

POLITECNICO DI MILANO

Scuola di Ingegneria Industriale e dell'Informazione

Corso di Laurea Magistrale in Management Engineering



**FAILURE, BIAS AND RISK-TAKING IN
SCIENTIFIC RESEARCH. A NEW
METHOD AND ESTIMATE**

Supervisor: Prof. Chiara Franzoni

MSc Dissertation by:
Fabrizio Tropea 914123
Davide Trusso 915454

ABSTRACT

It is recognized by literature that investing in R&D is not socially optimal, due to the expected market failure. Scientific research involves several actors, each one with its own objectives and priorities. In particular, it is demonstrated by the literature that governments have to fill the gap of underinvestment from corporations and privates in R&D. The most relevant market failures affecting investment decisions in scientific research field can be identified as the following:

- Uncertainty of returns: returns in R&D are uncertain, so the desired outcome is highly unpredictable. Since corporations could not, a priori, have a clear understanding of the risk/returns profile in R&D, they could decide to lower their efforts to reduce this uncertainty. In this sense, uncertainty of returns can lower the innovation level.
- Imperfect appropriability: in scientific research a company cannot always claim an invention through a patent, due to the essence of the research itself. As a consequence, corporations have no incentives to pursue the research.
- Short-termism: it is demonstrated by literature that corporations are redirecting their investments towards more marketable findings.

With our dissertation, we analyze how investors take their investment decisions under the constraints of market failures. One of the most impacting criteria, in order to evaluate how to take an investment decision, is the risk/return profile of

that investment. While in the financial world, the risk associated to an investment can be analyzed through historical track records and trends evaluation, this is not true for this field, where the effect resulting from the simultaneous presence of several factors is influencing the risk of failure of a trial. As a consequence, nowadays, it is not univocally possible to evaluate the risk associated to an investment in scientific research. In this sense, with our dissertation, we investigate which factors could let an actor to select a certain project and how these are influencing the final outcome of the investment.

ABSTRACT

Dai riferimenti letterari si evince che gli investimenti in R&D non sono ottimali sul piano sociale, poiché soggetti a market failures. Nello specifico, nella letteratura, viene trattato come i governi debbano colmare i gap lasciati dai mancati investimenti del settore privato nella ricerca scientifica. Le principali cause di market failure identificate dalla letteratura sono:

- Incertezza dei ritorni: i ricavi provenienti dagli investimenti in ricerca e sviluppo sono incerti, comportando l'imprevedibilità dei ritorni. Le compagnie, dal momento che non possono definire a priori il profilo di rischio/ritorno per gli investimenti in R&D, potrebbero decidere di ridurre i loro investimenti in questo settore per ridurre il rischio. Si evince dunque che l'incertezza dei ritorni potrebbe ridurre i loro investimenti nel settore.
- Imperfect appropriability: nella ricerca scientifica, non tutte le invenzioni e innovazioni sono appropriabili tramite brevetto. A conseguenza di ciò, le compagnie potrebbero necessitare di incentivi per intraprendere le ricerche in questo ambito.
- Short-termism: viene dimostrato dalla letteratura scientifica come le compagnie reindirizzino i loro investimenti verso soluzioni prossime alla commercializzazione.

Con il nostro lavoro di ricerca, abbiamo analizzato come gli investitori selezionano i progetti, soggetti alla presenza di market failures. Uno dei principali criteri nella valutazione degli investimenti è l'analisi del profilo di rischio/ritorno.

Mentre nell'ambito finanziario il rischio associato ad un investimento può essere quantificato tramite serie storiche o analisi dei trend, questo non è altrettanto vero nell'ambito della ricerca scientifica, dove la contemporanea presenza di più fattori influenza il rischio di fallimento del progetto. Di conseguenza, attualmente, non è possibile identificare un metodo univoco per la valutazione del rischio di una ricerca scientifica. Con il nostro lavoro vogliamo osservare quali di questi fattori spingono gli investitori a finanziare un determinato progetto e come questi influenzino l'outcome finale.

ACKNOWLEDGMENT

Questa tesi rappresenta la fine del nostro percorso Universitario, frutto di 5 anni intensi e formativi. Per questo motivo vorremmo ringraziare tutte le persone che ci sono state vicine e ci hanno accompagnato fin qui.

Un ringraziamento particolare va alla nostra relatrice, Chiara Franzoni, per l'opportunità e il continuo supporto prestatoci.

Ringraziamo le nostre famiglie per averci permesso di perseguire questo percorso, gli amici incontrati al Politecnico con i quali abbiamo condiviso questa avventura, in particolare Giorgio, Claudio, Gabri, Matte, Julian e Guido.

Davide: Fra tutte le persone che mi sono state vicine in questi anni, un ringraziamento particolare va a Tommi e Fabi, fratelli da sempre, è grazie a voi se ogni volta che tornavo a Rimini mi sentivo a casa di nuovo. Grazie Turuzzo per aver reso la vita a Milano leggera anche nelle giornate più pesanti. Grazie a tutti i Quatarelli per avermi ricordato cosa vuol dire far parte di una famiglia, vi voglio bene. Un ringraziamento speciale alla Gigia, per le tue attenzioni e i tuoi sorrisi, mi hai insegnato molto più di quanto imparato in questo percorso. Ringrazio i miei nonni per essermi sempre stati accanto, mi mancate molto.

Fabrizio: Un ringraziamento speciale a Paolo per i suoi consigli sulla letteratura medica. Grazie alle persone a me più care che hanno visto nascere questa tesi: Cri ed Enri. Terrei inoltre a ringraziare Bea per il suo contagioso buon umore.

TABLE OF CONTENTS

Chapter 1 – Introduction	11
1.1 Context	11
1.2 Overall Research Aim	13
Chapter 2 – Literature Review and General Overview ..	14
2.1 Literature Review	14
2.1.1 Market failure in scientific research	14
2.1.2 Public and private sector in scientific research.....	17
2.1.3 Short-termism of corporations	19
2.1.4 Risk-taking in scientific research.....	20
2.1.5 Selective Reporting.....	23
2.1.6 Evaluation of Clinical trials	26
2.1.7 Research objectives.....	27

2.2 Clinical trials: a theoretical framework	29
2.2.1 Phases of clinical trials	29
2.2.2 Sponsor of clinical trials	32
2.2.3 LPM (Likely Precision Medicines) trials.....	33
2.3 Interpreting clinical trials	36
2.3.1 Primary outcome measure	36
2.3.2 Statistical analysis.....	38
2.3.3 Confidence intervals	40
2.3.4 Significance Testing	40
2.3.5 Types of error.....	41
2.3.6 Further considerations	42
Chapter 3 - Database Creation.....	44
3.1 Data Source	44
3.1.1 Available information.....	45
3.2 Sample Selection	49

3.3 Data Extraction and Cleaning.....	50
3.4 Data Encoding	54
Chapter 4 – Data Analysis and Interpretation	63
4.1 Introduction and Research Strategy	63
4.2 Methodologies	66
4.2.1 Probit Regression.....	66
4.2.1.1 Probit Execution.....	68
4.2.2 Heckman Selection Model.....	72
4.2.2.1 New discrete dependent variable.....	72
4.2.2.2 Selection bias.....	73
4.2.2.3 Heckit model execution.....	74
4.2.3 Survival Analysis.....	76
4.2.3.1 Dataset adjustments.....	78
4.2.3.2 Survival Analysis Execution	79
4.2.3.3 Cox proportional hazard execution	79
4.3 Output Analysis.....	82

4.3.1 Probit Output Analysis	82
4.3.2 Heckit Model Output Analysis	85
4.3.3 Cox Proportional Hazard Output Analysis	86
4.3.2.1 Kaplan-Meyer.....	87
Chapter 5 - Conclusion	89
5.1 Hypotesis analysis	90
5.2 Limitations.....	91
5.3 Further discussion.....	92
Bibliography.....	94

Chapter 1 – Introduction

1.1 Context

In this work we analyzed the field of scientific research, considering it as a locus of market opportunities for investors. Just like every market, every time an actor wants to sponsor an investment, he has to take into account the presence of market failures. As we will see along this dissertation, market failures impact the behavior of investors and the selection process of each project. For the purpose of our work we decided to analyze empirically clinical trials as investments inside scientific research market. NIH (National Institute of Health), the primary US government agency responsible for biomedical and public health researches, gives a formal definition of what a clinical trial is:

“A research study in which one or more human subjects are prospectively assigned to one or more interventions (which may include placebo or other control) to evaluate the effects of those interventions on health-related biomedical or behavioral outcomes”

Most commonly, clinical trials are used to test the safety and effectiveness of drugs and devices. Different actors are involved in the sponsoring phase, such as pharmaceutical companies, universities, government authorities. In general, trials can be conducted by research teams or medical professionals. Clinical trials, thanks to their contribution, enlarge the set of information available for the scientific field, providing additional documentation on existing or innovative procedures, treatments or chemical composites.

In particular, clinical trials field is one of the most complete and discussed in terms of information available. Every time a clinical trial is performed, a well-

defined procedure, with milestones and requirements is requested by authorities who govern this process. One of the most reliable repositories of clinical trials is Clinicaltrials.gov, maintained by NIH. According to the requirements set by law, each study must present specific information regardless the final outcome, to document the entire process in the realization of the investment. Some of this information crucial to deeply analyze a trial are: its relevant dates (Start Date, Primary Completion Date and Completion Date of the trial), the result and the measures of each Outcome (both Primary and Secondary), the Intervention Type and all the aspects related to the Sponsor of the project (such as Class and Name). It is necessary to consider that clinical trials can be divided according to their therapeutic area. During our dissertation, we decided to focus only on one therapeutic area, that is the one of Oncology, which is the leading one for quality of data available and trials completed among years¹.

Cancer research and, more in general, the entire scientific research, need to be financed in order to be performed. There are mainly two categories of investors: the public sector and the private one. In our dissertation we analyzed the relation between these two actors, finding criticalities arising from the presence of market failures, that could lead to underinvestment for certain trials typologies.

¹ <https://pharmaintelligence.informa.com/~media/informa-shop-window/pharma/2019/files/pdf/trialtrove-2018-completed-trials-state-of-industry-sponsored-clinical-development.pdf>

1.2 Overall Research Aim

In paragraph 1.1 Context, we draft an overall picture of the domain of analysis. This step was useful to understand the main dynamics of clinical trials field. When dealing with trials' outcome, there is not a well-defined and unique relation that allows to state at priori that a certain combination of trials' attributes will lead to the success of it. This is also due to the high uncertainty in the R&D context and the presence of market failures. This led us to find hidden relations between key attributes of trials and the statistical success of trials.

Among the different actors involved in the financing of trials, Industries were one of the most present in terms of trials sponsored in our sample. This led us to further investigate this category, in order to find the presence of correlations between attributes and probability of success. Before interpreting any result, we gathered and studied a selected pool of papers treating this topic, discovering that corporations are redirecting their investments in R&D towards more marketable solutions. Moreover, they are also seeking for the minimization of the risk of failure of a trial. In order to do so, it's required to them to select only studies suitable for their purpose. In this sense we investigated which are the preferred categories of trials and the risk of failure associated to each of them.

Once discovered relations between corporation and sponsorship of trials, we looked for the most relevant recent trends in the clinical trial field. Among these, we found as extremely interesting and significant the one of using Mixed Methodologies as intervention type. Following the same approach of evaluation of risk used for corporation, we tried to determine whether the presence of this intervention type led to an increase in the final success rate of a trial.

The conclusion of our work is focused on the presentation of results obtained, the interpretation of these and suggestions for future developments of the model created.

Chapter 2 – Literature Review and General Overview

2.1 Literature Review

The literature review starts from the analysis of the findings about scientific research and the funders of it, focusing on the major economic problems related to it. Then the analysis focuses more on clinical trials, that are the focal point of our work. This chapter starts with a paragraph describing the main sources of market failure in scientific research field. Then, we discussed the context of investments in scientific research, both from private sector and public sector. One important point to be analyzed when dealing with markets is the external economy problem, a focal point for the dissertation, explaining why certain investments are not pursued.

After these premises, we focused on the attitude of corporations on short-termism, a behavior determining the under investment by corporations in scientific research. Afterwards we introduced the concept of risk-taking about the selection of the researches to fund. Then, we set our point of view on clinical trials, describing selective reporting, a problem that underline the difficulty to have reliable data about results of clinical trials. Finally, after having described these economical concepts regarding both scientific research in general and clinical trials, we set our hypotheses to test.

2.1.1 Market failure in scientific research

Market failure in R&D is a relevant issue for the aim of this work. Indeed, literature shows that allocating market resources for R&D is not socially optimal

due to the expected market failure (Choi, Lee, 2017). Among the reasons behind this phenomenon there is uncertainty, defined by F. Knight as the third type of probability. It describes a situation in which an actor cannot assign a priori a probability to an event (Knight, 1921). This is true especially for scientific research, where we cannot assign a probability of success or failure to a study, unlikely to what happens for the throw of a coin, where we have a priori probability.

The so called “Knightian uncertainty” theory relates risk to profitability, and the literature tried to model the effect of the uncertainty on R&D returns (Amoroso, Moncada-Paternò-Castello, Vezzani, 2017). The main finding is that uncertainty can slow down innovation and *“In this context, R&D policies could be particularly effective by preventing firms to lower their R&D efforts (as a consequence of uncertainty)”* (Amoroso, Moncada-Paternò-Castello, Vezzani, 2017).

The effect of policies is also discussed by Rao (2015), that shows how regulatory process of drugs and treatments approval is detrimental to innovation.

Thus, uncertainty of returns can lower innovation level, due to the unpredictability of a desired outcome, but this is not the only issue related to the market failure of scientific research sector.

The other issue to consider is the one of external economies, that play an important role in this scenario, especially for basic scientific research, that is more likely to generate external economies (Nelson, 1959). This is due to the essence of the scientific research itself, since it’s *“quite likely that a firm will be unable to capture through patent rights the full economic value created in a basic research project that it sponsors”* (Nelson, 1959).

This is essentially the problem of appropriability in R&D, that generates a negative externality in the market, indeed, if a company cannot appropriate of an invention through a patent, it has not incentives to pursue that research.

If an innovator can successfully capture the social benefits resulting from its innovation, we have perfect appropriability (Shapiro, 2011).

As a consequence, if a firm is able to protect the competitive advantage coming from its invention against competitors, it means that appropriability is high, vice versa we have poor or imperfect appropriability. The typical instrument to ensure the appropriability of an invention is the patent and intellectual property rights (IPR), more in general. As a consequence, in absence of these instruments to protect the knowledge that inventors and innovators create, they would have less incentives to innovate, since competitors would be able to imitate their findings at low or zero costs (Arrow, 1962).

Then it seems reasonable to state that *“Increased appropriability spurs innovation”* (Shapiro, 2011).

This theoretical framework refers to generic knowledge, without specifying the nature of the invention or innovation. If we refer to our context, the one of scientific research, there is one point we can add to the framework. In particular, the context of scientific research, that is basically knowledge based, is not reducible to pure codified knowledge (Foray, 2010). Indeed, research results and new treatments are very difficult to be formalized to be a set of codified instructions that can be simply reproduced by following the instructions, (Foray, 2010) in the same way as it happens for coding and IT in general.

Considering this point for our reference context, the imperfect appropriability externality could be considered as one of the causes for scientific research market failure.

To sum up, market failures in the scientific research market are generated mainly by uncertainty of returns and weak or absent appropriability of scientific research.

2.1.2 Public and private sector in scientific research

Now it's appropriate to consider more in detail the actors and dynamics of appropriability and patents in our reference context. Looking at the actors involved in our case, we have private companies and public sector investing in scientific research and society receiving the benefits of findings of scientific research.

For society, the advantages deriving from a research project would increase the total welfare. As a consequence, the non-realization of the research project is a loss for society. Since corporations have profit maximization as the primary goal of their activities, they are likely to decline to embark on these unprofitable projects and it is the duty of the government to fill the gap (Von Mises, 2010).

An important fact to underline is that society, only in the case in which the benefits of the research are relevant for the society, will benefit from that research only when companies share the findings and results of their studies. Indeed, profits generated for firms to keep research findings secret produce results that are economically inefficient. But if scientific knowledge is thus administered, the incentives of private firms to create new knowledge will be reduced. In absence of incentives to private firms to publish results quickly, a dollar spent on basic research in university is worth more than in an industry (Nelson, 1959).

Patentees, both U.S. and non-U.S., and corporations in particular, increasingly depend upon federally supported research as a source of scientific knowledge. Almost one-third of U.S. invention and the more important part as measured by future citations, renewals, and novelty rely on federal research investment (Fleming, Greene, Li, Marx, Yao, 2019).

As already explained in the paragraph 2.1.1 Market failure in scientific research, due to the uncertainty of returns and difficult appropriability of findings, companies financing scientific research have poor incentives to invest in R&D. Scientific research is fundamental, for its strict linkage with, economic growth (Adams 1990; Jaffe 1989; Stephan 1996), because uncertainty of returns and its features negatively affect the willingness of privates to invest. The result of this is that private sector underinvests in scientific research (Arrow 1962; Nelson 1959). This trend was also confirmed by empirical works that showed, even considering the high level of R&D investments before 1980s, that the vast majority of innovations comes from the government-based projects (Fleming et al. 2019). This led to fewer scientific publications over time (Arora, Belenzon, and Patacconi 2018) and its mainly directed towards the most marketable solutions (Budish, Roin, and Williams 2015).

