In addition, the way to simulate survival times for each patient was similar with the procedure of Gu Mi's study. His study used piecewise exponential distribution as a baseline survival function with 3 segments and 2 breakpoints at 193.33 and 350.67 days. The maximum follow up per subject is 547.5 days. If the exponential rates of every segment is known the survival times can be calculated. The median survival times for the control arm are $m_1 = 439.64$, $m_2 = 203.32$, $m_3 = 154.62$ (see Figure 3.2), so from the median of piecewise exponential $m = \frac{ln(2)}{\rho}$, it is possible to calculate (see below) the exponential rates used for simulating the survival time in each segment.

$$\begin{bmatrix} m_1 = 439.64 \\ m_2 = 203.32 \\ m_3 = 154.62 \end{bmatrix} \Longrightarrow \begin{bmatrix} \rho_1 = \frac{ln(2)}{m_1} = 0.0015 \\ \rho_2 = \frac{ln(2)}{m_2} = 0.003 \\ \rho_3 = \frac{ln(2)}{m_3} = 0.004 \end{bmatrix}$$

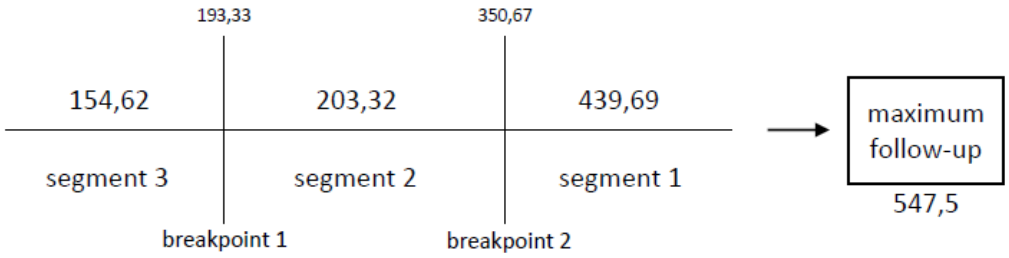

Figure 3.2: Median survival times in control group

As a result, the exponential rates used in Gu Mi's study for the control group are

$$\rho(t) = \begin{cases} 0.004, & 0 < t \leq 193.33 \\ 0.003, & 193.33 < t \leq 350.67 \\ 0.0015, & t \geq 350.67 \end{cases}$$

In the treatment group, sensitive (M+) and non-sensitive (M-) patients have different survival times and this is determined by the HR given in each group. Gu Mi examined 3 possibles scenarios for HR of treatment over control group and in the current study 2 quite different scenarios will be evaluated. The HR in the strong and weak scenario are depicted below:

Scenario	**HR in M+**	**HR in M-**
Strong	0.65	1.1
Moderate	0.65	1

Table 3.2: Hazard ratios in different predictive scenarios

In the "strong" scenario the HR of M- over control group is 1.1, meaning decrease in survival. This is in contrast to the 0.65 HR in the M+ group, that experiences positive impact from the new treatment. This scenario is the strong one because the HR of the two groups are in the limits so it is important to see if the design will detect the treatment effect.

The "moderate" scenario has HR of M- equal to 1 over control group, which means no effect at all. It is very essential to see the performance differences in both scenarios and in which cases they perform best. Having the HR for M+ and M- group, it is possible to simulate from the piecewise exponential by multiplying the control rate with the HR for each group.

Last but not least, the censoring rate is adjusted at 20% prior to maximum follow up, meaning that 20% of the patients left the study before experienced the event or for unknown reasons their behavior is not recorded. If a patient is not predicted to experience the event after the maximum follow up, will be automatically censored and thus the percentage of censoring rises up to approximately 30%.

3.4.2 Results

As already mentioned, our study is considering two different scenarios. Both strong and moderate scenario are evaluated for L=10000 genes and for K=10 "*sensitivity*" genes. For both scenarios, results in terms of statistical power are evaluated by alternating several parameters (see Table 3.3), that also proved to be significant in Gu Mi's study[Mi, 2017].