There are some reasons explaining this phenomenon. In particular, research results often are of little value to the firm sponsoring research, thought of great value to another firm, and, second, that research results often cannot be quickly patented (Nelson, 1959).

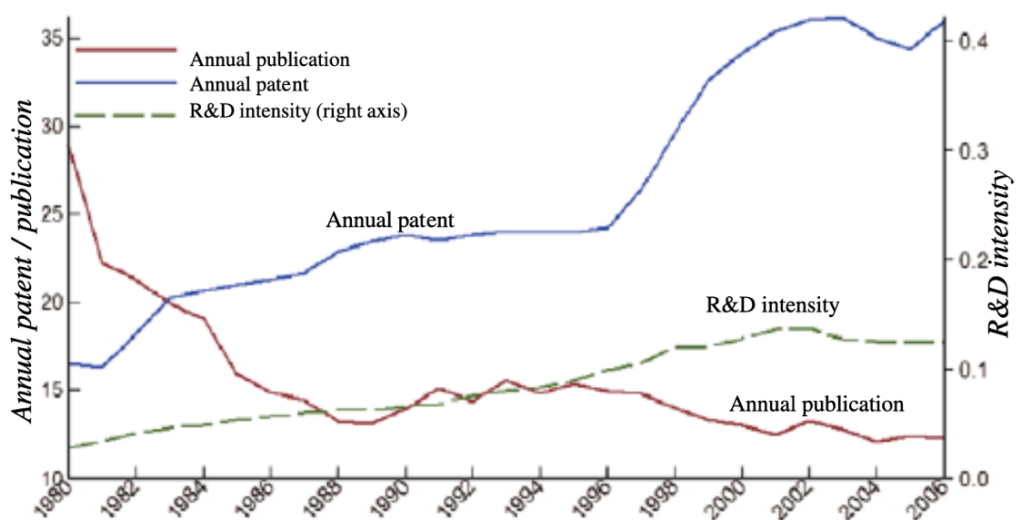


Figure 1 – Source: Arora, Belenzon, Patacconi (2017)

Figure 1 shows the declining over time of publications and at the same time, the increase of patented research, underlining the redirection to marketable researches, as explained before.

2.1.3 Short-termism of corporations

In addition to the uncertainty of returns and imperfect appropriability, it is possible to underline that there is another behavior of companies investing in scientific research, that is the one of redirection of investments towards more marketable findings, already anticipated in previous paragraph **Errore. L'origine riferimento non è stata trovata.** Indeed, labs of large corporations increasingly focus on developing existing knowledge and commercializing it, rather than creating new knowledge (Arora, Patacconi, Belenzon, 2015). As a consequence, we are witnessing from years a redirection by many leading firms, of resources and attention from more exploratory scientific research toward more commercially oriented projects (Arora, Patacconi, Belenzon, 2015).

So, in general we can state that private firms are willing to focus on commercial projects, the ones that aim at pursuing a profit opportunity in the short term, rather than focusing on exploratory scientific research. Corporations are in this way following short-termism, that is the preference for actions that secure short-term benefits. We can underline a reason explaining the short-termism of corporations: *“mounting evidence indicates that capital markets often apply short-term pressure on firms to gain short-term results by focusing primarily on reported financial performance”* (Dunk, Kilgore, 2000).

It is also possible to highlight another behavior, analyzed by the paper *“Do firms underinvest in Long-Term Research? Evidence from Cancer Clinical Trials, Budish, Roin, Williams, 2015”*, that states the following sentence:

“Private firms may be particularly likely to focus on the short term in the context of research and development (R&D) due to the structure of the patent system. Patents award innovators are subjected to a fixed period of market exclusivity (e.g., 20 years in the United States). Being the covering time of patents effective from the time of discovery (“invention”) rather than first sale (“commercialization”), the lag time reduces significantly the coverage due to the patenting. This means that the patent system provides, perhaps inadvertently, very little incentive for private firms to engage in long-term research.”

2.1.4 Risk-taking in scientific research

As already anticipated, the focus of this work is the scientific research, in particular the field of clinical trials, so from now on, the analysis will be focused in this section of scientific research.

A review of the literature suggests that analyst and shareholder bias against high-risk long-term research in favor of lower-risk, short-term product R&D influences organizations to reduce the time it takes to get products to market when the emphasis in the market place is on cost competition rather than product innovation (Dunk, Kilgore, 2000).

This finding introduces the second part of the literature review, that is focused on the risk-taking approach of investors in scientific research.

At this point, it’s useful to introduce another concept that is important for corporations deciding in which type of clinical trials to invest.

Indeed, we have evidence that corporations are focusing on short-term and marketable scientific research, but there is another element to consider, that is the one of risk-taking, because we have evidence that corporations treat investments in scientific research exactly as they treat a generic investment, so looking at risk/return profile (Campbell, 1996). In particular, the evaluation of

the risk of an asset or an investment, is a matter largely discussed in literature. The determination of how to measure the risk of an investment and the identification of factors determining the price of risk are two fundamental questions in this field (Campbell, 1996).

In the financial world, risk measurement is well established, also thanks to the availability of data. If, for example, we think about the stock market, we have the possibility to track the value of a certain stock in the time and as a consequence, evaluate the volatility of that stock and the risk associated to it. For this reason, an investor has some information to rely on to make an informed decision.

In particular, the costs of an investment in the financial world is typically disclosed to investors, while it is not the same for scientific research, because the factors impacting the costs of a research are a lot and not always easily computable. As a consequence, we can only provide an estimate of the costs of clinical trials. It is notable to state that data here presented are taken from a specialized source, that is Sofpromed, a European full-service contract research organization (CRO) specialized in the integral management of phase I-IV clinical trials and observational studies in oncology.²

They provide an exhaustive list of factors impacting clinical trials costs. Among them we have the study size (patients involved), locations (number of countries involved), number of clinical sites, therapeutic area, drug type.

As we will see more in detail in the next paragraphs, a clinical trial belongs to a specific phase and each one of them has a different time-horizon and different requirements. For instance, the study size is related to the study phase (i.e. Phase 1 trials require 20-80 patients, while Phase 3 studies may involve hundreds or thousands of subjects). Thus, it is difficult to give a single price answer for a clinical trial, it depends on the mentioned factors.

² <https://www.sofpromed.com/company/>

Just to give an idea of times and costs of a clinical trial, we report an example taken from Sofpromed³:

Clinical trial here reported belongs to phase 3, which is the closest to commercialization for a drug for advanced tumors. The study recruited 350 patients located in two countries with the following timeline:

- Start-up: 6 months
- Recruitment: 36 months
- Per-patient treatment: 6 months
- Survival follow-up: 12 months for last patient considered
- Close-out: 6 months
- Total study duration: 66 months

It is provided a cost for each activity of the trial, like regulatory affairs, site identification and selection, site management, onsite monitoring, drug logistics, medical writing, project management, document management, data management, quality control. The total budget is 12.900.000\$ but it is important to consider that this budget cannot be taken as reference for all the other studies belonging to phase 3. Anyway, a general rule of thumb suggests that the average cost of phase 1, 2 and 3 clinical trials is 4, 13 and 20 million \$ respectively⁴.

³ <https://www.sofpromed.com/how-much-does-a-clinical-trial-cost/>

⁴ <https://aspe.hhs.gov/report/examination-clinical-trial-costs-and-barriers-drug-development>

As we anticipated, each phase of clinical trial requires different timeframe for the trial. In particular, phase 1 trials typically require several months, at maximum one year; phase 2 trials require from one to two years and phase 3 trials require from one to four years.

This multiplicity of factors and criteria that determine the final cost and expected duration of a trial suggests us that differently from what we have in financial investments, we don't have the same availability of information about time and costs of clinical trials and this makes it harder to evaluate the risk of a clinical trial.

The aim of this work is also to find a way to evaluate risk associated to the failure of a clinical trial, since it could allow investors to evaluate investment opportunities more consciously.

2.1.5 Selective Reporting

In this section we analyze the role of selective reporting in clinical trials. It is defined as *“the incomplete publication of outcomes measured, or analyses performed in a study, that may lead to the over or underestimation of treatment effects or harms⁵”*. For example, a trial could be published omitting or misrepresenting outcomes, showing only those that have successful results for the study. This phenomenon is also known as outcome reporting bias. The selection of outcomes to show leads to the submission of trials of poor quality, where information is not totally available to doctors (Hemmiki, 1980).

The literature shows that *“selective reporting of research findings in clinical trials, it's real and it's diffused. Selective reporting can lead to concerns ranging from publishing flawed scientific knowledge, to skewing medical evidence, to wasting time and resources invested in the conduct of research”*. The risk of wasting time

⁵ <https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/s13643-015-0070-y>

and resources invested in research is due to the fact that *“Researchers may simply decide against the publication of entire studies, based on whether the results are ‘positive’.* This practice, concerning the entire suppression of a research paper, has been discussed in many studies, with recent estimates indicating that the results of half of clinical trials are never published” (Salandra, 2018). This practice leads to a reporting bias, as already said, that increases the information asymmetry suffered by investors when choosing a project to finance.

Just to give an idea of the magnitude of the phenomenon, the increasing proportion of studies in which at least one outcome is changed or omitted is up to 62% of the investigated trials that had major discrepancies in the outcomes (Dwan et al., 2013). In other words, ‘negative’ or ‘null’ outcomes have lower chances of being reported. *Indeed, “There is also evidence to suggest that study outcomes which are statistically significant are more likely to be published, with estimated odds for publication being two to four times greater than those not reaching significance.⁶”* This behavior makes us understand that it is more likely that positive results are published and negative ones are omitted from publication.

As a consequence, the implications of selective reporting are particularly serious in clinical research: efficacy of a treatment may be overestimated or, even more concerning, adverse effects may be underestimated.

According to Salandra (2018), *“there’s evidence that ‘softer’ fields of science report more positive outcomes, like in Mental Health and Dermatology trials, compared to Oncology. Although no direct measure of hardness is available, certain parameters may reflect theoretical and methodological consensus in a field”.* Inside the category of ‘softer’ fields are included mostly social sciences (such as sociology, psychology and so on) and all the sciences that based their

⁶ <https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/s13643-015-0070-y>

assumptions and evidences not with objectivity that, differently from ‘harder’ science, are usually expressed by mathematical models and verified data. *“In general, the soft sciences deal with intangibles and relate to the study of human and animal behaviors, interactions, thoughts, and feelings. Soft sciences apply the scientific method to such intangibles, but because of the nature of living beings, it is almost impossible to recreate a soft science experiment with exactitude. Some examples of the soft sciences, sometimes referred to as the social sciences, are Psychology, Sociology, Anthropology⁷”.*

It seems recognized by the literature that there are fields of medical research more affected by selective reporting than others, as already explained by (Salandra, 2018). Moreover, focusing on cancer R&D, it is possible to consider that:

1. *“High quality clinical data exists for cancer patients, which accurately tracks patient level characteristics, such as survival time”.*
2. *“The existence of a standardized classification system for cancer – namely standardized cancer organs of origin (such as breast and lung) and stages of cancer at the time of diagnosis (such as localized and metastatic) – facilitates a relatively clean match between aggregated patients level clinical data and information on clinical trial investments relevant to different groups of patients.”* (Budish, Roin, Williams, 2015)

As already anticipated, there are some fields of scientific research that are more affected by selective reporting and we will keep this in mind when choosing the field of our analysis.

⁷ <https://www.thoughtco.com/hard-vs-soft-science-3975989>

2.1.6 Evaluation of Clinical trials

So far, we analyzed the selective reporting of outcomes of clinical trials, but it is appropriate to describe the meaning of the outcome measure of clinical trials. In particular, clinical trials evaluate whether a treatment or a drug provides a clinical benefit to mortality, measured by overall survival or disease-free-survival, which measure times until cancer recurrence. (Budish, Roin, Williams, 2015).

Our work will focus on outcome measures of clinical trials, first of all by building a basic measure of the risk of failure of a clinical trial, through the identification of the most impacting factors on the outcome of a trial. After these settings, we would like to demonstrate that certain investors are more willing to invest in trials with lower risk associated and closer to commercialization.

Literature suggest us that there are some actions that actors are pursuing to minimize the risk. Some of these, consists in the monopolization of a line of research. Others, following the principle of minimization of risk by differentiating their portfolio, adopted a research portfolio that contains projects with varying degrees of uncertainty (Stephan, 1996).

Due to unavailability of data we already described, we decided to use the probability of success of a group of clinical trials as measure of the risk associated with that group. By risk, we mean the risk of failure of the trial outcome, measured by the Probability of Success, so excluding measures of risks like volatility and standard deviation, that as we mentioned required a dedicated structure of data (like historical series) not applicable for trials. For instance, by grouping clinical trials according to their phase, for example, we can obtain the historical probability of success of trials belonging to each phase.

The probability of success (POS) of a clinical trial is critical for clinical researchers and biopharma investors to evaluate when making scientific and economic

decisions. Prudent resource allocation relies on the accurate and timely assessment of risk.

Without up-to-date estimates of the POS, however, investors may misjudge the risk and value of drug development, leading to lost opportunities for both investors and patients (Wong, Siah, Lo, 2018).

We can underline that the estimation of probability of success of a clinical trial is a focal part of our analysis.

Previous estimates of success rates rely on relatively small samples from databases curated by the pharmaceutical industry and are subject to potential selection biases. Using a sample of 3374 entries of clinical trial data from January 1, 1995 to December 31, 2015, we estimate aggregate clinical trial success rates and durations. We also compute disaggregated estimates across several trial features including intervention type, clinical phase, industry or academic sponsor and time.

2.1.7 Research objectives

The first step of our empirical analysis is to measure the success rate for each phase of clinical trials to be able to estimate the risk of failure of trials. It is worth to consider that we extended the estimation of success rate also to other categories, such as the leading sponsor of the trial.

The primary goal of this work is to demonstrate the propensity to low-risk investments of corporations. To further analyze this point, we demonstrate that corporations are willing to invest in phase 3 trials more than other investors, since this phase is the most likely to succeed than the other. One possible explanation of this is that phase 3 trials have already passed the safety and efficacy tests of phase 1 and phase 2, so are the ones closer to commercialization. This reminds us the short-termism of corporations, as discussed in 2.1.3 Short-termism of corporations.

For these reasons, we are going to test these hypotheses:

H1: Corporations are more willing to invest in clinical trials with lower risk of failure than other investors.

To test this hypothesis, which is the principal, we had firstly to test other two sub-hypotheses:

H2: Phase 3 trials are most likely to succeed than other trials of other phases.

H3: Corporations are more likely than other investors to sponsor Phase 3 trials.

Another hypothesis we would like to prove is related to the success of implementing, for a clinical trial, mixed methodologies instead of a single treatment or intervention. By mixed methodologies we indicate a factorial trial, where the aim is to study two or more intervention methods applied alone or in combination (Evans, 2010). It is worth to underline that factorial trials have advantages such as high efficiency (possible to assess more intervention types) and the possibility to assess interaction between the different intervention types (Baker, Smith et al., 2017). From literature emerges that one field of clinical trials is particularly focusing on factorial design, that is the one of cancer treatment. It seems that factorial design is “the gold standard for definitive evaluation of new therapies” (Freidin, Korn, 2017). Considering these points, our aim is to investigate if a mixed methodology intervention type could lead to higher probability of success of the clinical trial, testing the following hypothesis:

H4: A trial with mixed methodologies of intervention is more likely to succeed than trials with a single intervention type or treatment.

2.2 Clinical trials: a theoretical framework

The success of the clinical trial⁸ enterprise relies on the public trust in scientific rigor, transparency, and ethical oversight. NIH is the largest federal funder of clinical trials in the United States, with a \$3 billion annual investment. NIH's role is mostly devoted on strengthening policies for each stage of a clinical trial, from the first collection of funding proposals to the publishing of results.

2.2.1 Phases of clinical trials

Clinical trials involving new drugs are commonly classified into five phases. Each phase of the drug approval process is treated as a separate clinical trial. The drug development process will normally proceed through phases 1 – 4 over many years, frequently involving a decade or longer. If the drug successfully passes through phases 1, 2, and 3, it will usually be approved by the national regulatory authority for use in the general population. Phase 4 trials are performed after the newly approved drug, diagnostic or device is marketed, providing assessment about risks, benefits, or best uses. We can analyze better each Phase⁹:

- Phase 0: The purpose of this phase is to help speed up and streamline the drug approval process. Phase 0 studies may help researchers find out if the drugs do what they're expected to do. This may help save time and money that would have been spent on later phase trials. Phase 0 studies are very small, often with fewer than 15 people, and the drug is given only for a short time. They're not a required part of testing a new drug.