Parameters	Possible Values or Levels
Sample Size	1500 | 1800 | 2000 | 2300 | 2500
Learn/Confirm set allocation	40/60 | 50/50 | 60/40 | 70/30
Splits of α	0.04/0.01 | 0.03/0.02 | 0.025/0.025
R	0.7 | 0.75 | 0.8 | 0.85
G	2 | 3 | 5
Sensitive percentage	10 | 25 | 40

Table 3.3: The parameters evaluated in this study

Because HR in M+, M- are adjusted between 0.6 and 1 or 1.1 respectively, the study firstly sets both scenarios with the tuning parameter R equal to

0.75, so as to examine whether the design responds to the goals of the study. To begin with, 5 different sample sizes are applied in the design. At this point, the percentage of sensitive patients is adjusted at 40%, the split of α is 0.025/0.025, the split of learn/confirm set is 50/50 and G=3. The behavior of both scenarios to sample size alterations is showed below.

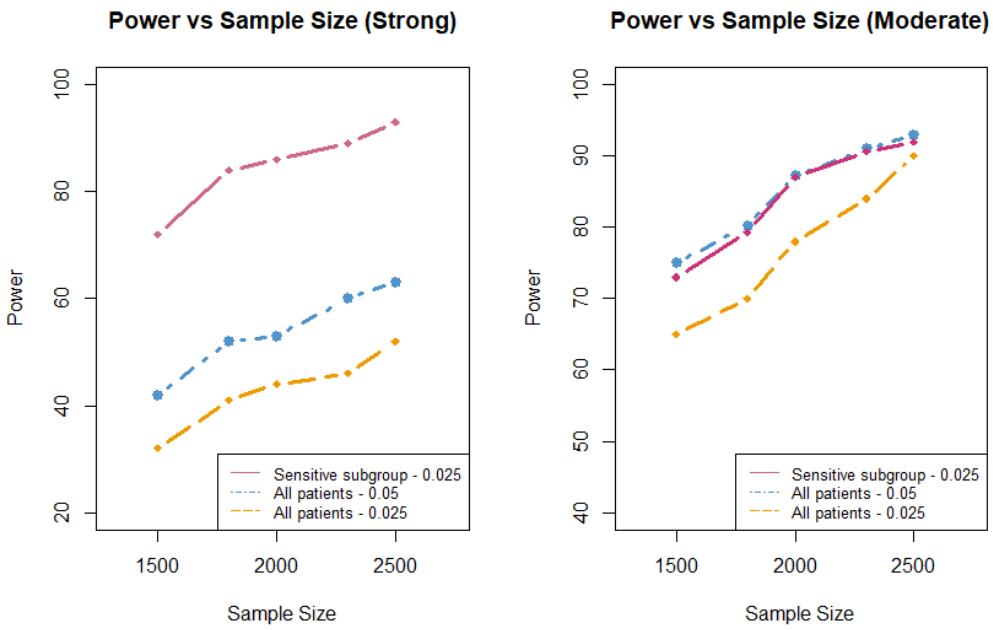

Figure 3.3: Power evaluation (%) for sample size alterations

The results of this parameter showed that the model in general had accuracy over 95% and by examining Figure 3.3 is obvious that:

- In both scenarios as the sample size increases, the power to detect any meaningful difference between the two arms increases.

- In the moderate scenario, the power of the test in all patients is very similar to the power of the test in the sensitive subgroup.
- In the strong scenario, the comparison test in sensitive subgroup has on average 30% more power to detect the treatment effect. The most realistic case that could be used in a real clinical study is when the sample is 1800, giving a power more than 80%.

Keeping up the sample size of 1800 patients, it is important to see if alternating the tuning parameter R will change the performance of the design.

Scenario		Strong			Moderate		
	Patients	All	All	Sensitive	All	All	Sensitive
	P-value	0.05	0.025	0.025	0.05	0.025	0.025
R	0.7	44	34	71	78	70	71
	0.75	52	41	**84**	80	70	**79**
	0.8	48	36	57	82	72	59
	0.85	53	43	34	82	75	46

Table 3.4: Power evaluation (%) for different tuning parameter R

It is obvious from the Table 3.4 that for R=0.75, both scenarios have their best performance and therefore R is a parameter that will be kept constant.

The study continues to the examination of the power loss or gain when changing the proportion of the learn and confirm set. The Figure 3.4 shows the performance for both scenarios.

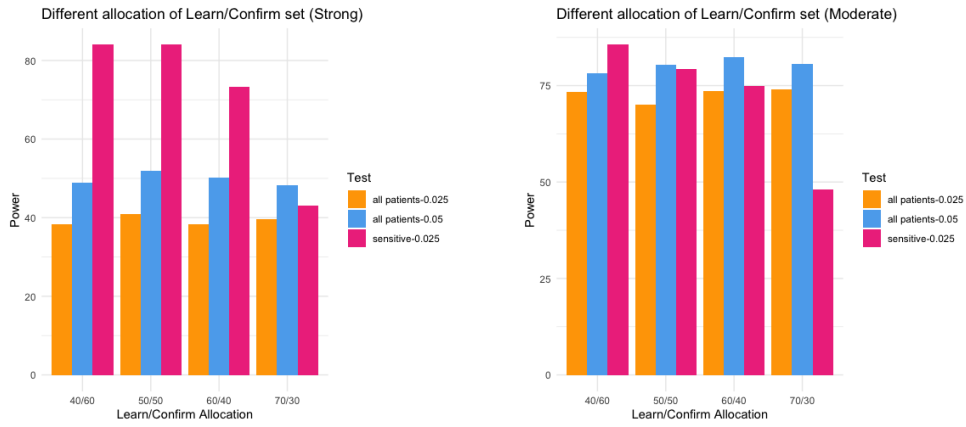

Figure 3.4: Power evaluation (%) for different learn/confirm allocation

Results shown that:

- The moderate scenario needs bigger confirm set that learn set in order to detect a meaningful difference in the subgroup with a power more than 80%.
- The strong scenario has on average the same performance with 50% or 60 % of confirm set.

In addition to the "learn/confirm split", interesting results showed the cases when the level of significance α is divided in different proportions. The Table **??** proves that:

- The moderate scenario outperforms with 0.03/0.02 α split. This is the only case that the subpopulation test is more powerful from the test in all patients.
- The strong scenario performs best for equal allocation in the significance level.