⁸ <https://grants.nih.gov/policy/clinical-trials/why-changes.htm>

⁹ <https://www.cancer.org/treatment/treatments-and-side-effects/clinical-trials/what-you-need-to-know/phases-of-clinical-trials.html>

- Phase 1: Phase 1 studies of a new drug are usually the first that involve people. Phase 1 studies are done to find the highest dose of the new treatment that can be given safely without causing severe side effects. Although the treatment has been tested in lab and animal studies, the side effects in people, “with appropriate health problems and medical histories” (Fink, Kokku et al., 2004), can’t be known for sure. Typically, the sample of people involved in these types of trials is around 20-80 subjects and could last several months. These studies also help to decide on the best way to give the new treatment.
- Phase 2: If a new treatment is found to be safe in phase 1 clinical trials, a phase 2 clinical trial is done to see if it works in certain types of cancer. The benefit the doctors look for depends on the goal of the treatment. It may mean the cancer shrinks or disappears. Or it might mean there’s a long period of time where the cancer doesn’t get any bigger, or there’s a longer time before the cancer comes back. In some studies, the benefit may be an improved quality of life. Many clinical trials look to see if people getting the new treatment live longer than most people do without the treatment. The groups are made by patients with the same type of cancer. A general reference for trial in this phase is to last for 1 or 2 years at maximum, involving hundreds of patients. If enough patients benefit from the treatment, and the side effects aren’t too bad, phase 3 clinical trials are begun.
- Phase 3: Treatments that have been shown to work in phase 2 clinical trials must succeed in one more phase before they’re approved for general use. Phase 3 clinical trials compare the safety and effectiveness of the new treatment against the current standard treatment. This is usually achieved testing with large groups of people (typically 1,000–3,000).

Usually, these trials last for a minimum of one year to a maximum of 4 years.

In the United States, when phase 3 clinical trials show a new drug is more effective or safer than the current treatment, a new drug application (NDA) is submitted to the Food and Drug Administration (FDA) for approval. The FDA reviews the results from the clinical trials and other relevant information. Based on the review, the FDA decides whether to approve the treatment for use in patients with the illness the drug was tested on. If approved, the new treatment often becomes a standard of care, and newer drugs may be tested against it before they can be approved. If the FDA feels that more evidence is needed to show that the new treatment's benefits outweigh its risks, it may ask for more information or even require that more studies be done. The FDA then has up to 10 months to review the application and determine whether to grant marketing approval (Chandra, Garthwaite, Stern, 2017). Anyway, even this step is subjected to uncertainty with unpredictable review times from FDA (Rao, 2015). This contributes to a delay in the commercialization process of a drug or a treatment, increasing the uncertainty of the research profitability.

- Phase 4: This Phase is the only one not necessarily requested by the FDA (Food and Drug Administration) as mandatory, because it's mostly related to safety studies during sales. In fact, it involves directly the company realizing the trial, differently from other phases that are most focused on the safety benefits and usages. These phase 4 studies analyze the safety of the treatment over time. Moreover, they can investigate other issues such as the cost effectiveness of the treatment ¹⁰. There are no general references of duration of these trials under this phase because it can

¹⁰ <https://www.cancer.org/treatment/treatments-and-side-effects/clinical-trials/what-you-need-to-know/phases-of-clinical-trials.html>

change accordingly to the purpose of the company that is performing the trial.

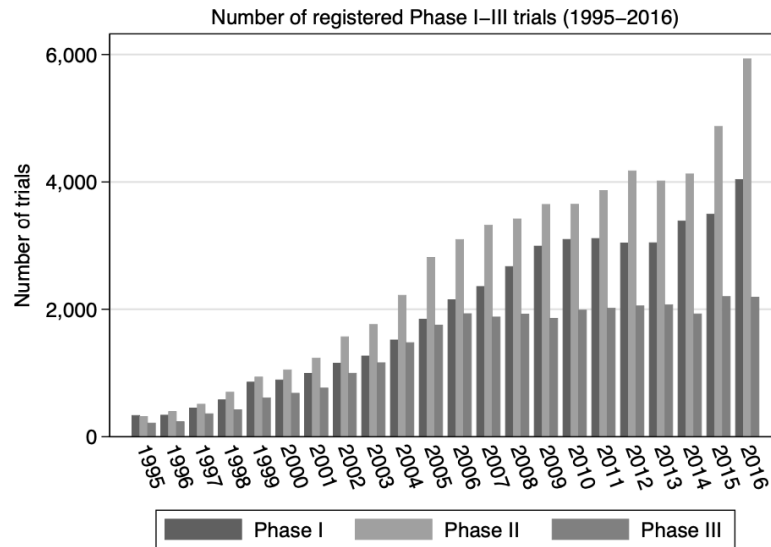


Figure 2 - from “characterizing the drug development pipeline for precision medicine, Chandra, Garthwaite, Stern, 2017”

2.2.2 Sponsor of clinical trials

Sponsor of clinical trial is a person, a company, institution, group or organization that oversees or pays for a clinical trial and collects and analyzes the data¹¹. The key responsibility of sponsors consists of informing local investigators and public authorities of relevant information about the trial (how it’s performed, how data are obtained and collected and so on). This is relevant for the collection of adverse events that, if present, can largely influence the development of the specific treatment or drug under analysis.

“Clinical trials can be funded by private companies – both small privately-financed and large publicly- listed organizations – as well as by

¹¹ <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/clinical-trial-sponsor>

universities/academic medical centers, and by public actors such as the NIH. The latter has historically been more focused on early-stage research, with a particular focus on basic science. This focus stems from the economic role of the NIH as not only the world's largest funder of biomedical research (with nearly \$32.3 billion invested in 2016), but also a provider of public goods in the form of investments in basic research" (Chandra, Garthwaite, Stern, 2017).

2.2.3 LPM (Likely Precision Medicines) trials

Precision medicine is a form of medicine that uses information about a person's own genes or proteins to prevent, diagnose, or treat disease. In cancer, precision medicine uses specific information about a person's tumor to help make a diagnosis, plan treatment, find out how well treatment is working, or make a prognosis¹².

"We identify clinical trials for likely precision medicines (LPMs) as those that use one or more relevant biomarkers. We then further segment trials based on the nature of the biomarker(s) used and other trial features with economic implications" (Stern, Alexander, and Chandra, 2017). This is an innovative frontier. The major constraint of LPMs is that, for each group of patients, it should be developed a personalized treatment. This leads to an increase in operational costs and complexity of running clinical trials. Due to these reasons, private firms may be disincentivized in running such operations. For boosting these practices, under the economic perspective, FDA introduces biomarkers.

"Biomarkers can be used predictively to determine ex ante how likely a given patient is to benefit from therapy. Biomarkers that constitute surrogate endpoints help manufacturers by speeding up clinical trials – e.g. through the use of the FDA's accelerated approval process, whereby a product can be approved

¹² <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/precision-medicine>

on the basis of intermediate patient outcomes that are a good proxy for a therapy’s ultimate effectiveness”.

In Figure 3, we have some examples of biomarkers:

Biomarker type	Official definition	Examples
Diagnostic Biomarker	A biomarker used to detect or confirm presence of a disease or condition of interest or to identify individuals with a subtype of the disease.	<ol style="list-style-type: none"> 1) Sweat chloride may be used as a diagnostic biomarker to confirm cystic fibrosis (Farrell et al. 2008). 2) Glomerular filtration rate (GFR) may be used as a diagnostic biomarker to identify patients with chronic kidney disease (National Kidney Foundation 2002).
Monitoring Biomarker	A biomarker measured serially for assessing status of a disease or medical condition or for evidence of exposure to (or effect of) a medical product or an environmental agent.	<ol style="list-style-type: none"> 1) HIV-RNA may be used as a monitoring biomarker to measure and guide treatment with antiretroviral therapy (ART) (AIDSinfo 2007). 2) Serial measurements of symphysis-fundal height during pregnancy can be used during antenatal screening to detect fetal growth disturbances (Papageorgiou et al. 2016).
Pharmacodynamic / Response Biomarker	A biomarker used to show that a biological response has occurred in an individual who has been exposed to a medical product or an environmental agent.	<ol style="list-style-type: none"> 1) Circulating B lymphocytes may be used as a pharmacodynamic/response biomarker when evaluating patients with systemic lupus erythematosus to assess response to a B-lymphocyte stimulator inhibitor (Stohl and Hilbert 2012). 2) Urinary level of glycosaminoglycans may be used as a pharmacodynamic/response biomarker when evaluating the effect of enzyme replacement therapy for patients with mucopolysaccharidosis type 1 (Jameson et al. 2016).

Figure 3 - from “characterizing the drug development pipeline for precision medicine, Chandra, Garthwaite, Stern, 2017”

“Biomarkers can facilitate a drug market being segmented into identifiable groups based on the expected efficacy of the product, and as a result a segmentation of patients by willingness to pay for the product. When pharmaceutical manufacturers are able to charge only a single price, the existence of known, distinct patient subgroups would effectively allow firms to choose which patients to serve. For example, where the population receiving lower (but positive) value is quite large, the manufacturer may choose to set a low price and sell to a larger market. However, when the lower-value population is quite small, the manufacturer may instead choose a higher price and forgo sales to those patients who derive the least value from the product. Economists will note that this represents the classic monopolist’s dilemma, where

pharmaceutical firms trade margins for quantity". (Stern, Alexander, and Chandra, 2017).

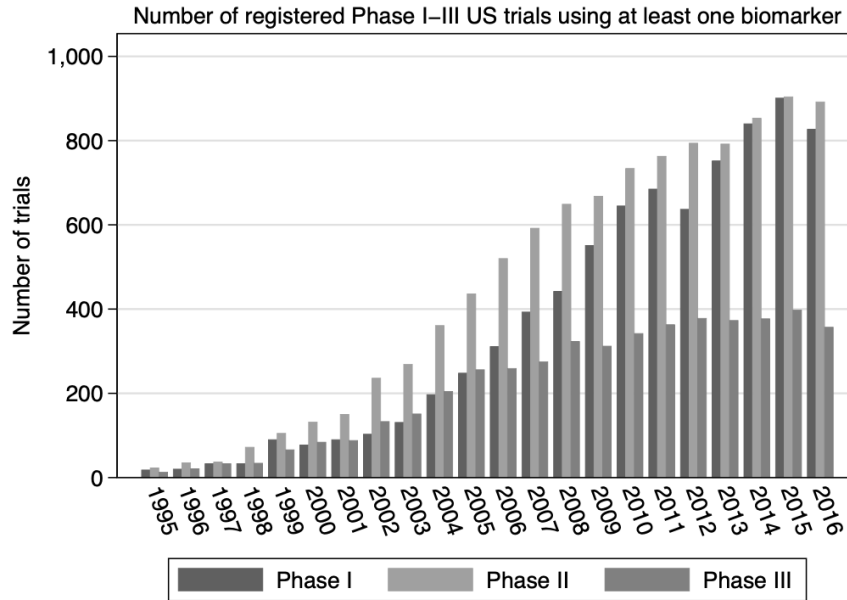


Figure 4 - from "characterizing the drug development pipeline for precision medicine, Chandra, Garthwaite, Stern, 2017"

"For this reason, firms often attempt to find ways to sell the same product to different customers based on their willingness to pay a strategy known as price discrimination. If firms develop a mechanism for charging indication-based prices, the existence of well-established, readily identifiable biomarkers will become an important tool for facilitating price discrimination. When such price discrimination is feasible, the most extreme outcome is that a manufacturer would be able to capture all of the surplus as profits. Depending on the distribution of patients, this could (but need not) expand access to lower-value indications. In a world where a product with a biomarker exists, an indication-based pricing strategy weakly increases the profits of firms" (Stern, Alexander, and Chandra, 2017).

2.3 Interpreting clinical trials

Since at the basis of our work there is the necessity to understand if a clinical trial is a success or a failure, we need to be able to interpret the results provided and then to encode them.

Once we perform a complete description of the elements and characteristics of clinical trials, it is necessary to understand which are the criteria that allow us to interpret a clinical trial and determine the success/failure of it.

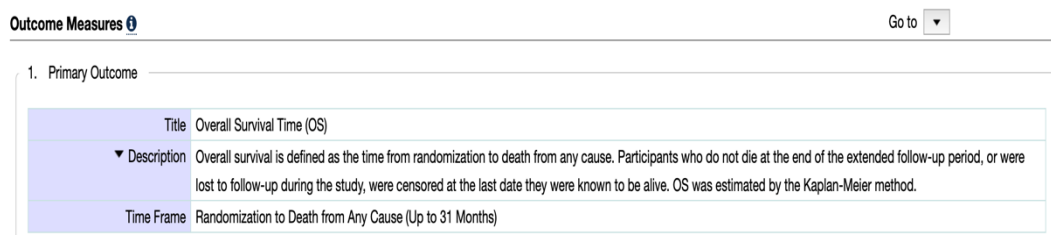
Biostatistics, that is the application of statistical methods to different topics in biology, gives us the tools for interpreting clinical trials.

Basically, since there is not a standard way to evaluate the outcome of clinical trials, we had to create a sort of rule of thumb suitable for our scope.

Indeed, while it's possible to obtain from the trial conductor a unique and standardized encoding about the phase, the intervention type and the lead sponsor, it's not possible to obtain the encoding about the results of the trial. In fact, for each trial we have the kind of results presented, with their unit of measures and sometimes the statistical analysis on the results.

2.3.1 Primary outcome measure

In Figure 5 we report an example of the outcome presented by a trial:



The screenshot shows the 'Outcome Measures' section of a clinical trial on ClinicalTrials.gov. It features a table with three rows: Title, Description, and Time Frame. The table is titled '1. Primary Outcome'.

1. Primary Outcome	
Title	Overall Survival Time (OS)
Description	Overall survival is defined as the time from randomization to death from any cause. Participants who do not die at the end of the extended follow-up period, or were lost to follow-up during the study, were censored at the last date they were known to be alive. OS was estimated by the Kaplan-Meier method.
Time Frame	Randomization to Death from Any Cause (Up to 31 Months)

Figure 5 - Example of Primary Outcome on Clinicaltrials.gov

In this case, the trial reports the primary outcome, that is the Overall Survival Time (OS). It represents the time to death of the patient from any causes in a given timeframe, that in this case is up to 31 months. Each clinical trial determines which are the primary and secondary outcome to measure and to which time frame to refer.

▼ Outcome Measure Data

▼ Analysis Population Description
All randomized participants. Censored participants: Nectinumab + Gemcitabine + Cisplatin = 127, Gemcitabine + Cisplatin = 106

Arm/Group Title	Nectinumab + Gemcitabine + Cisplatin	Gemcitabine + Cisplatin
▼ Arm/Group Description:	Nectinumab + Gemcitabine + Cisplatin Nectinumab: 800 mg I.V. infusion on Days 1 and 8 of every 3 week cycle. Continues until progressive disease, toxicity, noncompliance, or withdrawal. Gemcitabine: 1250 mg/m ² on Days 1 and 8 of every 3 week cycle. Continues for a maximum of six cycles. Cisplatin: 75 mg/m ² IV on Day 1 of every 3 week cycle. Continues for a maximum of six cycles.	Gemcitabine + Cisplatin Gemcitabine: 1250 mg/m ² on Days 1 and 8 of every 3 week cycle. Continues for a maximum of six cycles. Cisplatin: 75 mg/m ² IV on Day 1 of every 3 week cycle. Continues for a maximum of six cycles.
Overall Number of Participants Analyzed	545	548
Median (95% Confidence Interval) Unit of Measure: Months	11.5 (10.4 to 12.6)	9.9 (8.9 to 11.1)

Figure 6 - Example of Outcome Measure Data from Clinicaltrials.gov

Right below, the trial typically reports data about population analysis (Arm/Group Description). In this case, this trial wants to measure the Overall Survival Time (OS) of two groups of participants. Typically, one group is the treatment one, and the other is the control one. It means that the drug or intervention to test is given to the treatment group, and a placebo is given to the control group, to test the difference in outcome between the two different groups, and the effects of the drug/intervention, as a consequence. Anyway, it is necessary to underline that the two groups are made of randomized participants, so no one knows to which group they belong.

In this very case, we don't have a placebo group, but two groups testing different drugs. Indeed, it's important to underline that the universe of clinical trials has a wide variety, in terms of intervention type, outcome measure type, disease, time frame etc.

Under the description of the two groups, we have the sample size, so the number of participants analyzed for each group and the unit of measure of the primary outcome, that in this very case is expressed in months. In particular, we can notice an Overall Survival Time of 11.5 months for group 1 and 9.9 months for group 2.

2.3.2 Statistical analysis

At this point we don't know which is the better result, because qualitatively we can conclude that the group 2 has a better outcome, but not in statistical terms. For this reason, clinical trials report the statistical analysis performed on the outcome, like the one we report in Figure 7:

Statistical Analysis Overview	Comparison Group Selection	Necitumumab + Gemcitabine + Cisplatin, Gemcitabine + Cisplatin
	Comments	[Not Specified]
	Type of Statistical Test	Superiority or Other (legacy)
Statistical Test of Hypothesis	Comments	[Not Specified]
	P-Value	0.0120
	Method	Log Rank
	Comments	[Not Specified]
Method of Estimation	Estimation Parameter	Hazard Ratio (HR)
	Estimated Value	0.842
	Confidence Interval	(2-Sided) 95% 0.736 to 0.962
	Estimation Comments	[Not Specified]

Figure 7 - Example of Statistical Analysis on ClinicalTrials.gov

As we already stated, for each trial there could be different outcome measure data, and consequently different kinds of statistical analysis performed on that. In this very case, the statistical test of hypothesis is the Log Rank, for which it's reported the p-value. Log Rank method is typically used in clinical trials measuring Overall Survival Time, Progression Free Survival Time, Disease Free Survival Time, so the ones that typically perform a survival analysis. Indeed, the log-rank test, or log-rank test, is a hypothesis test to compare the survival distributions of two samples. It is a nonparametric test and appropriate to use

when the data are right skewed and censored (technically, the censoring must be non-informative). It is widely used in clinical trials to establish the efficacy of a new treatment in comparison with a control treatment when the measurement is the time to event (such as the time from initial treatment to a heart attack).