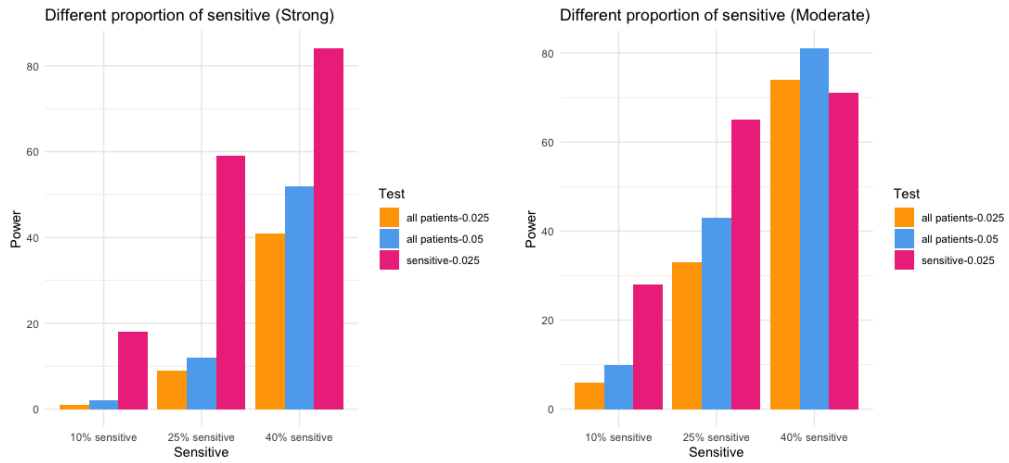

Figure 3.5: Power evaluation (%) for different percentage of sensitive patients

	Scenario	Strong			Moderate		
	Patients	All	All	Sensitive	All	All	Sensitive
	P-value	0.05	α_1	α_2	0.05	α_1	α_2
$\alpha_1 \vert \alpha_2$	0.04|0.01	51	46	72	82	79	67
	0.03|0.02	46	38	79	85	80	**88**
	0.025|0.025	52	41	**84**	80	70	79

Table 3.5: Power evaluation (%) for different split of significance level

When altering the sensitive percentage in 10%, 25% or 40% the Figure 3.5 shows that:

- As the proportion of the sensitive patients is decreased from 40%, the performance of both tests is decreased, but their power to detect a difference in the subpopulation is always bigger.

- The strong scenario has best performance when 40% of patients are sensitive.
- The moderate scenario, for the first time has 20% more power to detect the treatment effect in the subpopulation when the sensitive subgroup is 25%.
- In both cases, the power in the test with all patients, has the better performance with bigger proportion of sensitive patients.

Last but not less important, it was found (see Table 3.6) that for G=2 both scenarios had the best performance.

	Scenario	Strong			Moderate		
	Patients	All	All	Sensitive	All	All	Sensitive
	P-value	0.05	0.025	0.025	0.05	0.025	0.025
G	**2**	53	42	**86**	81	72	**84**
	3	52	41	84	80	70	79
	5	55	43	78	74	63	75

Table 3.6: Power evaluation (%) for different tuning parameter G


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Chapter 4

Discussion

In traditional randomized clinical trials, the way to examine the efficacy of a treatment was a hypothesis test over all the population. This comparison is not always statistically significant, as the specific treatment could be beneficial only for a small subset of patients. Examining treatment efficacy in this way, often leads to the withdrawal of new treatments without knowing whether it was truly useless. However, the development of biomarkers and in general the adaptive designs has contributed in a great extent to the evolution of medical research. In particular, they provide valuable information that accelerates the disease diagnosis and prediction of its response to therapies. In the previous Chapter, we proposed an adaptive design that identifies gene signatures. Our study was influenced in a great extent from the ASD design and the Gu Mi's enhancement of ASD design. We implemented 2 scenarios in the design with a view to examine two different cases.

The first scenario is about a situation where the treatment effect is positive in the sensitive subgroup, but with any effect to the rest of the patients

receiving the treatment (moderate scenario). Results in this scenario showed that, in order to reach a power over 80% in the subpopulation, bigger confirm set than learn set is required. The most impressive in this scenario was that with smaller percentage of sensitive patients, the power of the test in subpopulation was about 30% bigger than in all patients. For this scenario with 1800 patients, a satisfying power of 92% was accomplished with 40% learn set, significance level of 0.03 in the first test and tuning parameters R=0.75 and G=2.

The second scenario is when the treatment effect in the sensitive patients is as before but the rest of the patients experience a decrease in survival (strong). The test in the subpopulation outperformed in most of the cases with a difference at about 30% from the test in all patients. A possible explanation is that when a proportion of patients is benefited from the treatment and other patients are not, the overall test can not accumulate the size of the treatment effect. Whereas in the subpopulation are included all the patients that have increased survival times and so the test has a lot of power to detect the effect. For this scenario with 1800 patients, a satisfying power of 86% was accomplished with 50% learn set, equal split in significance level and tuning parameters R=0.75 and G=2.

We can not choose between these two scenarios, but what we can say that both of them are capable of detecting a treatment effect, even when the sensitive patients are limited. Of course, both scenarios could reach more than 95% if the sample size was over 3000 patients, but a case like that can not be realistic.

To make a conclusion, further research to our study could be a more

efficient way of internal and external validation than splitting the sample size in different proportions. Re-sampling techniques like cross-validation and bootstrap may be are capable of ameliorating the performance of the design and reducing prediction error. Moreover, we evaluated scenarios with 10 sensitivity genes from 10000 total genes, so a possible extension could be to find different combinations of L and K that could offer a better results.

Bibliography

[FDA,] U.s. food and drug administration. Accessed: 2019-03-04.

[NIH,] U.s. national institutes of health. Accessed: 2019-02-12.

[Aickin and Gensler, 1996] Aickin, M. and Gensler, H. (1996). Adjusting for multiple testing when reporting research results: the bonferroni vs holm methods. *American journal of public health*, 86(5):726–728.

[Altman et al., 2009] Altman, D. G., Vergouwe, Y., Royston, P., and Moons, K. G. (2009). Prognosis and prognostic research: validating a prognostic model. *Bmj*, 338:b605.

[Association et al., 2013] Association, W. M. et al. (2013). Wma declaration of helsinki-ethical principles for medical research involving human subjects.

[Barlow and Hayes, 1979] Barlow, D. H. and Hayes, S. C. (1979). Alternating treatments design: One strategy for comparing the effects of two treatments in a single subject. *Journal of applied behavior analysis*, 12(2):199–210.

[Bernard, 1957] Bernard, C. (1957). *An introduction to the study of experimental medicine*, volume 400. Courier Corporation.

[Cambon et al., 2015a] Cambon, A. C., Baumgartner, K. B., Brock, G. N., Cooper, N. G., Wu, D., and Rai, S. N. (2015a). Classification of clinical outcomes using high-throughput informatics: Part 1–nonparametric method reviews. *Model Assisted Statistics and Applications*, 10(1):3–23.

[Cambon et al., 2015b] Cambon, A. C., Baumgartner, K. B., Brock, G. N., Cooper, N. G., Wu, D., and Rai, S. N. (2015b). Classification of clinical outcomes using high-throughput informatics: Part 2-parametric method reviews. *Model Assisted Statistics and Applications*, 10(2):89–107.

[Cambon et al., 2017] Cambon, A. C., Baumgartner, K. B., Brock, G. N., Cooper, N. G., Wu, D., and Rai, S. N. (2017). Properties of adaptive clinical trial signature design in the presence of gene and gene-treatment interaction. *Communications in Statistics-Simulation and Computation*, 46(10):8233–8250.

[Chalmers et al., 2008] Chalmers, I., Milne, I., Tröhler, U., Vandenbroucke, J., Morabia, A., Tait, G., and Dukan, E. (2008). The james lind library: explaining and illustrating the evolution of fair tests of medical treatments. *The journal of the Royal College of Physicians of Edinburgh*, 38(3):259–264.

[Chen et al., 2014] Chen, J. J., Lu, T.-P., Chen, D.-T., and Wang, S.-J. (2014). Biomarker adaptive designs in clinical trials. *Translational Cancer Research*, 3(3):279–292.

[Chen et al., 2018] Chen, Y.-C., Lee, U. J., Tsai, C.-A., and Chen, J. J. (2018). Development of predictive signatures for treatment selection in

precision medicine with survival outcomes. *Pharmaceutical statistics*, 17(2):105–116.

[Chow and Liu, 2008] Chow, S.-C. and Liu, J.-p. (2008). *Design and analysis of clinical trials: concepts and methodologies*, volume 507. John Wiley & Sons.

[Cox, 1972] Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220.

[Department of Health et al., 2014] Department of Health, E. et al. (2014). The belmont report. ethical principles and guidelines for the protection of human subjects of research. *The Journal of the American College of Dentists*, 81(3):4.