The last observation of the statistical analysis is the method of estimation, that in survival analysis is typically the hazard ratio (HR), which is the ratio of the hazard rates corresponding to the conditions described by two levels of an explanatory variable. For example, in a drug study, if the treated population dies at twice the rate as the control population, the hazard ratio is 2, indicating a higher hazard of death from the treatment¹³.

To sum up, taking into consideration this very case, the null-hypothesis is that the two groups have identical hazard functions, and we also have p-value and confidence intervals reported, but how to interpret this outcome?

For our scope, we need to encode the success/failure of a clinical trial, but it's not already available and ready to use, so we need to understand the logic beyond it and be able to assign an encoding that could be as correct as possible.

It's important to underline that the same structure of the primary outcome of the trial is repeated also for its secondary outcomes, but they are outside of our lens of analysis. Biostatistics can help us for our scope, since it gives some important guidelines to follow to correctly interpret the outcome of the trial. Before the definition of guidelines for interpretation of a clinical trial, it's necessary to set and go deep in some important definitions.

¹³ <https://docs.teradata.com/reader/JtLhZxnZVIJAs8pZG1VVfg/9uN2cAIGzBvB~b4tK5oA1w>

2.3.3 Confidence intervals

Confidence intervals are a way of admitting that any measurement from a sample is subject to errors. Although the estimate given from the sample is likely to be close, the true values for the population may be above or below the sample values. A confidence interval specifies how far above or below a sample-based value the population value lies within a given range, from a possible high to a possible low. The true mean, therefore, is most likely to be somewhere within the specified range (USMLE Step 2, Kaplan Book).

2.3.4 Significance Testing

To test hypotheses, it is necessary to draw a random sample from a population and make an inference. Before the sample is necessary to set a significance level, alpha, which is the risk of error you are willing to tolerate. Usually the level of significance is set at 0,05 and the risk is associated with the rejection of the null hypothesis, even though it is true (e.g. type I error).

Both p-value and alpha represent significance, but the difference is that p-value measures the strength or magnitude of the data against the null hypothesis, whereas alpha level represents risk and is independent of data.

The confidence interval provides a direct, numeric measurement of the imprecision of the estimate of the response to treatment that is due to sampling variability. The larger the sample, the narrower the interval; the larger the variability, the wider the interval (Bigby, Gadenne, 1996).

The acceptance of a significance level of 0.05 as the cutoff for rejecting the null hypothesis is a tradition based on quality control standards and is not an absolute truth (Bigby, Gadenne, 1996).

In Figure 8 we report an example explaining the decision making:

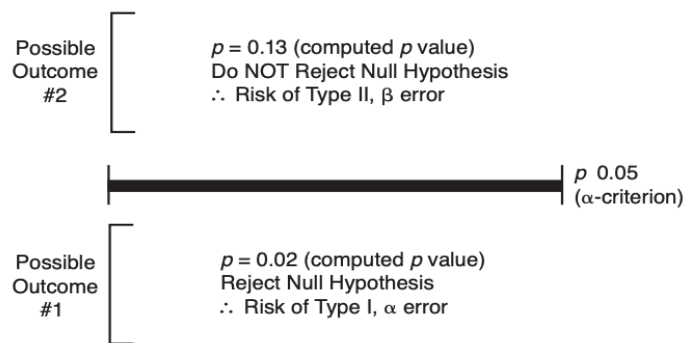


Figure 8 - from USMLE Step 2, Kaplan Book

2.3.5 Types of error

In case of rejection of the null hypothesis, there is no certainty of the trustiness of the assumption. For some reason, the results given by the sample may be inconsistent with the full population. Considering this fact, there are 2 types of error we could make:

- Type I error (alpha error): rejecting the null hypothesis when it is really true, i.e., assuming a statistically significant effect on the basis of the sample when there is none in the population or asserting that the drug works when it does not. The chance of a type I error is given by the p -value.
- Type II error (beta error): failing to reject the null hypothesis when it is really false, i.e., declaring no significant effect on the basis of the sample when there really is one in the population or asserting the drug does not work when it really does (USMLE Step 2, Kaplan Book).

At this point, it is important to underline some relevant facts regarding the p-value. As already explained, it provides criterion for making decisions about the null hypothesis and, as a consequence, allow us to identify a rule of thumb to determine whether the clinical trial was successful or not, exactly in the same way we described before. It's also necessary to state an important characteristic of the p-value, since it tells us statistical significance, not clinical significance or likelihood of benefit. As a consequence, we will always refer to the outcome of clinical trials in terms of statistical significance or not.

The importance of statistical analyses must be kept in proper perspective. Statistics are a tool for trying to ensure that results of clinical trials are not due to chance or sampling variation alone. The combination of hypothesis testing and the use of confidence intervals give a measure of the likelihood that results of a trial are due to chance and the precision of the estimated difference in treatments, respectively. Statistical analyses cannot tell you the medical significance of differences in treatments. In other words, "statistically significant" should not and cannot be equated with "medically significant". (Bigby, Gadenne, 1996).

2.3.6 Further considerations

It's now important to do some considerations about the statistical analysis provided by clinical trials. As we mentioned before, not all the clinical trials report the statistical analysis on the outcome, making it hard for us to analyze the outcome and classify it. It's a quite diffused and ancient problem, as mentioned in "*Understanding and evaluating clinical trials*, Bigby, Gadenne, 1996": "Surprisingly, statistical analyses are often omitted from published clinical trials".

Another important remark to do is that simply stating that the difference was statistically significant ($p < 0.050$ does not constitute an adequate description of the statistics used). The exact procedure used (e.g., t-test or X²) and the results obtained must be specified (Bigby, Gadenne, 1996). Indeed, in the example provided in this chapter, the procedure used was the survival analysis, but sometimes, the unavailability of data didn't permit us to properly perform the encoding of the outcome of the trial.

For which concerns the confidence interval provided in the study, in particular in comparative studies, confidence intervals should be reported for the differences between groups, not for the results of each group separately. The advantage of using confidence intervals instead of or in addition to p-values is that confidence intervals provide an indication of the size of the differences in treatments and give numeric measurements of the inexactness in our knowledge of the real differences in a treatment. (Bigby, Gadenne, 1996).

Chapter 3 - Database Creation

3.1 Data Source

In every database creation, the most critical part is the one regarding the data source. The presence of clear and unique data is fundamental for testing hypotheses without incurring in major bias or mistakes. In particular, in this section we wish to test if, in presence of certain characteristics of trials (like for example the presence of a certain sponsor or a certain treatment applied), the success or the failure of this is statistically influenced.

For our scope, we need a big sample of data that can give us variance and a data structure that could be standard or in a format that is suitable to enable statistical investigation of the observational variables.

For these reasons, we defined Clinicaltrials.gov as our data source, the database for trials of the US government. The database is public, comprehensive and is maintained by the NIH (National Institute of Health). The choice of this data provider could be also explained by its evolution in time

(Chandra, Garthwaite, Stern, 2017): the database has been created by virtue of a law of the US congress, passed in 1997, called FDAMA (Food and Drug Administration Modernization Act), requiring the NIH to create a public information resource on certain clinical trials regulated by the FDA. In compliance with the law, NIH released in 2000 the Clinicaltrials.gov website. The completeness of the database was further reinforced by a resolution of the WHO (World Health Organization), which in 2006 stated that all clinical trials should be registered, and it identified a minimum trial registration dataset of 20 items. The last important milestone was registered in 2008, where Clinicaltrials.gov began allowing sponsors and principal investigators to submit the results of clinical studies.

Clinicaltrials.gov was chosen as the main and unique data source due to its completeness and the presence of a standardized and recurrent format of data divided into keywords, which allows us to identify the most important and significant data to be analyzed. Indeed, an important aspect of the data source selection was the presence of a data universe of around 350.000 different trials, registered from 1964 (it's the date of the first published clinical trial). Again, our scope was to have a high-variety database with a large selection of data, but at the same time with a standardized structure based on a keyword categorization, in order to allow us to automatically extract huge sets of data.

3.1.1 Available information

The database has a graphical web-based interface. Each clinical trial is a record in the database. For each record, the interface presents the following information, as shown in Figure 9:

First-line Treatment of Participants With Stage IV Squamous Non-Small Cell Lung Cancer With Necitumumab and Gemcitabine-Cisplatin (SQUIRE)

ClinicalTrials.gov Identifier: NCT00981058

Recruitment Status: Active, not recruiting
First Posted: September 22, 2009
Results First Posted: June 27, 2016
Last Update Posted: July 28, 2020

Sponsor:
Eli Lilly and Company

Collaborators:
Parexel
PPD
Medidata Solutions
Laboratory Corporation of America
University of Colorado, Denver
Thermo Fisher Scientific
ICON Clinical Research
Pacific Biomarkers
Symanex Inostics GmbH
Intertek

Information provided by (Responsible Party):
Eli Lilly and Company

Figure 9 - Example of a trial on the platform

- In the top of the page there is the title of the trial of reference with the first description of treatments/drugs used.

- In the red box at the top-right of the page there is the recruitment state, so in which state is the trial, if it's still recruiting volunteers, if it's completed or in an unknown state (so probably abandoned or dismissed).
- At the top-left of the page, right under the title, there is the reference of the sponsor, and all the collaborators or, if present, the responsible party.

Every record presents information concerning the content and the result of the trial. This is divided into three main sections, each one graphically represented on the database as a flyer:

1. Study Details: In the first flyer there is the summary of the trial, so its description and procedure, all the possible references in terms of institution involved, important dates in the course of the trial, its primary and secondary outcomes and other relevant information.
2. Tabular View: in this flyer there is a table synthesizing all the information of the methods used for the trial.
3. Study Results: this last flyer contains all the numbers and information regarding the trial, so every results of the outcomes, the information on the sample of people used to test these, and all the statistical analysis used to test results, as shown in Figure 10.

Outcome Measures 6 Go to

1. Primary Outcome

Title		Overall Survival Time (OS)	
Description: Overall survival is defined as the time from randomization to death from any cause. Participants who do not die at the end of the extended follow-up period, or were last to follow-up during the study, were censored at the last date they were known to be alive. OS was estimated by the Kaplan-Meier method.			
Time Frame: Randomization to Death from Any Cause (Up to 31 Months)			
Outcome Measure Data			
Analysis Population Description			
All randomized participants. Censored participants: Nectumumab + Gemcitabine + Cisplatin = 127, Gemcitabine + Cisplatin = 106			
Arm/Group Title	Nectumumab + Gemcitabine + Cisplatin	Gemcitabine + Cisplatin	Gemcitabine + Cisplatin
Arm/Group Description:	Nectumumab + Gemcitabine + Cisplatin Nectumumab: 800 mg I.V. infusion on Days 1 and 8 of every 3 week cycle. Continues until progressive disease, toxicity, noncompliance, or withdrawal. Gemcitabine: 1250 mg/m ² on Days 1 and 8 of every 3 week cycle. Continues for a maximum of six cycles. Cisplatin: 75 mg/m ² IV on Day 1 of every 3 week cycle. Continues for a maximum of six cycles.	Gemcitabine + Cisplatin Gemcitabine: 1250 mg/m ² on Days 1 and 8 of every 3 week cycle. Continues for a maximum of six cycles. Cisplatin: 75 mg/m ² IV on Day 1 of every 3 week cycle. Continues for a maximum of six cycles.	Gemcitabine + Cisplatin Gemcitabine: 1250 mg/m ² on Days 1 and 8 of every 3 week cycle. Continues for a maximum of six cycles. Cisplatin: 75 mg/m ² IV on Day 1 of every 3 week cycle. Continues for a maximum of six cycles.
Overall Number of Participants Analyzed	545		548
Median (95% Confidence Interval)			
Unit of Measure: Months	11.5 (9.4 to 12.6)		8.9 (8.9 to 11.1)
Statistical Analysis 1			
Statistical Analysis Overview	Comparison Group	Nectumumab + Gemcitabine + Cisplatin, Gemcitabine + Cisplatin	
	Selection	[Not Specified]	
	Comments	[Not Specified]	
	Type of Statistical Test	Superiority or Other (region)	
	Comments	[Not Specified]	
Statistical Test of Hypothesis	P-Value	0.0120	
	Comments	[Not Specified]	
	Method	Log Rank	
	Comments	[Not Specified]	
Method of Estimation	Estimation Parameter	Hazard Ratio (HR)	
	Estimated Value	0.842	
	Confidence Interval	0.5068-0.9516	
	Estimation Comments	[Not Specified]	

Figure 10 - Study results

As shown in Figure 10, results are divided for outcomes (in this case results are expressed for the Primary Outcome Overall Survival Time). Inside the box “Analysis Population Description” there are two columns¹⁴: one is usually used for the main treatment under analysis (like a new drug or a new usage of this), and the other refers to the Placebo to be compared with. For each column, the Overall Number of Participants Analyzed and the result of the trial are expressed. If present, immediately after this box, there is the one dedicated to statistical analysis, which was fundamental to perform statistical analysis on trials, as explained in the Chapter 4 – Data Analysis and Interpretation.

As we can see, for each trial, the set of data presents is various and large. Of this set, for our analysis, the key aspect was in selecting what was really meaningful. According to our research scope and our hypothesis, the most important data to extract, are the following:

- NCTid: unique alphanumeric identifier of the trial on the platform.

¹⁴ Anyway, this structure is not always present; there could be trials with a single column, so just analyzing the main treatment, or even with multiple ones.

- Start Date: starting date of the trial.
- Primary Completion Date: effective date in which the primary outcome of the trial was completed. In case of multiple primary outcomes, it's the date in which they are all completed.
- Completion Date: effective date in which all the outcomes (primary and secondary) of the trial are completed.
- Phase: phase of progression of the trial. Each phase typology was further analyzed in paragraph 2.2.1 Phases of clinical trials.
- Lead Sponsor Name and Class: name of the primary sponsor of the trial and its belonging class (E.g. Gynecologic Oncology Group is a main sponsor under the class of Networks, while Bayer is a sponsor under the class of Industries).
- Intervention Type: classification of the treatment adopted in the trial, based on the type of drug, medical treatment or combination thereof (E.g. chemotherapy, radiotherapy, immune therapy).
- Primary Outcome Measure and Value: under the measure there could be identified the purposes of the trial, and in particular what they want to analyze. The value of the primary outcome is simply the result of this analysis (E.g. with reference to the Figure 10 the Primary Outcome Measure was the Overall Survival Time, with results of 11.5 months within a confidence level of 95%).
- P-Value: it is the P-Value associated to the null hypothesis tested with the statistical analysis and performed on the Primary Outcome.

- Statistical Method: it is the type of the statistical analysis performed to assess results. Examples of this are Cox Regression Analysis, Log-Rank, Chi-Squared and so on.
- Parameter Type and Value: it's the type of parameter chosen to assess the statistical validity of the analysis, and the corresponding value. Examples of this are Hazard Ratio, Percentage of Participants and so on.
- Confidence Interval (CI) Value, Lower and Upper Bound: this is the value of the confidence interval chosen to perform the statistical analysis, in addition to the values of lower and upper bound of this.

3.2 Sample Selection

First of all, we had to set a therapeutic area to analyze inside clinical trials. In order to select it, we evaluated some aspects like the quality of data reported, the quantity of clinical trials performed, and all the premises made in paragraph 2.1.5 Selective Reporting. Following these considerations, we decided to perform our analysis on cancer clinical trials. Inside this field, the outcome measured by clinical trial can vary according to the scope of the research. To standardize our analysis, and obtain a common frame of results to study, we selected only a subset of outcome measures. In particular, we are going to consider the ones that are more important and frequent in cancer research, that are the ones regarding survival analysis.

To understand which are the most important outcome measures related to overall survival, we extracted a sample database of trials to look at the most recurrent measure comprised under the primary outcome analysis, finding that OS (Overall Survival, defined as the time from registration to death, or censored

at last date known alive. Kaplan-Meier method was used to estimate the overall survival rate at 24 months) and PFS (Progression-Free-Survival, defined as the time from randomization to date of first documented PD or date of death, whichever occurred first) were the most frequent ones, and also the ones that present the highest number of trials completed and with data to be analyzed. Anyway, this sample database was related to our choice to run the analysis only focusing on primary outcomes, because they were the most significant in respect to secondary measures.

Another important selection was the one regarding the time frame of our analysis. The sample we created included all clinical trials registered between 1995 and 2015, related to the treatment of cancer. These amounted to 3374 different records. The key rationale behind this decision, is that there are two potential censoring problems in the data: first, data prior to 1995 present missing information or the complete absence to the registered portals, probably due to the absence of the compulsory publication of every trial (which was declared in 1997 by the FDAMA); second, recent trials are in the database, but their results are often incomplete, either because the trial is still in progress or because it's closed, but results have not been yet reported. In order to overcome this problem, we have limited the selection to trials with a Start Date within 2015.

3.3 Data Extraction and Cleaning

In order to automatically extract the data needed, we used the Application Programming Interface (API) of the Clinicaltrials.gov. This specific function provides a toolbox for programmers and other technical users to access all posted information on Clinicaltrials.gov study records data. The API is designed for encoding simple and complex search expressions, and parameters in URLs.