[Diao et al., 2018] Diao, G., Dong, J., Zeng, D., Ke, C., Rong, A., and Ibrahim, J. G. (2018). Biomarker threshold adaptive designs for survival endpoints. *Journal of biopharmaceutical statistics*, 28(6):1038–1054.

[Dijkland et al., 2018] Dijkland, S., Retel Helmrich, I., and Steyerberg, E. (2018). Validation of prognostic models: challenges and opportunities. *Journal of Emergency and Critical Care Medicine*, 2(91):1–4.

[Dodgson, 2006] Dodgson, S. J. (2006). Evolution of clinical trials. *The Write Stuff. In press. Will be accessible at http://www. emwa. org.*

[Fedorov and Liu, 2014] Fedorov, V. V. and Liu, T. (2014). Enrichment design. *Wiley StatsRef: Statistics Reference Online.*

[Feinstein, 1995] Feinstein, A. R. (1995). Clinical epidemiology: the architecture of clinical research. *Statistics in Medicine*, 14(11):1263–1263.

[Ferreira and Patino, 2017] Ferreira, J. C. and Patino, C. M. (2017). Types of outcomes in clinical research. *Jornal brasileiro de pneumologia: publicacao oficial da Sociedade Brasileira de Pneumologia e Tisilogia*, 43(1):5–5.

[Festing, 2003] Festing, M. F. (2003). Principles: the need for better experimental design. *Trends in pharmacological sciences*, 24(7):341–345.

[Fox, 2002] Fox, J. (2002). Cox proportional-hazards regression for survival data. *An R and S-PLUS companion to applied regression*, 2002.

[Freidlin et al., 2010] Freidlin, B., Jiang, W., and Simon, R. (2010). The cross-validated adaptive signature design. *Clinical Cancer Research*, 16(2):691–698.

[Freidlin and Simon, 2005] Freidlin, B. and Simon, R. (2005). Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clinical cancer research*, 11(21):7872–7878.

[Giuffrida, 2016] Giuffrida, M. A. (2016). Defining the primary research question in veterinary clinical studies. *Journal of the American Veterinary Medical Association*, 249(5):547–551.

[Gonzalez et al., 2009] Gonzalez, C. D., Bolaños, R., and de Sereday, M. (2009). Editorial on hypothesis and objectives in clinical trials: superiority, equivalence and non-inferiority.

[Govindan et al., 2015] Govindan, R., Mandrekar, S. J., Gerber, D. E., Oxnard, G. R., Dahlberg, S. E., Chaft, J., Malik, S., Mooney, M., Abrams, J. S., Jänne, P. A., et al. (2015). Alchemist trials: a golden opportunity to transform outcomes in early-stage non–small cell lung cancer. *Clinical Cancer Research*, 21(24):5439–5444.

[Green, 2000] Green, S. B. (2000). Hypothesis testing in clinical trials. *Hematology/oncology clinics of North America*, 14(4):785–795.

[Grodin, 1992] Grodin, M. A. (1992). Historical origins of the nuremberg code. *Nazi doctors and the Nuremberg code*.

[He and Allen, 2010] He, M. and Allen, A. (2010). Testing gene–treatment interactions in pharmacogenetic studies. *Journal of biopharmaceutical statistics*, 20(2):301–314.

[Hirakawa et al., 2018] Hirakawa, A., Asano, J., Sato, H., and Teramukai, S. (2018). Master protocol trials in oncology: Review and new trial designs. *Contemporary clinical trials communications*, 12:1–8.

[Hoering et al., 2008] Hoering, A., LeBlanc, M., and Crowley, J. J. (2008). Randomized phase iii clinical trial designs for targeted agents. *Clinical Cancer Research*, 14(14):4358–4367.

[Jandhyala et al., 2013] Jandhyala, V., Fotopoulos, S., MacNeill, I., and Liu, P. (2013). Inference for single and multiple change-points in time series. *Journal of Time Series Analysis*, 34(4):423–446.