Thanks to this function, we were to extract the fields needed. This is the syntax used to download data is showed in Figure 11:

Study Fields Request

Specify the [query parameters](#) for a [Study Fields query URL](#):

Query URL: /api/query/study_fields

Search Expression: expr= cancer AND AREA[StartDate]RANGE[January 2015, February 2015]

[\(See API Search Expressions and Syntax\)](#)

Study Fields: fields= NCTId,BriefTitle,StartDate,PrimaryCompletionDate,CompletionDate,Phase,LeadSponsorName,LeadSponsorClass,InterventionType,PrimaryOutcomeMeasure,OutcomeMeasurementValue,OutcomeAnalysisGroupID,OutcomeAnalysisNonInferiorityType,OutcomeAnalysisPValue,OutcomeAnalysisStatisticalMethod,OutcomeAnalysisParamType,OutcomeAnalysisParamValue,OutcomeAnalysisCIPctValue,OutcomeAnalysisCILowerLimit,OutcomeAnalysisCIUpperLimit

[\(See Available Study Fields\)](#)

Minimum Rank: min_rnk=

Maximum Rank: max_rnk=

Format: fmt= csv

Figure 11 - Example of API syntax on ClinicalTrials.gov

As shown in the Figure 11, the API interface allows us to extract data according to our specific filters. In particular:

- Search Expression: in this example, we extracted all the trials containing the cancer keyword, having the start date between January and February 2015.
- Study Fields: in this field we insert all the keywords we wanted to download from the database. Each keyword corresponds to a column in the Excel file of destination.

After the extraction, the structure of the database obtained was like the one represented in Figure 12:

Rank	NCTId	Brief Title	Start Date	Primary Completion Date	Phase	Lead Sponsor	Lead Sponsor	Intervention	Primary Outcome	Measurement	Group Id	Non Inferiority	PValue	Statistical Meth	Param Type	Param Value	CI Lower	CI Upper
1	NCT01074424	Studying Bio	February 2010	March 2010	NC	NA	Gynecologic NETWORK	Genetic Oth	Generation c	NC	NC	NC	NC	NC	NC	NC	NC	NC
2	NCT01157962	Sentinel Con	January 2011	January 2011	Not Applicable	Charite Univ	OTHER	Procedure	overall survi	NA	NA	NA	NA	NA	NA	NA	NA	NA
3	NCT01074398	Biomarkers i	February 2010	March 2010	NC	NA	Gynecologic NETWORK	Genetic Ger	Association I	NC	NC	NC	NC	NC	NC	NC	NC	NC
4	NCT01367353	Characteriza	January 2011	December 21	December 21	Not Applicable	National Tai	OTHER	Procedure	overall survi	NA	NA	NA	NA	NA	NA	NA	NA
5	NCT01056809	Treatment S	January 2011	May 2012	May 2012	Not Applicable	Vrinnevi Hos	OTHER	Procedure	overall survi	NA	NA	NA	NA	NA	NA	NA	NA
6	NCT01086618	Chemothera	January 2011	January 2011	July 2013	Phase 2	Pha University	OTHER	Drug Proc	Accrual rate	NA	NA	NA	NA	NA	NA	NA	NA
7	NCT03648151	Influence of	January 1, 21	December 3,	December 3,	NA	The First Affi	OTHER	NA	Overall survi	NA	NA	NA	NA	NA	NA	NA	NA
8	NCT00981058	First-Line Tre	January 7, 21	June 17, 2011	July 31, 2021	Phase 3	Eli Lilly and	INDUSTRY	Biological D	Overall Survi	11.5 9 9 5 7	OG000	OG0 Superiority < 0.0120	0.02	Log Rank Log R	Hazard Ratio 0.842	0.85195 95 95 90.736	0.74 0.962 0.975
9	NCT02423278	The Long-ter	January 2011	June 2016	December 21	Phase 2	Pha First Affili	OTHER	Procedure F	Overall Survi	NA	NA	NA	NA	NA	NA	NA	NA
10	NCT01102517	A Trial on Vic	January 2011	December 21	December 21	Not Applicable	Sun Yat-sen	OTHER	Procedure F	Disease-free	NA	NA	NA	NA	NA	NA	NA	NA
11	NCT01392981	An Observati	February 2010	October 2011	October 2011	NA	Hoffmann-L	INDUSTRY	NA	Median Over	2.51 0.85	0	NA	NA	NA	NA	NA	NA
12	NCT01075555	Sorafenib To	February 2010	September 2	September 2	Phase 3	Federation F	OTHER	Drug Drug	Overall survi	NA	NA	NA	NA	NA	NA	NA	NA
13	NCT01076751	Observation	February 2010	October 2011	October 2011	NA	Sanofi	INDUSTRY	NA	Overall survi	NA	NA	NA	NA	NA	NA	NA	NA
14	NCT01078311	Drug Monit	February 2010	February 2010	February 2010	NA	South West	OTHER	NA	To demonst	NA	NA	NA	NA	NA	NA	NA	NA
15	NCT01086345	Radiosurger	February 2010	December 21	December 21	Early Phase 1	Case Compr	OTHER	Radiation B	Overall survi	NA	NA	NA	NA	NA	NA	NA	NA
16	NCT01095523	Prospective	January 14, 1	February 9, 1	February 9, 1	Phase 2	National Car	NIH	Procedure	To compare	NA	NA	NA	NA	NA	NA	NA	NA
17	NCT03879395	Surgery for	January 1, 21	December 3,	December 3,	NA	Sahlgrenska	OTHER	Procedure	Overall survi	NA	NA	NA	NA	NA	NA	NA	NA
18	NCT03978039	National Clin	January 2011	December 21	December 21	NA	Institut Berg	OTHER	Drug Drug	Time to next	NA	NA	NA	NA	NA	NA	NA	NA
19	NCT01069801	A Study of W	January 2011	December 21	July 2015	Phase 3	Hoffmann-L	INDUSTRY	Drug Drug	Overall Survi	43 75 293 2	OG000	OG0 Superiority < 0.0001	< 0	Log Rank Log R	Hazard Ratio 0.37	0.26 0.26 0.20	0.55 0.33
20	NCT01095993	Efficacy and	January 2011	July 2012	July 2012	Phase 3	Abbott	INDUSTRY	Drug Drug	Overall Survi	NA	NA	NA	NA	NA	NA	NA	NA
21	NCT01587430	Anthracycl	January 2011	April 2014	January 2011	Phase 4	National Res	NETWORK	Drug Drug	overall survi	NA	NA	NA	NA	NA	NA	NA	NA
22	NCT00893999	Study of Che	February 2010	July 2020	July 2020	Phase 3	University of	OTHER	Drug Drug	Overall survi	NA	NA	NA	NA	NA	NA	NA	NA
23	NCT04273215	Overweight	January 1, 21	December 3,	December 3,	NA	Coordinació	OTHER_GOV	Other	Overall survi	NA	NA	NA	NA	NA	NA	NA	NA
24	NCT01266668	The Impact	February 2010	May 2011	June 2011	NA	Chonbuk Na	OTHER	NA	Overall Survi	NA	NA	NA	NA	NA	NA	NA	NA
25	NCT01055197	Radiation Th	March 2010	March 2015	December 21	Phase 2	Radiation Th	NETWORK	Radiation R	Overall Survi	60.1 50.8 24	OG000	OG0 Superiority < 0.2103	0.24	Log Rank Fish	Hazard Ratio 1.44	95 0.82	2.53
26	NCT01103323	Patients Wit	April 2010	July 2011	January 2011	Phase 3	Boyer	INDUSTRY	Drug Drug	Overall Survi	196 151 159	OG000	OG0 Superiority < 0.009178	< 0	Log Rank Log R	Hazard Ratio 0.774	0.49 95 95 90.636	0.41 0.942 0.582

Figure 12 - Final Database Structure

From the raw data extrapolated from the software, few adjustments were required in order to have a standardized and manageable database. In particular, the most critical one was the management of void spaces in the database. This situation was frequent whenever a trial had some missing data into its information. To solve these unfilled gaps in our lines of data, we used two types of approach:

1. For those trials where the Completion Date was not present (so there was a blanket cell in the dedicated space), we inserted a NC (Not Completed) value both in the Completion Date cell, both in the subsequent cells of values (trials are disposed on a single row, where each column shows a different field associated to the same NCTid).

Rank	NCTId	Brief Title	Start Date	Primary Completion Date	Phase	Lead Sponsor	Lead Sponsor	Intervention	Primary Outcome	Measurement	
1	NCT01074424	Studying Bio	February 2010	March 2010	NC	NA	Gynecologic NETWORK	Genetic Oth	Generation c	NC	
2	NCT01157962	Sentinel Con	January 2011	January 2011	Not Applicable	Charite Univ	OTHER	Procedure	overall survi	NA	
3	NCT01074398	Biomarkers i	February 2010	March 2010	NC	NA	Gynecologic NETWORK	Genetic Ger	Association I	NC	
4	NCT01367353	Characteriza	January 2011	December 21	December 21	Not Applicable	National Tai	OTHER	Procedure	overall survi	NA

Figure 13 - Example of NC Attribution

This solution was adopted because, in case of absence of Completion Date, the trial could be considered not completed (this is true for our assumption of taking only trials with a Start Date before 2015), and so all the values related to that

trial could be considered as not completed (for example, the Primary Outcome Measurement).

2. For those trials where there was a Completion Date, but the total or partial absence of values in the subsequent cells, we replaced blanket cells with NA (Not Available), because the trial should be completed, but still there are absence of information.

Rank	NCTid	Brief Title	Start Date	Primary Com	Completion D	Phase	Lead Sponso	Lead Sponso	Intervention	Primary Outc	Measurement
1	NCT01074424	Studying Bio	February 20	March 2010	NC	NA	Gynecologic	NETWORK	Genetic Oth	Generation c	NC
2	NCT01157962	Sentinel Con	January 201	January 201	January 201	Not Applicat	Charite Univ	OTHER	Procedure	overall survi	NA
3	NCT01074398	Biomarkers i	February 20	March 2010	NC	NA	Gynecologic	NETWORK	Genetic Ger	Association I	NC
4	NCT01367353	Characteriza	January 201	December 2	December 2	Not Applicat	National Tai	OTHER	Procedure	overall survi	NA

Figure 14 - Example of NA Attribution

Another correction, less frequent than the previous one here described, was present due to how we built our database. Selecting trials having Primary Outcome Measure of Overall Survival or Progression-Free-Survival, a possible mistake could be made if a single trial present, as Primary Outcome Measures, both OS and PFS. In this case, the trial is double counted. To solve this problem, we decided to keep only one row of the trial, while before the correction there could be two rows with the same NCTid (so a copy of the same trial).

Each time we had to code a specific field, data needed to be adjusted as well. In particular, if there was, for a single trial, the presence of more than one Primary Outcome, data were showed grouped in a single cell, divided by the term “|”:

Rank	NCTid	Brief Title	Start Date	Primary Com	Completion D	Phase	Lead Sponso	Lead Sponso	Intervention	Primary Outc	Measurement
1	NCT01074424	Studying Bio	February 20	March 2010	NC	NA	Gynecologic	NETWORK	Genetic Oth	Generation c	NC
2	NCT01157962	Sentinel Con	January 201	January 201	January 201	Not Applicat	Charite Univ	OTHER	Procedure	overall survi	NA
3	NCT01074398	Biomarkers i	February 20	March 2010	NC	NA	Gynecologic	NETWORK	Genetic Ger	Association I	NC
4	NCT01367353	Characteriza	January 201	December 2	December 2	Not Applicat	National Tai	OTHER	Procedure	overall survi	NA

Figure 15 - Example of Aggregation in the same cell

For example, in the first trial, more than one Intervention Type were present at the same time due to the multiplicity of Primary Outcomes. The fact is that, for example, when we need to code Intervention Type, we need to have all the values on a single row, not in a single cell. The solution for this task was to explode cells and obtain more than one Intervention Type column.

3.4 Data Encoding

After completing the data collection, the situation was the following: 3374 unique trials (one for each row in the database), each one with 18 trial attributes (one for each column of the database). In order to analyze the data, we had to perform a data-coding to identify some of the fields of interest not readily available from the data. Our purpose in this phase was to categorize the information, such that each category in the field of interest was represented by a numerical value.

The first encode we applied to the database was related to the outcome variable, describing the statistical success of the trial. In particular, for each trial, according to its P-Value, a binary classification was obtained in this way:

- Statistically successful P-Value (value 1 in the encode): the trial achieved the statistical validity in its analysis. The number of trials under this category are 101 over 3374 (3.0% of successful trials). In this category are present trials whose outcome involves rejecting the null hypothesis of no difference between treated and untreated with a statistically significant level of confidence.
- Statistically unsuccessful P-Value (value 0 in the encode): the trial did not achieve statistical validity or has a missing information in the P-Value field

(so the trial never reached the statistical analysis phase). The number of trials under this category are 3273 over 3374 (97.0% of unsuccessful trials).

Reference	Success Encode
1	Statistically Successful
0	Statistically Unsuccessful

Table 1 - Boolean encoding

Next, we moved-on to encoding the main trial attributes, including Phase, Sponsor, State of the trial, Intervention Type and 5-years period in which the trial has started (for example, if the trial has a Start Date between 1995 and 2000, it is the first category, from 2001 to 2005 in the second, and so on). Before entering in detail in each one of these, an important consideration needs to be underlined. For each of these five attributes, we encoded with an increasing numerical value all the possible categories under that specific sector. One problem encountered is that several categories, especially the small ones, had no variance in the associated outcome variables. For example, all Phase 1 trials were non-successful. As a consequence, we combined several of the categories coded in larger ones, or in residual groups (“Others”). An example is represented in Figure 16:

Type_cod_Phase			Code_Stats_Phase	under the voice "Other" there are Phases with missing information
0	NA	empty	1	Phase 2
1	Early Phase 1	empty	2	Phase 3
2	Phase 1	empty	3	Phase 1 Phase 2
3	Phase 2		4	Other: NA, Early Phase 1, Phase 1, Phase 4, Phase 2 Phase 3
4	Phase 3			
5	Phase 4	empty		
6	Phase 1 Phase 2			
7	Phase 2 Phase 3	empty		

Figure 16 - Phase Encoding

In this case, all the categories with empty aside had the absence of trial statistically successful, so can be grouped in the code 4 of the right table, which was the encode used to perform the regression. This procedure was repeated also for the other fields. Let’s go deeper now in each field:

1) **Type_Cod_Phase**: under this encode the Phases of each trial were coded 8 different categories, each one corresponding to the performed Phase of the trial under analysis. In particular, the code 0 was present for those trials in which there was a blanket space (so a missing information) in the column of Phases, while those with two information (as example Phase 1 | Phase 2) were those trials that advanced from Phase 1 to Phase 2, so evolved from one Phase to the successive one. In Table 2 there is the situation before the adjustment considering the statistical success or not, while in Table 3 there is the new coding after the adjustment:

Reference	Type_cod_Phase
0	NA
1	Early Phase 1
2	Phase 1
3	Phase 2
4	Phase 3
5	Phase 4
6	Phase 1 Phase 2
7	Phase 2 Phase 3

Table 2 - First Phases Encoding

Reference	Code_Stata_Phase
1	Phase 2
2	Phase 3
3	Phase 1 Phase 2
4	Other: NA, Early Phase 1, Phase 1, Phase 4, Phase 2 Phase 3

Table 3 - Final Phases Encoding

The situation in this code is the following: the most densely populated attributes are the first two, so Phase 2 and Phase 3, with respectively 40.3% and 30% of

trials, while the last two are respectively 6.5% and 23.1% of the total (which is always 3374 unique trials).

- 2) **Type_Cod_5Y**: here, according to which was the starting date of each trial, having a database on a 20 years time frame, we grouped trials on a 5-year period, obtaining four different codes:

Reference	Code_Stata_5Y
1	1995-2000
2	2001-2005
3	2006-2010
4	2011-2015

Table 4 - Summary of 5Y Encoding

This encoding, differently from the others, did not require adjustments, because there were statistically successful trials in each of the four categories. Being always present the Start Date as information of the trial, there are not missing values in this encode.

As predictable, the last two attributes are the biggest ones in terms of population (due to the growing trend of reporting trials and the highest expenditures on this sector), with respectively 34.1% and 44.4%, while the first two attributes show a density of 6.3% and 15.2%.

- 3) **Type_Cod_Sponsor**: to group different Sponsors under the same category, we had to code a little more in respect to the previous coding. In this case, we had to build a code that looks for keywords in the information presents under Sponsor name (for example, if the Sponsor was Sun Yat-sen university, the code finds the keyword University, so automatically define that Sponsor under the collegial institutions, while for Gynecologic Oncology Group, the keyword of reference was Group). If

in the Sponsor name column there was OTHER, another column was extrapolated from Clinicaltrials.gov, called Lead Sponsor Class, in which four possibilities were present: Network, Other, Industry and NIH (National Institute of Health). Following this logic, we were able to identify 6 different categories showed in Table 5, which became 4 after the statistical reclassification in Table 6:

Reference	Type_Cod_Sponsor
1	Group and Network
2	Inc., Corporation, Industry
3	University, College
4	Hospital
5	Institute
6	Centre

Table 5 - First Sponsor Encoding

Reference	Code_Stata_Sponsor
1	Group and Network
2	Inc., Corporation, Industry
3	Institute
4	Others: Centre, Hospital, University, College

Table 6 - Final Sponsor Encoding

Here the situation is more balanced in respect to other coding, in fact the first, the second and the last attributes present the 28.7%, the 27.4% and the 35.7% of all trials, while the third code only the 8.2%.