[Kalow et al., 1998] Kalow, W., Tang, B., and Endrenyi, L. (1998). Hypothesis: comparisons of inter-and intra-individual variations can substitute for twin studies in drug research. *Pharmacogenetics*, 8(4):283–290.

[Lewis, 1999] Lewis, J. A. (1999). Statistical principles for clinical trials (ich e9): an introductory note on an international guideline. *Statistics in medicine*, 18(15):1903–1942.

[Lind, 1980] Lind, J. (1980). *A Treatise on the Scurvy*. Classics of Medicine Library.

[Maitournam and Simon, 2005] Maitournam, A. and Simon, R. (2005). On the efficiency of targeted clinical trials. *Statistics in medicine*, 24(3):329–339.

[Matsui et al., 2012] Matsui, S., Simon, R., Qu, P., Shaughnessy, J. D., Barlogie, B., and Crowley, J. (2012). Developing and validating continuous genomic signatures in randomized clinical trials for predictive medicine. *Clinical Cancer Research*, 18(21):6065–6073.

[Mi, 2017] Mi, G. (2017). Enhancement of the adaptive signature design for learning and confirming in a single pivotal trial. *Pharmaceutical statistics*, 16(5):312–321.

[Miranta Antoniou,] Miranta Antoniou, Andrea Jorgensen, R. K.-D. Enrichment design. Accessed: 2019-05-20.

[Nellhaus and Davies, 2017a] Nellhaus, E. M. and Davies, T. H. (2017a). Evolution of clinical trials throughout history. *Marshall Journal of Medicine*, 3(1):41.

[Nellhaus and Davies, 2017b] Nellhaus, E. M. and Davies, T. H. (2017b). Evolution of clinical trials throughout history. *Marshall Journal of Medicine*, 3(1):41.

[Paul and Brookes, 2015] Paul, C. and Brookes, B. (2015). The rationalization of unethical research: revisionist accounts of the tuskegee syphilis study and the new zealand "unfortunate experiment". *American journal of public health*, 105(10):e12–e19.

[Piantadosi, 2017] Piantadosi, S. (2017). *Clinical trials: a methodologic perspective*. John Wiley & Sons.

[Rabaglio et al., 2009] Rabaglio, M., Sun, Z., Price, K., Castiglione-Gertsch, M., Hawle, H., Thürlimann, B., Mouridsen, H., Campone, M., Forbes, J. F., Paridaens, R., et al. (2009). Bone fractures among postmenopausal patients with endocrine-responsive early breast cancer treated with 5 years of letrozole or tamoxifen in the big 1-98 trial. *Annals of oncology*, 20(9):1489–1498.

[Renfro and Sargent, 2016] Renfro, L. and Sargent, D. (2016). Statistical controversies in clinical research: basket trials, umbrella trials, and other master protocols: a review and examples. *Annals of Oncology*, 28(1):34–43.

[Rosenwald et al., 2002] Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B., Giltnane, J. M., et al. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. *New England Journal of Medicine*, 346(25):1937–1947.

[Sanchez and Binkowitz, 1999] Sanchez, M. M. and Binkowitz, B. S. (1999). Guidelines for measurement validation in clinical trial design. *Journal of biopharmaceutical statistics*, 9(3):417–438.

[Simon and Simon, 2013a] Simon, N. and Simon, R. (2013a). Adaptive enrichment designs for clinical trials. *Biostatistics*, 14(4):613–625.

[Simon and Simon, 2013b] Simon, N. and Simon, R. (2013b). Adaptive enrichment designs for clinical trials. *Biostatistics*, 14(4):613–625.

[Simon, 2017] Simon, R. (2017). Critical review of umbrella, basket, and platform designs for oncology clinical trials. *Clinical Pharmacology & Therapeutics*, 102(6):934–941.

[Zhang et al., 2017] Zhang, Z., Li, M., Lin, M., Soon, G., Greene, T., and Shen, C. (2017). Subgroup selection in adaptive signature designs of confirmatory clinical trials. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(2):345–361.