- 4) Type_Cod_Int_Type: for this coding, the logic used is the same as Type_Cod_Sponsor, we had to look for keywords in the column related to information on the Intervention Type. The only difference is that, while for Sponsor there was just one information (so just one Sponsor name),

here trials could present more than one Intervention Type, one for each Primary Outcome to realize. That’s why we encode in this way: if there was just one Intervention Type, or in case of homogeneous multiplicity (like for example a trial which has 4 Primary Outcomes all using Drugs), the code could be univocally identified, while in case of multiplicity but with different treatments (for example a trial with 4 Primary Outcomes, but 2 using Drugs, one using Procedure and one Radiation), we created a classification called Mixed Methodologies. As usual, if there was missing information under the Intervention Type column, the code adopted is NA (Not Available, so missing information).

Reference	Type_Cod_Int_Type
0	NA
1	Behavioral
2	Biological
3	Combination Product
4	Device
5	Diagnostic Test
6	Dietary Supplement
7	Drug
8	Genetic
9	Procedure
10	Radiation
11	Other
12	Mixed Methodologies

Table 7 - First Encoding of Intervention Type

Reference	Code_Stata_Int_Type
1	Biological
2	Drug
3	Mixed Methodologies
4	Others: NA, Behavioral, Combination Product, Device, Diagnostic, Genetic, Procedure, Radiation, Other, Dietary

Table 8 - Final Encoding of Intervention Type

Here there is a large predominance of drug treatment, present in the 51.9% of trials, while the other categories have a density of, respectively, 2.9%, 27.2% and 18.0%.

5) Type_Cod_Trial_State: the objective of this encode was to define whether the trial under analysis is abandoned, still in time to be completed or statistically successful or not. In order to define in which category the trial is, we used as reference the theoretical completion time¹⁵ of the Phase in which the trial is, adding one year to consider possible delays (for example, a trial in Phase 3 has a theoretical duration comprised between 1 and 4 years, so we kept as maximum reference of time 5 year). This theoretical duration was then compared with the difference in year between the Primary Completion Date and the Start Date. Following this ratio, these are the possible situations in which a trial could be:

- Abandoned: a trial is presumed abandoned in two different cases. The first case is when the trial has neither a Primary Completion Date, nor a Completion Date, and the Start Date was before 2015. Given that the maximum time available for a trial completion is 5 years, so the

¹⁵ Theoretical Phase durations were taken from the paper “Characterizing the drug development pipeline for precision medicines”, Chandra, Garthwaite, Stern, 2017

maximum of 4 years, we consider 1 year of time buffer for possible delays. The second case is present when the trial presents no information of its Phase (NA is the value of the Phase in this case). If so, we cannot identify a maximum theoretical duration. The presence of abandoned trials is around 11.6% of the total.

- Coded with Success: is when the trial presents a statistical analysis and it results as statistically successful. This encode is the same as having a 1 encode under Success Encode, in fact these are always 101 over 3374 trials (3.0%).
- Coded with Unsuccess: opposite reasoning of the previous point, the same of having a 0, which are present in the 7.7% of total cases.
- Incomplete: a trial could be incomplete if there is the presence of some values or numerical results in the trial information extrapolated from Clinicaltrials.gov. Trials in this category can be distinguished from the others because, even if are not completed, does not present a full row of NA values in its information, as are the net two encodes. This encode counts the 14.9% of total trials.
- ODT (Omission Data Trial) ongoing: this is the case of a trial that presents a full row of NA values, but still is theoretically in time, because has a difference between the Primary Completion Date and the Start Date less or equal to the theoretical duration. In addition to these trials, were also included in this category trials with a Start Date in 2015 and with a future completion date, so planned in 2020 or successive years (so these can be considered the most recent trials in our database, so we preferred to consider them still in time). This is the least populated at all, just 2.0% of the total.

- ODT out of time: this is the case of trials that present no values in their information and have a difference between the Primary Completion Date and the Start Date greater to the maximum duration theoretically assumed. In this category were also considered trials with a future completion date but that have a Start Date previous to 2015, so have more than 5 years of activity in which didn't publish any results. This could result from the fact that, when the trial was registered, a fictitious future Completion Date was inserted, but the trial never reached its completion. This, instead, is the most densely populated at all, with 60.8% of unique trials.

For this encode, the final situation is showed in Table 9:

Reference	Type_Cod_Trial_State
0	Abandoned
1	Coded with Success
2	Coded with Unsuccess
3	Incomplete
4	ODT in time
5	ODT out of time

Table 9 - Trial State Encoding

After the completion of all these encodes, the final database is ready to be statistically analyzed with regressions performed on Stata software, but this will be further and deeply explained in the following chapter.

Chapter 4 – Data Analysis and Interpretation

4.1 Introduction and Research Strategy

Starting from our research objectives, identified through four hypotheses, we have to define a strategy to test them. In Chapter 3 - Database Creation we demonstrated how to encode the database in order to obtain the variables needed to test hypotheses.

It is necessary to highlight that each of our hypothesis is focused on the concept of risk of failure of a trial, for which the probability of success represents a proxy. As a consequence, the first thing to clarify in this section is the choice of the dependent variable, that has a direct impact on the measure of the probability of success of trials.

In particular, for the aim of our work, we considered two kind of dependent variables indicating the statistical success of a trial:

1. The former is expressed by the encoding *Success Encode*, a Boolean variable where 1 corresponds to the success of that trial and 0 to failure.
2. The latter is represented by the encoding *Type_Cod_Trial_State*, which, in addition to the distinction between successful trials and not successful ones, explores different possibilities of what could be considered as abandoned or dismissed trial.

Before adopting different models to prove our hypotheses, we tested if our data sample was distributed as one of the most common probability distributions. Intuitively, since a trial can only have as outcome (in the case of *Success Encode*) the success or the failure of the testing, we tried to test if the sample could be

distributed as a Binomial distribution. This choice is due to the fact that the presence of a Binomial distribution allows us to determine the average probability of success of a clinical trial belonging to our dataset, which aims to proxy the entire database of clinical trials.

As a consequence, by dividing the dataset in groups (sponsor, intervention type, phase), it became possible compare the probability of success of a given subset with the probability of success of the entire sample, underlining qualitatively which are the most/least performing groups of clinical trials, and, most importantly, to have a reliable measure of the risk of failure associated to a particular type of clinical trial.

To test the adherence of the sample to a Binomial distribution we followed the Chi-Square goodness of fit test. It is worth to consider this test only under the reference of the *Success Encode* encoding. The hypotheses tested for Chi-Square are:

- H0: The data are consistent with a specified distribution.
- H1: The data are *not* consistent with a specified distribution.

By following the procedure¹⁶ of the goodness of fit of the Chi-Square test, we can accept the null hypothesis, according to which our sample is distributed as a binomial. As a consequence, all the formula and properties of binomial distribution are valid in our database.

The identification of a probability distribution of our sample allows us to investigate, with appropriate models, our hypotheses:

¹⁶ For the whole procedure we followed the book "*Metodologie sperimentali in fisica, Cannelli, 2010*".

- H1: Corporations are more willing to invest in clinical trials with lower risk of failure than other investors.
- H2: Phase 3 trials are most likely to succeed than other trials of other phases.
- H3: Corporations are more likely than other investors to sponsor Phase 3 trials.
- H4: A trial with mixed methodologies of intervention is more likely to succeed than trials with a single intervention type or treatment.

To prove so, we had to focus on what are the factors strictly correlated with the risk of failure of a trial. In this sense, the encoding of the entire sample into categories, that numerically shows the presence or not of certain parameters, was fundamental in the adoption of models. As also mentioned in Chapter 3 - Database Creation, the main categories of reference are the Phase of the trial, the Lead Sponsor Class, the Intervention Type and the 5 years period containing the Start Date of the trial, as summarized in Table 10.

Independent Variables	Brief Description
Sponsor Type	The trial could be sponsored by a corporation, university, institution, or a group.
Intervention Type	The trial could test a drug, or a biological treatment, or mixed types.
Phase	It identifies which is the phase of that trial
Start Date	It identifies the starting year of the trial. We created 4 different variables, each one for a timeframe of 5 years, from 1995 to 2015.

Table 10 - Independent Variables

In order to consider simultaneously more aspects of our problem setting, we decided to perform regression analyses through different methods, such as Probit, Heckman Selection model and Survival analysis. The first one is the simplest logistic analysis that could be performed on the dataset, showing correlations between independent and dependent variables. The results obtained in this model can be also commented and explained with the results of the other two models.

In particular, we decided to use Heckit model, because it takes into consideration the issue of selection bias in the dataset, a topic analyzed and mentioned in paragraph 4.2.2 Heckman Selection Model. Moreover, we decided to perform a survival analysis to understand what variables influences the most the “lifecycle” of our clinical trials across time. It is worth to consider that, in the implementation of survival analysis and Heckit model, we decided to slightly change the dependent variable, from Boolean (*Success Encode*) to a discrete one (*Type_Cod_Trial_State*), in order to be able to capture the different states of clinical trials and to have a deeper understanding of the dataset.

4.2 Methodologies

4.2.1 Probit Regression

Linking to paragraph 4.1 Introduction and Research Strategy, we chose to prove the correlation of the dependent variable, *Success Encode* with independent variables through a probit regression.

The choice of performing a probit regression comes from the observation that this model is a way to perform regression for binary outcome variables.

Binary outcome variables are independent variables with two possibilities, like yes/no, positive test result/negative test result or single/not single¹⁷.

Graphically, a probit model estimates a curve that is an S-shaped cumulative normal distribution, that below we report as an example in Figure 17. For instance, in the Y axis we have the dependent variable (*Success Encode*), that assumes 0 or 1 value, and with the *probit* we transform Y from {0,1} to the real line in red. This line represents the cumulative normal distribution Φ , that is $\Phi(Z) \in [0,1]$ with z as the z-score of the normal distribution¹⁸.

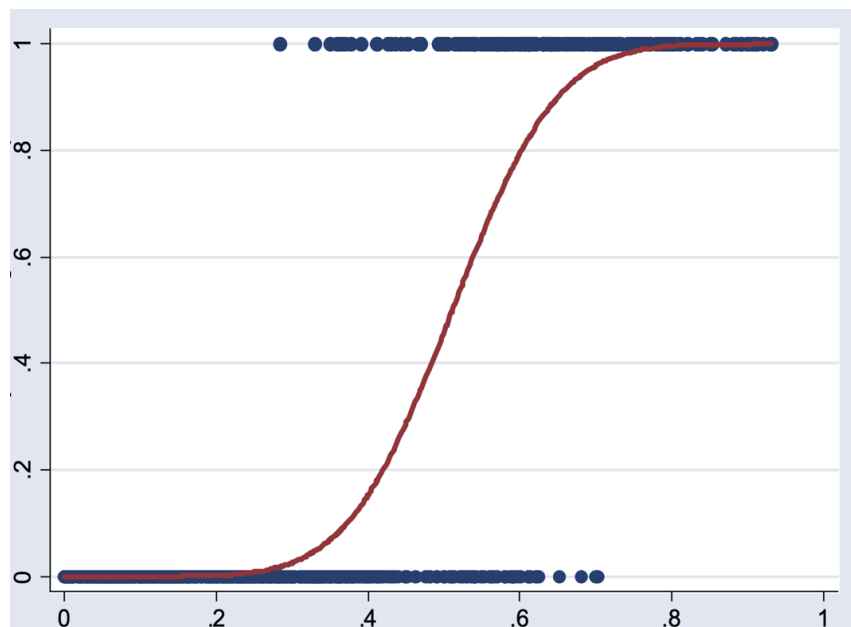


Figure 17 - Probit Graphical Representation

Y represents the value of the dependent variable, that in our case is the success/failure of the trial, and the betas are the regression coefficients for the explanatory variables (also called independent variables).

¹⁷ <https://www.statisticshowto.com/probit-model/>

¹⁸ *Sustainable Development U9611 Econometrics II, O'Halloran*

4.2.1.1 Probit Execution

On the practical sight, to perform the probit regression we imported our encoded database on Stata, a licensed software dedicated to statistical analysis. This process starts with the import, from our Excel database, on Stata, taking the right spreadsheet of reference. After this initial operation, the software presents seven different variables: Rank, NCTid, Success Encode, Code_Stata_Sponsor, Code_Stata_IntType, Code_Stata_Phase and Code_Stata_5Y. Code_Stata variables are distributed like showed in Table 11:

	Obs.	Mean	Std.Dev.
Code_Stata_Sponsor	3,374	2.508299	1.240591
Code_Stata_IntType	3,374	2.603142	0.8112036
Code_Stata_Phase	3,374	2.124778	1.174447
Code_Stata_5Y	3,374	3.166568	0.9049701

Table 11 - Variables Population

After doing this, we transform the last four variables, that will become independent variables in our regression, in dummies, using the Stata command *tab*, as shown in the following example: imagine transforming into dummies the variable Code_Stata_Sponsor, composed by 4 categories. The output of this operation results as showed in Table 12:

Code_Stata_Sponsor	Freq.	Percent	Cum.
Group	970	28.75	28.75
Corporation	923	27.36	56.11
Institute	277	8.21	64.32
Other_Sponsor	1,204	35.68	100.00
Total	3,374	100.00	

Table 12 - Example of Code_Stata_Sponsor

This procedure was performed also for Phase, Intervention Type and 5Y, transforming each of the four Code_Stata into 4 different dummies, obtaining a final set of variables as showed in Table 13:

Code_Stata_Sponsor	Code_Stata_IntType	Code_Stata_Phase	Code_Stata_5Y
Group	Biological	Phase_2	From_1995_ to_2000
Corporation	Drug	Phase_3	From_2001_ to_2005
Institute	Mixed_Methodologies	Phase_1_and_2	From_2006_ to_2010
Other_Sponsor	Other_Intervention	Other_Phase	From_2011_ to_2015

Table 13 - Dummies of each Variable

At this point, everything is set to perform the probit on the software from the point of view of input data, with all the dummies that will perform regression in respect to the main variable which is the binary one of *Success Encode*.

The last adjustments left were devoted to the presentation of the output, like inserting after the probit expression the command *rob*, that performs the estimation considering robust standard errors¹⁹ and marginals coefficients. Another possibility is that, instead of visualizing coefficients values, to use the command *dydx(*) atmeans post* which estimates all the marginal effects of variables and margins at the means of covariates. In particular, this second regression gave to us, for each category, marginal effects from the baseline.

¹⁹ When the homogeneity of variance assumption is violated the ordinary least squares (OLS) method calculates unbiased, consistent estimates of the population regression coefficients. In this case, these estimates won't be the best linear estimates since the variances of these estimates won't necessarily be the smallest.

According to all of these considerations, the output obtained is represented in Table 14:

Log likelihood = -315.38891
Number of obs. = 1,701
Prob > chi2 = 0.0000
Pseudo R2 = 0.1769

Success Encode	Coefficients ²⁰	Stat. Significance ²¹
Group	- 0.1511451 (0.2705358)	
Corporation	0.9099713 (0.2448507)	***
Other_Sponsor	0 (omitted)	
Biological	- 0.1415685 (0.3760926)	
Mixed_Methodologies	0.2515005 (0.128755)	*
Other_Intervention	0 (omitted)	
Phase_2	- 0.1699644 (0.2432404)	
Phase_3	0.6323805 (0.2263271)	***
Other_Phase	0 (omitted)	
from_1995_to_2000	- 0.1694209 (0.2851366)	
from_2001_to_2005	- 0.1370107 (0.1750415)	
from_2011_to_2015	0.1009123 (0.1214607)	
_cons	- 2.552914 (0.3104816)	

Table 14 - Coefficients Output of Probit

Instead, the one adopting the command *dydx(*)* is showed in Table 15:

²⁰ Between brackets are reported standard error values.

²¹ If there are present “***”, the value $P > |z|$ is lower than 0.01, with “**” this value is lower than 0.05, while with “*” this value is lower than 0.1, while with no indications there is no statistical significance.

Success Encode	dy/dx Coefficients ²²	Stat. Significance ²³
Group	- 0.0098977 (0.0178606)	
Corporation	0.0595892 (0.0155097)	***
Other_Sponsor	0 (omitted)	
Biological	- 0.0092706 (0.0226104)	
Mixed_Methodologies	0.0164694 (0.008309)	**
Other_Intervention	0 (omitted)	
Phase_2	- 0.0111301 (0.0172157)	
Phase_3	0.0414112 (0.0142933)	***
Other_Phase	0 (omitted)	
from_1995_to_2000	- 0.0110945 (0.0190374)	
from_2001_to_2005	- 0.0089721 (0.0111514)	
from_2011_to_2015	0.0066082 (0.0078931)	

Table 15 - dydx Coefficients Output

The first output is the one showing β coefficients, here described²⁴:

- β Coefficients: These are the regression coefficients. The predicted probability of admission can be calculated using these coefficients. For a given record, the predicted probability of admission is:

$$\begin{aligned}
 F(-2.55914 - 0.1511451 \textit{Group} + 0.9099713 \textit{Corporation} \\
 - 0.1415685 \textit{Biological} + 0.2515005 \textit{Mixed_Methodologies} \\
 - 0.1699644 \textit{Phase_2} + 0.6323805 \textit{Phase_3} \\
 - 0.1694209 \textit{from_1995_to_2000} \\
 - 0.1370107 \textit{from_2001_to_2005} \\
 + 0.1009123 \textit{from_2011_to_2015})
 \end{aligned}$$

where F is the cumulative distribution function of the standard normal. A positive coefficient means that an increase in the predictor leads to an

²² Between brackets are reported Delta-method standard error values.

²³ If there are present “***”, the value $P > |z|$ is lower than 0.01, with “**” this value is lower than 0.05, while with “*” this value is lower than 0.1, while with no indications there is no statistical significance.

²⁴ <https://stats.idre.ucla.edu/stata/output/logistic-regression-analysis/>

increase in the predicted probability. A negative coefficient means that an increase in the predictor leads to a decrease in the predicted probability.

The situation is different for the second performed regression, the one with marginal coefficients, with the only difference in the meaning of the coefficients obtained:

- They indicate the effect of a unit change of that variable on the probability $P(Y=1|X=x)$, given that all other variables are constant. For example, Corporation in this particular regression has a meaningful marginal variation of 5.96%.

4.2.2 Heckman Selection Model

4.2.2.1 New discrete dependent variable

Until now we treated the dependent variable as a Boolean one, considering only the success or failure of a trial. The reality is slightly different, since a clinical trial cannot simply be considered as a success or not, but we have to take into account some other possibilities:

- Abandoned trial: the trial won't be concluded, and results are not available.
- Coded as Success: same encoding as *Success Encode*.

- Coded with Unsuccess: the trial has been concluded and data about the outcome are available, but the trial is considered as a failure.
- Incomplete: the trial presented only some results, so it's not possible to completely and effectively evaluate his outcome.
- ODT (Omission-Data Trial) in time: it means that the trial is not concluded and there are not available data.
- ODT out of time: the trial has reached the completion date, but there are not available data about the outcome.

In Table 16 we have the list of encoding just described. As a consequence, it's possible to consider not a dichotomous variable anymore as a dependent variable, but a discrete one, with the 6 classes described before.

<i>Reference</i>	Type_Cod_Trial_State
0	Abandoned
1	Coded with Success
2	Coded with Unsuccess
3	Incomplete
4	ODT in time
5	ODT out of time

Table 16 - Encoding of dependent variables

4.2.2.2 Selection bias

Starting from this encoding, we can realize that the trials encoded as 0,3,5 can be considered as a failure, since they can't be considered concluded. In particular, they are abandoned, or it's not possible to effectively evaluate them, or they did not present results. Vice versa, trials with the remaining encodings can be

considered differently, because they are completed trials. In particular, they could be successful, they could be still in time to be evaluated as success or not, or they could present a failure outcome.

From a distinction like that, it is possible to understand that regarding the sample selection, it is not properly a random sample and there could be a selection bias: “Selection bias is a kind of error that occurs when the researcher decides who is going to be studied. It is usually associated with research where the selection of participants isn’t random.”²⁵

Indeed, it is clear that only trials having 1,2,4 as encodings could give significant results to interpret, while 0,3,5 are quite insignificant, and trials belonging to this class will never be able to be empirically evaluated, as happens for the other group.

In order to overcome this limit and obtain unbiased estimates, it is possible to perform a regression analysis using the Heckman Selection Model (1976). It is sometimes called the Heckit model and it’s a method for estimating regression models which suffer from sample selection bias. Under the Heckman selection framework, the dependent variable is only observable for a portion of data.

4.2.2.3 Heckit model execution

We performed the analysis on STATA and by using the command *heckman*, we have the following outcome, where *Success Encode* is our binary variable y_i and *Code_035* represents our selection equation z_i .

By looking at the Prob>chi2, that is 0,0689, we can conclude that it is possible to reject the null hypothesis at 95% of confidence interval. In particular, the H0 is that all coefficients estimated are equal to 0, so they don’t generate any effect on the regression.

²⁵ <https://www.iwh.on.ca/what-researchers-mean-by/selection-bias>

Heckman assumes that Success Encode is the dependent variable and that the first variable list (Corporation and Phase3) are the determinants of Success Encode. The variables specified in the select() option (Group, Corporation, Other_Sponsor, Mixed_Methodologies, Phase3, from_1995_to_2000, from_2001_to_2005) are assumed to determine whether the dependent variable is observed (the selection equation). Thus, we fit the model:

$$Success\ Encode = \beta_0 + \beta_1 * Corporation + \beta_2 * Phase_3 + u_1$$

And we assumed that Success Encode is observed if:

$$\begin{aligned} \gamma_0 + \gamma_1 * Group + \gamma_2 * Corporation + \gamma_3 * Other_Sponsor + \gamma_4 \\ * Mixed_Methodologies + \gamma_5 * Phase_3 + \gamma_6 \\ * from_1995_to_2000 + \gamma_7 * from_2001_to_2005 + u_2 \end{aligned}$$

where u_1 and u_2 have correlation ρ , where ρ is constrained between [-1,+1] Similarly to what we obtained from *probit* regression with *Success Encode*, the most significant variables that explains the dependent variable are *Corporation* and *Phase3*, suggesting us that trials having a company as lead sponsor and trial belonging to phase 3 are more likely to succeed than others, exactly as we obtained in paragraph 4.2.2.2 Selection bias.

For which concerns the *Mixed_Methodologies* variable, we decided to keep it only on the selection equation, because by adding it into the treatment equation, the Prob>Chi2 was definitely worse, and *Mixed_Methodologies* was not a significant variable for the model. The output of this model is showed in Table 17:

Number of obs. = 3,374
Censored obs. = 2,946
Uncensored obs. = 428
Prob > chi2 = 0.0689

Success Encode	Coefficients	Stat. Significance ²⁶
Corporation	0.1617802 (0.0790962)	**
Phase_3	0.1340493 (0.0640487)	**
_cons	0.2002853 (0.2277081)	
Code_035		
Group	- 0.4086896 (0.1139773)	***
Corporation	0.428751 (0.1099816)	***
Other_Sponsor	- 0.6694965 (0.1181084)	***
Mixed_Methodologies	0.2584239 (0.0697482)	***
Phase_3	0.5905642 (0.0634372)	***
from_1995_to_2000	- 0.4133883 (0.1483495)	***
from_2001_to_2005	- 0.2353321 (0.0901344)	***
_cons	- 1.283361 (0.105103)	
mills		
lambda	- 0.1043004 (0.1118508)	
rho	- 0.25740	
sigma	0.40520008	

Table 17 - Heckit Output

4.2.3 Survival Analysis

For what concern the last analysis performed, we changed the perspective completely. Until now, we have studied trials within given data in time, while here we want to study how trials evolved on a certain timeline. To properly complete this task, we decided to perform a survival analysis. The reason behind the inclusion of this model is that, in Probit, we investigated how risk factors were correlated to the success or failure of a trial. Instead, could be interesting to understand how a risk factor affects time to failure or other events.

²⁶ If there are present “***”, the value $P > |z|$ is lower than 0.01, with “**” this value is lower than 0.05, while with “*” this value is lower than 0.1, while with no indications there is no statistical significance.

We may have study dropout, and therefore, studies that should be further investigated to understand if they failed or not. In these cases, Probit regression is not appropriate²⁷.

Survival analysis²⁸ is the analysis of time-to-event data. Such data describe the length of time from a time origin to an endpoint of interest. For example, individuals might be followed from birth to the onset of some disease, or the survival time after the diagnosis of some disease might be studied. Survival analysis methods are usually used to analyze data collected prospectively in time.

One of the reasons why survival analysis requires “special” techniques is the possibility of not observing the event of interest for some individuals. For example, individuals may drop out of a study, or they might have a different event, which is not part of the endpoint of interest. Another possibility is that there might be a time point at which the study finishes and thus if any individuals have not had their event yet, their event time will not have been observed. These incomplete observations cannot be ignored but need to be handled differently. This is called *censoring*. The objectives of survival analysis include the analysis of patterns of event times, the comparison of distributions of survival times in different groups of individuals and examining whether and by how much some factors affect the risk of an event of interest.

The most commonly encountered type of censoring and easiest to handle in the analysis is right censoring. Right censoring occurs when an individual is followed up from a time origin t_0 up to some later time point t_c and he/she has not had the event of interest, such that all we know is that their event has not occurred up to their censoring time t_c . This may occur, for example, if an individual drops out of a study before the event of interest occurs.

²⁷ <http://www.stat.columbia.edu/~madigan/W2025/notes/survival.pdf>

²⁸ <https://www.sciencedirect.com/science/article/pii/S1756231716300639>

Commonly studies are terminated at some specified time and at the end of the study some individuals have not yet had their event.

4.2.3.1 Dataset adjustments

To perform survival analysis, we had to transform our database inserting a dedicated spreadsheet with this structure:

- Time: to build the timeline of analysis, we created a column with the difference, in years, between the Completion Date of the trial and the Start Date of it. The only data cleaning needed in this phase was to remove those that didn't show a completion date, so had a NC value in the column, and so could not present a reference in the timeline.
- EVENT: we built a specific encode forming three categories:
 - Censored, so trials that didn't achieve a completion in terms of results (in this category we inserted abandoned trials, ODT out of time and ODT in time);
 - Failed, so trials that presents results but that are not successful (in this category we inserted incomplete trials and Coded but with statistically unsuccess);
 - Positive, so trials that are coded and achieved statistical success, so trials coded with 1 in Success Encode.

4.2.3.2 Survival Analysis Execution

To perform a survival analysis, the first step is to declare data to be treated as survival analysis data. After this first operation, the software recognizes as failure event the values 1 and 2 in the encode EVENT. These values are those recognized as “death” occurrences of the trial (its failure or abandon). These values, in addition, are all spotted along the timeline, which goes from zero to 30 years of difference between the completion date of the trial and starting date. To perform the survival analysis, we used one of its most famous models, the Cox Proportional Hazard model. We decided to use this method because Cox proportional hazards regression²⁹ is used to relate several risk factors, considered simultaneously, to survival time. In a Cox proportional hazards regression model, the measure of effect is the hazard rate, which is the risk of failure (i.e. the risk or probability of suffering the event of interest), given that the trials “has survived” up to a specific time.

If the hazard ratio for a predictor is close to 1 then that predictor does not affect survival. If the hazard ratio is less than 1, then the predictor is protective (i.e., associated with improved survival) and if the hazard ratio is greater than 1, then the predictor is associated with increased risk (or decreased survival).

4.2.3.3 Cox proportional hazard execution

To perform the Cox Proportional Hazard Model, we had to set all the independent variables (which are always the 12 different ones also used in other models) and decide whether to show Coefficients (Table 18) or Hazard Ratios (Table 19):

²⁹ https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Survival/BS704_Survival6.html

	Coefficients	Stat. Significance ³⁰
Group	- 0.0653463 (0.0910817)	
Corporation	0.1728282 (0.0913502)	*
Institute	0.1024803 (0.1335162)	
Other_Sponsor	0 (omitted)	
Biological	0.2441017 (0.2269113)	
Drug	0.1570782 (0.1151311)	
Mixed_Methodologies	0.0304252 (0.1225669)	
Other_Intervention	0 (omitted)	
Phase_2	- 0.19336 (0.1044239)	*
Phase_3	- 0.2366771 (0.1093631)	**
Phase_1_and_2	- 0.1037081 (0.1571531)	
Other_Phase	0 (omitted)	
from_1995_to_2000	- 0.4862613 (0.2009914)	**
from_2001_to_2005	- 0.0613617 (0.1104582)	
from_2006_to_2010	0.0070729 (0.0773744)	
from_2011_to_2015	0 (omitted)	

Table 18 - Coefficients of Cox Regression

Log likelihood = -5974.8614

No. of subjects = 3,056

No. of failures = 837

Prob > chi2 = 0.0112

³⁰ If there are present “***”, the value $P > |z|$ is lower than 0.01, with “**” this value is lower than 0.05, while with “*” this value is lower than 0.1, while with no indications there is no statistical significance.

	Haz. Ratio	Stat. Significance ³¹
Group	0.936743 (0.0853201)	
Corporation	1.188662 (0.1085845)	*
Institute	1.107915 (0.1479246)	
Other_Sponsor	1 (omitted)	
Biological	1.276474 (0.2896464)	
Drug	1.170087 (0.1347134)	
Mixed_Methodologies	1.030893 (0.1263533)	
Other_Intervention	1 (omitted)	
Phase_2	0.8241852 (0.0860647)	*
Phase_3	0.7892461 (0.0863144)	**
Phase_1_and_2	0.9014884 (0.1416717)	
Other_Phase	1 (omitted)	
from_1995_to_2000	0.6149211 (0.1235938)	**
from_2001_to_2005	0.940483 (0.103884)	
from_2006_to_2010	1.007098 (0.0779236)	
from_2011_to_2015	1 (omitted)	

Table 19 - Hazard Ratios of Cox Regression

Coefficients³² in our Cox regression relate to hazard; a positive coefficient indicates a worse prognosis and a negative coefficient indicates a protective effect of the variable with which it is associated. The hazard ratio associated with a predictor variable is given by the exponent of its coefficient; this is given with a confidence interval under the "95% Conf. Interval".

The most significant coefficients to be analyzed are those that present differences one with another. For example, Corporation and Institute present both positive coefficients, and we want to see if these categories present

³¹ If there are present "****", the value $P > |z|$ is lower than 0.01, with "***" this value is lower than 0.05, while with "*" this value is lower than 0.1, while with no indications there is no statistical significance.

³² https://www.statsdirect.com/help/survival_analysis/cox_regression.htm

differences. This can be achieved using post estimation commands as showed in Table 20:

Test Corporation-Institute = 0	
chi2	0.26
Prob > chi2	0.6087

Table 20 - Post Estimation Command

As we can see, prob>chi2 gives back a value greater than our smallest reference (5%), so we can affirm that these two categories differ one with another. This test can be checked for all the variables of interest.

4.3 Output Analysis

4.3.1 Probit Output Analysis

Referring to the hypotheses we want to test mentioned in paragraph 2.1.7 Research objectives, it is possible to express these considerations:

H1: Corporations are more willing to invest in clinical trials with lower risk of failure than other investors.

Considering this hypothesis, we can look at the output of the Probit regression. The β coefficient of the Sponsor variable Corporation is statistically meaningful ($P > |z| < 0.05$) and it's largely positive (0.9099713). Also looking at marginal coefficients, we can underline that, with an increase of one unit of the variable Corporation, the success is increased by 0.0595892, in reference to the omitted

variable. This means that the variable Corporation is positive correlated to the probability of success of the trials, so the presence of this type of Sponsor could lead to positive results of the outcome of the trial. This led us to consider as verified the hypothesis number one, under the condition of Success Encoding (the binary dependent variable). Indeed, in the following part we are going to test the same hypothesis under other perspectives.

A possible interpretation of this result could be given by considering the nature of Corporations. In fact, with the reference of the premises made in paragraph 2.1.3 Short-termism of corporations, we have looked at how Corporations are likely to maximize investments profits, under a perspective of short-termism. In other words, the fact that the probability of success is higher for trials sponsored by Corporations, suggests that this type of Sponsor is selecting and investing efficiently, minimizing the risk of failure of the trial.

H2: Phase 3 trials are most likely to succeed than other trials of other phases.

Under the same conditions previously mentioned of Success Encoding, we can look at how β coefficient of Phase 3 has a largely positive value (0.6323805) with statistical meaning. The same scenario is present for marginal coefficients, with a value of 0.0414112, in reference to the omitted variable. These values, just like what happened for Corporation before, let us assume a positive correlation between the presence of Phase 3 and the success of the trial.

We can interpret this result considering that Phase 3 trials are most likely to be successful probably due to the fact that they have already passed to stages (Phase 1 and Phase 2), and that is the Phase closest to the commercialization.

H3: Corporations are more likely than other investors to sponsor Phase 3 trials.

To demonstrate this hypothesis, we performed a descriptive statistics analysis on the database in this way: we filtered, from the total set of trials (3374 unique ones), only those belonging to Phase 3, which are 1012 unique ones. Of these, we counted how many trials were sponsored by Corporations, as showed in Table 21:

Reference	Type_Cod_Sponsor	Count	% on the Total
1	Group and Network	307	30,34%
2	Inc., Corp. and Industry	420	41,50%
3	University and College	154	15,22%
4	Hospital	29	2,87%
5	Institute	70	6,92%
6	Center	32	3,16%

Table 21 - Proportion of Sponsors on trials

We can see from the results that, among the different typologies of Sponsor present, Corporations (so the encode 2 in Type_Cod_Sponsor) are the most present, as suggested by our hypothesis H2.

This suggests us that companies have a deep interest on the success of trials, in fact we could prove the presence of a link between H1 and H2, because while Corporations are willing to risk less, and Phase 3 trials are likely to be the most successful, we found that Phase 3 trials are mainly sponsored by companies.

H4: A trial with Mixed Methodologies of intervention is more likely to succeed than trials with a single intervention type or treatment.

Both outcomes of the regression of Mixed_Methodologies, the one showing β coefficient (0.2515005) and the one showing marginal coefficient (0.0164694, with reference to the omitted variable), underline the positive relation between

the usage of Mixed Methodologies as Intervention Type and the success of the trial, both with statistical meanings.

This lead us to think that, even if the usage of Mixed Methodologies involves higher computational and operational efforts to realize the trial (sometimes, when two different methodologies are enrolled, efforts are duplicated if there are no synergies between treatments), like we described in the paragraph 2.1.7 Research objectives, the final results seems to be positive for the success of the trial. A possible interpretation of this scenario could be that, involving different methodologies in the same procedure, weaknesses of each treatment could be flattened by the strength of the others, achieving a final positive effect on the trial success, highlighting that Mixed Methodologies are more efficient as described in the Literature.

4.3.2 Heckit Model Output Analysis

Results obtained from Heckit Model are in line, as mentioned in paragraph 4.2.2.3 Heckit model execution, with the results obtained in the Probit regression. In particular, is confirmed the correlation between success of the trial and the variables Corporation and Phase 3. Instead, we found that Mixed Methodologies variable is not correlated to the dependent variable.

4.3.3 Cox Proportional Hazard Output Analysis

These are the considerations on hypotheses that comes from the analysis of the output of the Cox Regression:

H1: Corporations are more willing to invest in clinical trials with lower risk of failure than other investors.

Even if Corporation presents a $P > |z|$ slightly higher than the limit (0.059), we have not the statistical reference to correctly interpret its coefficient with a confidence of 95%.

A possible explanation of this phenomenon is discussed in the paper “Use and Misuse of Statistical Significance in Survival Analysis, Furuya, Wijesundara, Neeman, Metzger, 2014”, that states that the biological outcome from the experiment should be considered first, and then statistics applied to determine if the results are likely to be due to chance. In this process, it should be remembered that a cutoff P value of 0.05 is relative; a P value of 0.1 indicates that a particular result would occur by chance 10% of the time. This could still reflect a biologically important effect.

If, following their suggestions, we adopt a P value of 0.1, we find that Corporation has a positive coefficient and a Hazard Ratio greater than 1, both statistically meaningful. This could be interpreted in this way: the presence of Corporation reduces the “survival” of the trial, following what we mentioned in our premises, so that companies prefer to follow short-time investments and possibly closer to commercialization. This result is exactly in line with the findings of Probit about H1.

H2: Phase 3 trials are most likely to succeed than other trials of other phases.

With the presence of a Coefficient of -0.2366771 and a HR of 0.7892461, both with statistical influence, we can state that the presence of the variable Phase 3 indicates a protective effect of the variable to which it is associated, so the success of the trial. This result is perfectly coherent with what we found from the Probit regression. As result of this, we can state that hypothesis number three is always verified, both under the Success Encode perspective, both under the different Encodes of the State in which the trial is.

H4: A trial with Mixed Methodologies of intervention is more likely to succeed than trials with a single intervention type or treatment.

Regression values for Mixed Methodologies shows the absence of the statistical reference to correctly interpret the outcome. Referring to the same paper mentioned in H1, so just looking at values and not at the statistical significance, we can observe that the presence of Mixed Methodologies as Intervention Type is not significative, because presents both a coefficient close to zero, and a Hazard Ratio close to 1. A possible interpretation of this result is that, using different methodologies for a trial, positive effects that could possibly arise from a treatment could be offset by the weaknesses of another, obtaining a final result that flattened results.

4.3.2.1 Kaplan-Meyer

The output can be also analyzed graphically for each field using Kaplan-Meyer survival functions, that graphically underline, for each category, how trials move in time. For example, for the Sponsor Type, where its four categories are Group, Corporation, Institute and Other Sponsors, as reported in Figure 18:

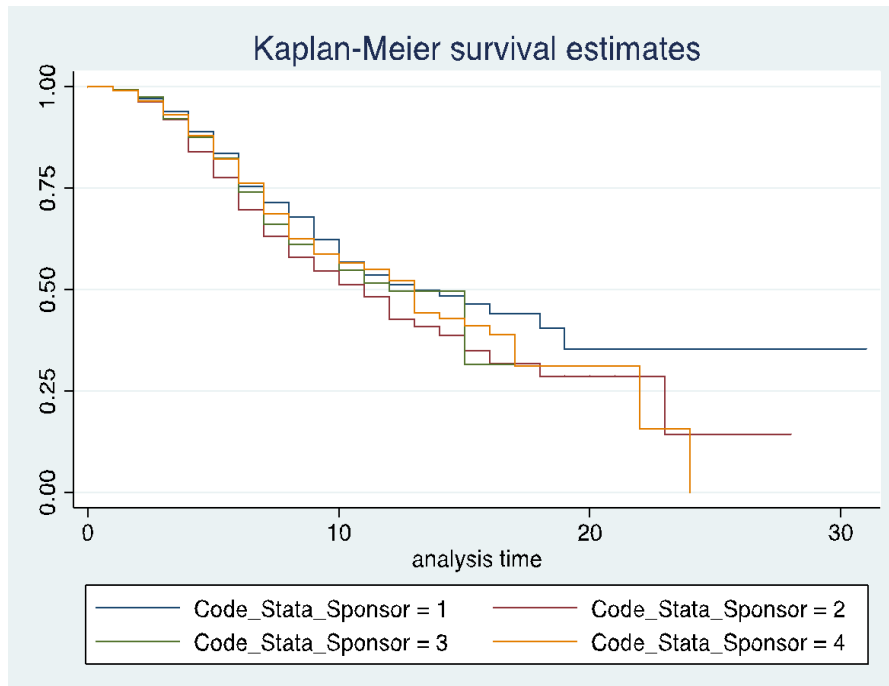


Figure 18 - Kaplan-Meier survival estimate

Kaplan–Meier curves³³ can be used in simple analyses of which the aim is to compare survival times of two or more generally a small number of groups. For example, in a clinical trial the researchers might want to look at the survival times of individuals allocated to treatment A and of those allocated to treatment B. This can be examined by plotting two Kaplan–Meier curves, one for treatment A and one for treatment B in the first example. In our case, one might want to compare the survival times in order to determine whether having a certain type of sponsor is linked to shorter analysis time. The upper figure shows the two curves. Note that each event time appears as a ‘jump’ on a Kaplan–Meier curve. Censoring times are also commonly plotted on a Kaplan–Meier curve, to visualize the amount and patterns of censoring with time.

³³ <https://www.sciencedirect.com/science/article/pii/S1756231716300639>

Chapter 5 - Conclusion

Investment opportunities in scientific research are always under the attention of the academic and the corporate world. In particular, our attention was focused on clinical trials. With this dissertation, we analyzed how factors could affect with larger extent the success or not of trials, and how investors behave when choosing an investment alternative belonging to this field. We studied the literature treating this argument, looking for information about the nature of trials, how they are performed, in what environment they are inserted, which are all the possible types existing and the most critical issues related to the market. Thanks to information gathered from papers, we were able to formulate our hypotheses, so what we wanted to prove statistically in order to express considerations about the relation between investors and investment.

To demonstrate these hypotheses, different steps were needed. First of all, there was a deep study of trials data gathered from the data source, both in order to understand which of these could be meaningful to be analyzed, both to understand which fields had the most stable and trustable information (freedom from bias). To perform the data extraction, we used [Clinicaltrials.gov](https://clinicaltrials.gov) as data source, which was one of the most reliable and complete at all. Data extracted from the website were raw data, so need to be cleaned, elaborated and coded to create a final database.

5.1 Hypotesis analysis

Basing on the database, we applied different statistical methodologies to test our hypotheses. First of all, we considered two different approaches in considering the dependent variables: the first, Boolean, considers the success or not of the trial (Success Encode); the other one considers the possible states in which the trial could be (Type_cod_Trial_State).

To understand which set of variables were correlated to the trial success (Boolean), we performed a Probit regression. Moreover, to confirm and further analyze the outcome of the regression, we performed two more analysis (based on the trial state): Heckman selection model and Survival Analysis. Results from these two additional methods were useful because they confirmed the interpretation of the regression, giving more value to what we wanted to prove. In particular, what we found is in line with what mentioned in the literature review, which is:

- H1: we found as true that Corporations are more willing to invest in least risky clinical trials than other investors. This confirms what is expressed in literature by the concept of short-termism of companies in investment decisions, which want to maximize profit in the shortest time horizon as possible. Following these considerations, we can say that the presence of this type of Sponsor is correlated to the probability of success of a trial.
- H2: We proved that Phase 3 trials are most likely to succeed than other phases, from the statistical point of view. The rationale behind derives from the definition of Phase 3 trials, which is the last and the most expensive stage of trial possible, has already passed two previous stages (Phase 1 and Phase 2) and it's the closest to commercialization.

Results obtained in these two hypotheses were also confirmed by the analysis of H3, which shows a relation between the choice of Corporations to sponsor Phase 3 trials.

The last hypothesis we wanted to prove was H4, so that trials adopting Mixed Methodologies are most likely to success than trials with a single intervention type. This hypothesis was partially proved, because the result of the probit suggest us to consider as true this statement, but this was not verified in case of Heckman and Cox regression. We suggest to further investigate this point (maybe adopting other models than the ones here used) to fully understand the rationale of the influence of this Intervention of the trial success.

5.2 Limitations

The main finding of this work focuses on the risk taking of corporations sponsoring investments in scientific research. Our findings demonstrated that companies are not only investing in short-term researches, but also in least risky ones. Anyway, our work presents some limitations:

- 1) First of all, we limited our dataset to existing data presents on Clinicaltrials.gov. Our view on the dataset was restricted on trials belonging to the Oncology sector, and in particular that analyzed as primary outcome the Overall Survival and Progression Free Survival of patients. Of these, we limited the timeframe to 20 years (from 1995 to 2015), in order to consider just completed trials.
- 2) All of our considerations were based on data presented on publications on the data provider, potentially affected by bias and by opportunistic behavior of publishers, as already underlined by Salandra (2018). We

cannot know if data reported truly represented the reality of things, because we were limited to the observation of trials' results disclosed.

- 3) All of our assumptions, in interpreting the hypothesis, were demonstrated by models adopted. So, there is the possibility that, adopting other tools, results may vary. We suggest, for those interested in analyzing deeper these aspects, to perform more analysis to obtain a robust version.

5.3 Further discussion

Starting from our findings, it could be interesting to expand our model and database. Possible implementations could regard the selection of trials. In particular, we could suggest enlarging our point of view, which was focused on trials with the two main measures of survival as Primary Outcome. An addition could be presented both on the Outcome to consider as primary measures, both in considering Secondary Outcomes. We have to remember that we limited our analysis on oncology trials, but more fields could be analyzed with the framework we proposed in our dissertation.

As time passes, to update the analysis and prove the validity even in most recent data, trials from 2016 on could be added to verify if hypothesis are still valid and so confirmed as trend in time. Another addition could be that, in our work, we didn't consider the presence of hybrid sponsors, as example trials sponsored by the conjoined work of privates and government, which could be relevant to be analyzed.

Even the addition of variables could lead the model to evolve and to capture more information. As an example, we didn't mention the aspect of the

population composing the trial sample. Coding and inserting this variable as independent one in the analysis could influence the model and maybe could open new scenarios and hypothesis to be tested inside this framework.

Thanks to this dissertation, we aimed at amplifying the existing knowledge on a relevant topic such as investments in scientific research. This field, nowadays, is under the public attention (due to the pandemic situation of Covid19), and, more than ever, making an informed decision in this context could make a difference.

Bibliography

Adams, J. D. (1990). Fundamental Stocks of Knowledge and Productivity Growth. *Journal of Political Economy*, 98(4), 673–702.

Amoroso, S., Moncada-Paternò-Castello, P., & Vezzani, A. (2016). R&D profitability: the role of risk and Knightian uncertainty. *Small Business Economics*, 48(2), 331–343.

Arora, A., Gambardella, A., Magazzini, L., & Pammolli, F. (2009). A Breath of Fresh Air? Firm Type, Scale, Scope, and Selection Effects in Drug Development. *Management Science*, 55(10), 1638-1653.

Arora, A., Belenzon, S., & Pataconi, A. (2017). The decline of science in corporate R&D. *Strategic Management Journal*, 39(1), 3–32.

Arrow, K. J. (1972). Economic Welfare and the Allocation of Resources for Invention. In *Readings in Industrial Economics* (pp. 219–236). Macmillan Education UK.

Baker, T. B., Smith, S. S., Bolt, D. M., Loh, W.-Y., Mermelstein, R., Fiore, M. C., Piper, M. E., & Collins, L. M. (2017). Implementing Clinical Research Using Factorial Designs: A Primer. *Behavior Therapy*, 48(4), 567–580.

Bigby, M., & Gadenne, A.-S. (1996). Understanding and evaluating clinical trials. *Journal of the American Academy of Dermatology*, 34(4), 555–590.

Bound, J., Clint Cummins, Zvi Griliches, Bronwyn H. Hall, Adam B. Jaffe, 1982. *Who Does R&D and Who Patents?* NBER Working Papers 0908, National Bureau of Economic Research, Inc.

Budish, Eric, Benjamin N. Roin, and Heidi Williams. 2015. Do Firms Underinvest in Long-Term Research? Evidence from Cancer Clinical Trials. *American Economic Review*, 105 (7): 2044-85.

Campbell, John Y. 1996. Understanding risk and return. *Journal of Political Economy* 104, no. 2: 298-345.

Chandra, A., Garthwaite, C., & Stern, A. D. (2017). Characterizing the Drug Development Pipeline for Precision Medicines. National Bureau of Economic Research.

Choi, J., & Lee, J. (2017). Repairing the R&D market failure: Public R&D subsidy and the composition of private R&D. *Research Policy*, 46(8), 1465–1478.

Dunk, A. S., & Kilgore, A. (2001). Short-term R&D bias, competition on cost rather than innovation, and time to market. *Scandinavian Journal of Management*, 17(4), 409–420.

Dwan, K., Gamble, C., Williamson, P. R., & Kirkham, J. J. (2013). Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias — An Updated Review. *PLoS ONE*, 8(7), e66844.

Evans, S. R. (2010). Clinical trial structures. *Journal of Experimental Stroke and Translational Medicine*, 3(1), 8–18.

Fink, E., Kokku, P. K., Nikiforou, S., Hall, L. O., Goldgof, D. B., & Krischer, J. P. (2004). Selection of patients for clinical trials: an interactive web-based system. *Artificial Intelligence in Medicine*, 31(3), 241–254.

Fleming, M. (1955). External Economies and the Doctrine of Balanced Growth. *The Economic Journal*, 65(258), 241.

Fleming, L., Greene, H., Li, G., Marx, M., & Yao, D. (2019). Government-funded research increasingly fuels innovation. *Science*, 364(6446), 1139–1141.

Foray, D. (2010). A primer on patent and innovation. *Management international*, 14(3), 19.

Freidlin, B., & Korn, E. L. (2017). Two-by-Two Factorial Cancer Treatment Trials: Is Sufficient Attention Being Paid to Possible Interactions? *JNCI: Journal of the National Cancer Institute*, 109(9).

Furuya, Y., Wijesundara, D. K., Neeman, T., & Metzger, D. W. (2014). Use and Misuse of Statistical Significance in Survival Analyses. *MBio*, 5(2).

Hemmiki E. *Br Med J*. 1980 Mar 22; 280(6217): 833–836. Study of information submitted by drug companies to licensing authorities.

Jackson, R. R. (1985). The evaluation of clinical trials. *Postgraduate Medical Journal*, 61(712), 133–139.

Knight, Frank H., *Risk, Uncertainty and Profit* (1921). University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship.

Krieger, Joshua L. *Trials and Terminations: Learning from Competitors' R&D Failures*. Harvard Business School Working Paper, No. 18-043, November 2017.

López, A. (2009). *Innovation and Appropriability: Empirical Evidence and Research Agenda*.

Ludwig von Mises, *Human Action: A Treatise on Economics*, ed. Bettina Bien Graves (4th revised edition) (Irvington-on-Hudson: Foundation for Economic Education, 1996).

Nelson, R. (1959). The Simple Economics of Basic Scientific Research. *Journal of Political Economy*, 67(3), 297-306.

Rao, A. (2015). Entry and investment decisions in the pharmaceutical industry.

Reid, E. K., Tejani, A. M., Huan, L. N., Egan, G., O'Sullivan, C., Mayhew, A. D., & Kabir, M. (2015). Managing the incidence of selective reporting bias: a survey of Cochrane review groups. *Systematic Reviews*, 4(1).

Salandra, R. (2018). Knowledge dissemination in clinical trials: Exploring influences of institutional support and type of innovation on selective reporting. *Research Policy*, 47(7), 1215–1228.

Scitovsky, T. (1954). Two Concepts of External Economies. *Journal of Political Economy*, 62(2), 143–151.

Shapiro, C., 2011. Competition and Innovation: Did Arrow Hit the Bull's Eye? NBER Chapters, in: *The Rate and Direction of Inventive Activity Revisited*, pages 361-404, National Bureau of Economic Research, Inc.

Stephan, P. (1996). The Economics of Science. *Journal of Economic Literature*, 34(3), 1199-1235.

Wong, C. H., Siah, K. W., & Lo, A. W. (2018). Estimation of clinical trial success rates and related parameters. *Biostatistics*, 20(2), 273–286.